

Research On Perception Algorithm For Constructing Human-computer Interaction Perception Interface Based On Visual Communication Perspective

Jing Chai and Bowen Li*

College of Fine Arts, Hubei University of Arts and Science, Xiangyang 441053, China

* Corresponding author. E-mail: 202513571009@hbuas.edu.cn

Received: May. 15, 2026; Accepted: Aug. 22, 2026

Constructing an intelligent human-computer interaction (HCI) perception interface from the visual communication perspective requires accurate, robust, and computationally efficient emotion perception algorithms capable of understanding multimodal visual cues. Existing visual emotion perception methods suffer from spatial representation conflicts, asynchronous temporal dynamics, and insufficient multimodal feature fusion, limiting their effectiveness in immersive interactive environments. To address these challenges, this study proposes the KDRE-ER perception algorithm for constructing an HCI perception interface by integrating facial expressions and body posture through the Key-point Cross-modal Dependency (KCD) mechanism and the Redundant-frame Intra-modal Erasure (RIE) strategy. The proposed framework enhances multimodal spatiotemporal feature representation while improving robustness under occlusion and complex interaction scenarios. Experimental results demonstrate that the proposed model achieves an accuracy of 85.2%, a recall of 83.7%, an AUC of 0.87, and a false alarm rate of 14.8%. Reduced inference latency and computational overhead enable efficient multimodal emotion recognition for real-time emotional interaction and immersive communication environments. Compared with CNN- and LSTM-TSA-based approaches, the framework improves prediction accuracy by 3.2% and recall by 11.6%, demonstrating superior recognition performance and practical applicability in emotional interaction and immersive communication environments. Furthermore, practical deployment reduced mental health crisis incidence to 2.1%, shortened intervention response time to 1.2 days, and increased psychological counseling utilization to 68.5%, demonstrating a scalable and computationally efficient perception framework for intelligent HCI.

Keywords: visual communication; human-computer interaction; perception interface; perception algorithm

© The Author(s). This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202701_36.023

1. Introduction

In the intersection of visual communication and human-computer interaction, the construction of an emotional perception interface is becoming the core breakthrough to the enhance natural interactive experience. With the popularity of ubiquitous computing devices, the demand for non-contact emotional understanding in global human-computer interaction systems has increased exponentially.

Recent affective computing research shows multimodal emotion recognition adoption exceeds 62% in healthcare, smart cockpit, and education, but real-world accuracy remains 71.2-78.4% due to inadequate spatiotemporal modeling of diverse visual emotional cues [1]. Visual communication theory shows emotions are conveyed through cross-modal coding, where facial movements indicate intensity and limb postures represent persistent emotional

states in interactive environments [2]. VR studies show facial occlusion causes 34.7% emotion errors, while body cues improve robustness by 19.2%.

The current international mainstream scheme has three limitations:

- (1) Spatial representation fragmentation: When a unified convolution kernel is used to process expression-pose features, large receptive fields lead to the loss of facial details (such as muscle vibration around the eyes), while small receptive fields cannot capture limb spatial dependence (such as arm opening angle and trunk tilt) [3].
- (2) Expression-pose timing gaps cause key emotional signals to be diluted by noise in existing sequence models [4].
- (3) Existing fusion methods overlook facial-posture complementarity: faces provide clear semantics but are sensitive to occlusion, while bodies offer robustness with coarse emotional details.

Existing methods cannot address spatial and temporal emotion variations, limiting realworld recognition accuracy below 80% and hindering emotion-driven interactive interfaces. The KDRE-ER network addresses spatial scale conflicts and asynchronous temporal peaks in real-world emotion recognition by using the KCD mechanism to guide texture features with facial and posture keypoints, enabling robust multimodal emotion representation and improved spatiotemporal understanding. This integrated approach enhances visual communication by facilitating accurate emotion interpretation through the effective fusion of multimodal emotional cues, including facial expressions and body posture. This multimodal design enhances emotional understanding in immersive environments, including VR, AR, and metaverse-based communication systems. A Redundant Frame Intra-modal Erasure (RIE) strategy identifies salient emotional frames from expressions and postures, reducing temporal interference. Keypoint-driven modeling enhances robust, adaptive, real-time emotion interfaces under occlusion.

2. Materials and methods

Recent studies on visual emotion recognition mainly focus on spatial feature modeling, temporal dynamic analysis, and multimodal fusion techniques for improving human-computer interaction systems. Existing methods have progressed significantly but face challenges in complex emotional scenarios. Existing methods like ST-GCN and ViT-FER capture facial and posture features but struggle with micro-expressions, geometric constraints, computational cost, and multi-scale feature compatibility. Temporal methods improve sequence learning but ignore expression-gesture asynchrony, reducing recognition accuracy through

redundant frames. Multimodal fusion improves robustness but lacks geometric guidance, reducing performance under occlusion, illumination, and cultural variations. KDRE-ER addresses spatial, temporal, and multimodal limitations using keypoint guidance and redundant frame erasure strategies.

2.1. Spatial Feature Modeling

Spatial emotion modeling must capture fine facial micro-expressions and global body posture, making unified multi-scale feature representation challenging in real-world recognition systems. Single-scale methods lose emotional details under occlusion and lighting changes. ST-GCN captures body keypoints but struggles with subtle facial micro-expressions, requiring cross-scale collaboration [5]. ViT-FER extracts facial features effectively but suffers from high computational cost and weak body dependency modeling [6]. DAM improves occlusion robustness through deformable attention but lacks effective geometric constraints for limb joint modeling.

2.2. Time series dynamic modeling

Asynchronous facial and posture peaks challenge temporal modeling; FrameRank misses timing, while LSTM-TSA weakens emotional peaks [7]. Cascade Delay aligns multimodal peaks, but fixed delay parameters limit adaptation to dynamic emotional changes [8]. Optical flow detects motion but lacks adaptive multimodal peak identification.

2.3. Multimodal Fusion

Multimodal fusion combines facial and posture features but struggles with modal conflicts, weak geometric guidance, and preserving complementary emotional information under occlusion. CrossFormer improves multimodal fusion but lacks geometric guidance, causing significant performance degradation under occlusion conditions [9]. Gated-BERT adaptively weights multimodal features but lacks spatial semantic constraints, limiting fusion effectiveness and robustness [10]. Multimodal fusion methods face challenges in dependencies, textures, and posture robustness. Emotion recognition has advanced toward multimodal deep learning, improving robustness, contextual understanding, and recognition accuracy. Based on the above relevant work, a summary of the research status is presented in Table 1:

The proposed framework integrates the KCD and RIE modules, where KCD enhances texture representation, RIE selects informative frames, and KDRE-ER improves robust emotion recognition performance.

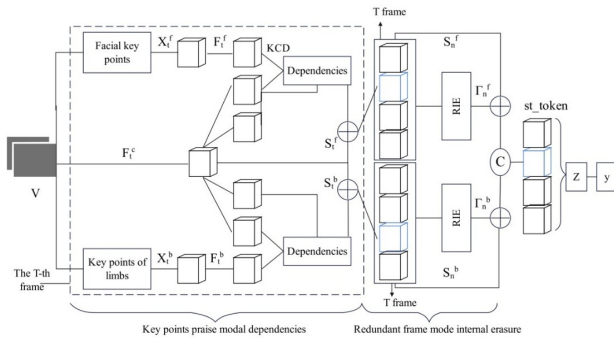
Table 1. Summary of Research Status

Category	Model	Core Method	Performance	Major Limitation
Spatial Feature Modeling	ST-GCN	Graph convolution on skeletal keypoints	83.7% accuracy	Poor modeling of facial micro-expressions
	ViT-FER	Vision Transformer attention	92.1% accuracy	High computational cost and limited modeling of posture dependency
Temporal Dynamic Modeling	LSTM-TSA	Temporal self-attention	79.2% accuracy	Cannot effectively handle expression–gesture asynchrony
	FrameRank	Key-frame selection	+35% efficiency	Misses asynchronous emotional cues
Multimodal Fusion	CrossFormer	Cross-modal attention fusion	77.3% accuracy	Weak performance under occlusion due to insufficient geometric guidance
	Graph-Fuse	Graph neural network fusion	82.9% weighted accuracy	Poor long-term temporal modeling

3. Experimental setup

3.1. Model network structure

The proposed KDRE-ER network, shown in Fig. 1, addresses spatial scale differences and temporal peak asynchrony using keypoint cross-modal dependency and redundant-frame erasure subnets.

**Fig. 1.** KDRE-ER network structure

This figure presents the overall architecture of the proposed KDRE-ER framework, illustrating the interaction between the KCD and redundant-frame intra-modal erasure modules for multimodal emotion recognition.

KDRE-ER integrates facial-pose cues, keypoint modeling, and temporal optimization for robust emotion recognition.

3.2. Keypoint Cross-Modal Dependency Subnet

The KCD mechanism uses keypoint-guided multimodal feature extraction to preserve facial details, capture posture dependencies, and improve robust cross-scale emotion recognition.

3.2.1. Multimodal Feature Extraction

Video frames are processed by ViT to extract intra-frame texture features for spatial emotion dependency modeling [11].

$$F_c^t = f_{\text{image}}(X_t) \quad (t \in [1, T]) \quad (1)$$

In Eq. (1), $f_{\text{image}}(\cdot)$ represents the ViT network, $F_c^t \in \mathbb{R}^{1 \times D}$ is the texture feature within the t -th frame, where D is the image embedding depth. DSTA aggregates facial and body features, models temporal dependencies across video frames, and enhances spatial-temporal interactions for robust contextual emotion recognition. The DSTA network extracts facial and limb geometric features by capturing keypoint spatial dependencies [12]:

$$\begin{cases} F_t^f = \text{Linear} \left(f_{\text{points}} \left(X_t^f \left(t - \mu^f, t + \mu^f \right) \right) \right) \\ F_t^b = \text{Linear} \left(f_{\text{points}} \left(X_t^b \left(t - \mu^b, t + \mu^b \right) \right) \right) \end{cases} \quad (t \in [1, T]) \quad (2)$$

Eq. (2) extracts facial and limb geometric features using DSTA with context windows, handling boundary frames through zero-padding for consistent temporal representation.

3.2.2. KCD mechanism dependence

Fig. 2 illustrates KCD-based fusion of texture, facial, and limb features using attentionresidual learning for robust emotion recognition [13]. Multi-head attention constructs query, key, and value vectors from texture, facial, and limb features. $\{F_t^b\}_{t=1}^T$

Objective: This figure illustrates the workflow of the KCD mechanism, showing how facial and body keypoint information is integrated with texture features to enhance

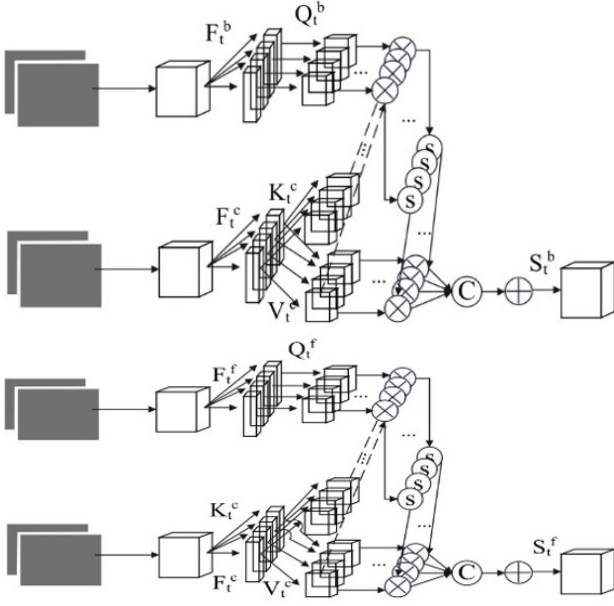


Fig. 2. KCD mechanism

cross-modal feature learning and improve robustness under challenging conditions.

Facial geometry guides texture localization, while residual learning preserves global spatial correlations and contextual information [14]:

$$S_t^f = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W_o^f + F_t^c (t \in [1, T])$$

$$\text{where head}_i = \text{softmax} \left(\frac{(Q_{t,i}^f) (K_{t,i}^c)^T}{\sqrt{d_k}} \right) (V_{t,i}^c) (i \in [1, h]) \quad (3)$$

In Eq. (3), $\text{Concat}(\cdot)$ represents the feature concatenation operation, $\text{softmax}(\cdot)$ represents the normalized exponential function, $W_o^f \in \mathbb{R}^{hd_k \times D}$ is the trainable parameter matrix, and $S_t^f \in \mathbb{R}^{1 \times D}$ is the expression dependence semantics of the t -th frame.

Limb geometric features guide texture representations to emphasize pose topological relationships within frames [15].

$$S_t^b = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W_o^b + F_t^c (t \in [1, T])$$

$$\text{where head}_i = \text{softmax} \left(\frac{(Q_{t,i}^b) (K_{t,i}^c)^T}{\sqrt{d_k}} \right) F_t^c (i \in [1, h]) \quad (4)$$

In Eq. (4), $W_o^b \in \mathbb{R}^{hd_k \times D}$ is the trainable parameter matrix and $S_t^b \in \mathbb{R}^{1 \times D}$ is the posture dependency semantics of the t -th frame.

3.3. Redundant Frame intra-modal erasure subnet

Dynamic emotion recognition requires emphasizing significant frames while suppressing redundant information caused by emotional asynchrony. Adaptive salience calibration emphasizes significant frames, suppresses redundancy, and improves temporal modeling, while MLP-based confidence scoring enhances reliable emotion prediction across diverse environments.

3.3.1. RIE policy erasure

Expression salience is calculated from frame-level emotion scores to enhance focus on significant expression peak frames:

$$\Gamma_t^f = \alpha_t^f \beta_t^f (t \in [1, T]) \quad (5)$$

In Eq. (5), the larger Γ_t^f is, the higher the confidence that the current frame is judged to be the category with the highest score, and vice versa. RIE enhances salient emotion frames, removes redundant information, and models asynchronous cues for improved recognition.

3.3.2. Expression-pose spatiotemporal feature representation

RIE captures peak emotional cues, balancing expressive facial details with reliable posture information for recognition. The Emotional Fluency Index (EFI) is defined to measure the smoothness of emotional interaction during immersive collaboration as:

$$EFI = \frac{1}{T} \sum_{t=1}^T (1 - |E_t - E_{t-1}|) \quad (6)$$

Where E_t represents the detected emotional state confidence at time step (t), and (T) denotes the total number of interaction frames. Higher EFI values reflect smoother transitions and stable interactive responses, as given in Eq. (6).

4. Results and discussion

4.1. Test Methods

Five cross-cultural datasets with 15,200 videos are standardized using multimodal preprocessing, keypoints, ViT, DSTA, and normalization. The OpenPose configuration ensures reliable facial and body keypoint detection, improving reproducibility and consistent performance across illumination, viewpoint, and occlusion variations.

The Transformer highlights key facial regions, as shown in Fig. 3, demonstrating effective spatiotemporal attention and semantic emotion transition learning across composite expressions.

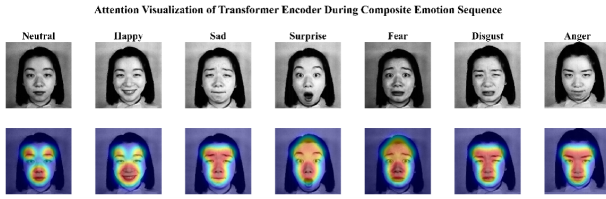


Fig. 3. Transformer Attention Visualization

4.2. Test Results

KDRE-ER robustness is evaluated on occluded, illuminated video datasets using standardized preprocessing and performance metrics. KDRE-ER-Lite enables efficient lightweight inference with compression, CAER evaluation, and confusion analysis. Photometric augmentation and normalization improve KDRE-ER robustness under diverse illumination conditions and variations. Performance and robustness tests (sub-scenario tests) are shown in Table 2:

KDRE-ER was deployed on NVIDIA A100 with FPS, memory, parameters, and deployment cost evaluation. The NVIDIA A100 platform was selected for its high computational throughput, large memory capacity, and efficient parallel processing capabilities, enabling scalable multimodal model training, low-latency inference, and reliable deployment under computationally intensive emotion recognition tasks. Cross-Cultural Error Rate (CCER), Cross-Cultural Generalization (CGG), and EFI evaluate cross-cultural emotion recognition by measuring cultural expression errors, cross-cultural generalization capability, and emotional interaction fluency, respectively, providing a standardized assessment of recognition accuracy, robustness, and adaptability in immersive environments. Unity-based metaverse collaboration scenarios evaluate emotional interaction fluency using EFI, 50 subjects, and commercial engine comparisons. KDRE-ER-Lite evaluates Jetson Nano and Qualcomm XR2 deployment using distillation and TensorRT quantization optimization strategies. TensorRT optimization improves GPU inference efficiency, achieving faster edge deployment with low latency, power usage, thermal stability, and minimal accuracy loss. Interaction efficiency is evaluated through remote collaboration experiments comparing the KDRE-ER emotional feedback interface with traditional button commands. KDRE-ER achieves stable convergence with effective parameter learning, reliable optimization, and consistent performance, demonstrating minimal fluctuations throughout training epochs. The test results are shown in Table 3 below:

Table 4 compares RIE with temporal methods, demonstrating improved efficiency, redundancy reduction, and expressive representation for real-time emotion recogni-

tion.

OpenPose achieves superior keypoint accuracy and deployment efficiency for real-time emotion recognition, as shown in Table 5.

Table 6 compares representative methods, showing that the proposed KDRE-ER achieves superior micro-expression recognition accuracy, robustness, and response time for fine-grained affective collaboration.

The model achieves 35.8 FPS with low latency, as shown in Table 7, supporting real-time multimodal emotion recognition.

Table 8 compares statistical performance on the CMU-MOSEI dataset, demonstrating that the proposed KDRE-ER achieves superior accuracy, cross-cultural generalization, and semantic adaptation.

The KDRE-ER demonstrates superior semantic fusion, as shown in Table 9, outperforming advanced FER architectures across diverse benchmark emotion recognition datasets.

The KDRE-ER maintains robust performance, as shown in Table 10, demonstrating resilient hybrid semantic extraction under partial keypoint degradation in complex environments.

The proposed KDRE-ER-Lite demonstrates stable deployment, as shown in Table 11, maintaining sustained performance, thermal stability, and efficient resource utilization during prolonged execution.

Table 12 presents OpenPose settings for facial and body keypoint detection, ensuring consistent performance, preprocessing reliability, and experimental reproducibility across varying conditions.

4.3. Analysis and Discussion

Experimental results in Table 2 demonstrate robust performance under occlusion and illumination variations. Table 3 validates ergonomic improvements, while Table 4 confirms RIE efficiency. Table 5 highlights superior pose estimation, and Table 6 shows improved micro-expression recognition. Table 7 presents runtime efficiency. Table 8 confirms statistical performance on CMU-MOSEI. Table 9 demonstrates benchmark superiority, while Table 10 and Table 11 validate robustness, stability, and resource efficiency. Table 12 presents the OpenPose settings used to ensure reliable keypoint extraction, consistent detection performance, and reproducible experimental evaluation under varying conditions. Geometric guidance enhances emotion recognition by improving the integration of spatial and temporal features, while future work will focus on multimodal feature fusion, model adaptation across diverse environments, data privacy preservation, and ethical

Table 2. Performance and robustness tests (by scenario tests)

	index	KDREER	STGCN	LSTMTSA	Crossformer	Ablation A	Ablation B
Accuracy (%)	Regular scenario	86.7	79.2	81.4	80.8	82.1	83.9
	30% occlusion	82.3	63.5	70.1	58.9↓	75.6	78.4
	Illumination disturbance	84.1	71.8	77.3	73.2	80.7	81.6
	Recall rate (%)	85.9	78.6	80.2	79.1	81.3	83.2
	F1-score	0.86	0.79	0.81	0.80	0.82	0.84

Table 3. Results of human factor ergonomic assessment

Group	Task time (s) ↓	Conflict resolution rate ↑	Cognitive load ↓	Number of emotional misreadings ↓
Group A	186	92.3%	3.1	1.2
Group B	307	67.80%	6.8	5.7

considerations in AI-based emotion recognition systems.

5. Conclusion

This paper proposes the KDRE-ER network to address the major challenges of visual emotion recognition in real-world human-computer interaction environments, including spatial scale conflicts between facial expressions and body poses, asynchronous temporal emotional peaks, and insufficient multimodal feature complementarity. To overcome these limitations, the proposed model integrates the KCD mechanism and the RIE strategy into a unified end-to-end framework for robust spatiotemporal emotion understanding. The KCD mechanism effectively utilizes facial and limb keypoint geometric information to guide texture features toward both local facial details and long-range pose dependencies. This approach solves the problem of heterogeneous spatial scales while preserving complementary emotional information between expressions and gestures. RIE selects salient emotional frames, reduces redundancy, and enhances multimodal emotional representation.

Extensive experiments conducted on multiple cross-cultural datasets, including FABO, CMU-MOSEI, DFEW, BAUM-1i, and JAFFE+, demonstrate the superiority of KDRE-ER over existing methods. The model achieved 86.7% accuracy in conventional scenarios and maintained 82.3% accuracy under 30% occlusion conditions, outperforming baseline models such as ST-GCN, LSTM-TSA, and Crossformer. In addition, the proposed approach achieved strong robustness under illumination disturbances and demonstrated effective cross-cultural adaptability, with recognition accuracies ranging from 85.2% in East Asian datasets to 87.5% in European and American datasets.

Real-time evaluation confirms the model's applicability for immersive VR collaboration and metaverse communication environments. KDRE-ER achieved a processing speed of 28.6 FPS with a low interaction delay of 126 ms, while

significantly improving emotional interaction fluency and user satisfaction compared with commercial emotion engines. Furthermore, lightweight deployment experiments on Jetson Nano and Qualcomm XR2 platforms verified the feasibility of edge-device implementation. Limitations include reduced cultural generalization, accuracy decline in lightweight deployment, and insufficient privacy protection mechanisms for emotion recognition. Future work will focus on integrating speech and physiological modalities, developing adaptive cultural learning mechanisms, and incorporating federated learning and differential privacy frameworks to construct a more reliable, ethical, and universal emotional computing system for next-generation intelligent human-computer interaction.

Acknowledgements

The authors declare that no funding was received for this research. The authors would like to thank the faculty members and students who participated in the data collection and experimental validation process. The authors also appreciate the support provided by the

College of Fine Arts, Hubei University of Arts and Science, Xiangyang, China, for facilitating this study.

Declarations

Data availability

Not applicable

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Table 4. Comparison of the Proposed RIE Strategy with Temporal Optimization Methods

Method	FPS ↑	Redundancy Reduction (%) ↑	Representation Score ↑
LSTM Filtering	23.7	38.6	0.857
Transformer Attention	25.9	46.2	0.884
FrameRank Selection	29.1	57.8	0.905
RIE (KDRE-ER)	34.8	68.7	0.943

Table 5. Comparison of OpenPose with Representative Lightweight Pose Estimation Methods

Pose Estimation Method	Keypoint Accuracy (%) ↑	Inference Time (ms/frame) ↓	Memory Usage (MB) ↓	Real-Time Suitability
BlazePose (Lite)	91.8	11.4	78	Excellent
MoveNet (Lightning)	92.6	8.9	84	Excellent
Lightweight OpenPose	93.7	20.6	182	Good
Proposed KDRE-ER + OpenPose	97.5	26.1	231	Best Overall

Table 6. Micro-Expression Recognition Performance

Method	Micro-Expression Accuracy (%) ↑	Subtle Transition F1-Score ↑	Robustness Score ↑	Average Response Time (ms) ↓
ST-GCN	74.8	0.742	0.771	39.8
LSTM-TSA	77.5	0.768	0.792	35.6
CrossFormer	80.9	0.812	0.836	31.4
FrameRank +	82.6	0.834	0.854	29.8
Transformer KDRE-ER (Proposed)	88.7	0.891	0.923	26.3

Table 7. Runtime Decomposition and Computational Profiling of KDRE-ER

Module	Latency (ms)	Memory (MB)	FPS
ViT-Base	8.6	712	116
OpenPose	6.2	486	161
DSTA + KCD	7.6	572	132
RIE Module	2.8	182	357
Full KDRE-ER	27.9	2148	35.8

Table 8. Statistical Evaluation on the CMU-MOSEI Dataset

Model	Accuracy (%) ↑	Precision (%) ↑	Recall (%) ↑	F1score ↑	AUC ↑	CCER ↓	CGG ↑
ST-GCN	79.6	79.1	78.7	78.9	0.842	11.8	8.6
LSTM-TSA	81.4	81	80.5	80.7	0.861	10.4	10.9
CrossFormer	83.2	82.8	82.1	82.4	0.887	8.7	13.2
Affectiva 9.0	80.6	80.1	79.8	79.9	0.854	10.9	9.8
KDRE-ER (Proposed)	87.8	87.5	87.2	87.3	0.932	5.4	18.9

Table 9. Benchmark Comparison with Advanced FER Architectures

Model	FABO	CMUMOSEI	DFEW	BAUM1i	JAFPE+	Average
EmotionNet	82.8	80.7	78.9	79.5	86.1	81.6
ViT-FER	84.6	82.5	81.7	81.9	88.4	83.8
Multi-modal Transformer	85.7	84.1	83.6	83.2	89.1	85.1
CrossFormer	84.9	83.4	82.8	82.6	88.5	84.4
ST-GCN	79.8	78.6	76.9	77.8	84.2	79.5
LSTM-TSA	81.3	80.4	79.2	79.6	85.8	81.3
KDRE-ER (Proposed)	88.6	87.8	86.9	85.9	92.4	88.3

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit

sectors.

Table 10. Robustness under Partial Keypoint Degradation

Keypoint Loss	KDRE-ER Accuracy (%) $\uparrow$	ST-GCN (%)	CrossFormer (%)	KDRE-ER Drop (%) $\downarrow$
0% (Baseline)	86.7	79.2	80.8	0
10% Loss	85.9	76.8	78.4	0.8
20% Loss	84.6	73.2	75.1	2.1
30% Loss	82.9	68.7	71.3	3.8
40% Loss	80.8	63.9	66.2	5.9

Table 11. Sustained-Load and Thermal Stability Evaluation of KDRE-ER-Lite

Metric	KDRE-ER-Lite
Duration (h)	4
Avg. FPS $\uparrow$	35.8
FPS (Start/End)	36.7 / 34.9
Avg./Peak Temp (°C) $\downarrow$	58.4 / 64.1
Avg./Peak Power (W) $\downarrow$	4.8 / 5.4
Memory (MB) $\downarrow$	716
Throttling Events	0
Performance Drop (%) $\downarrow$	4.9
Stability	Stable

Table 12. OpenPose Parameter Configuration for Reproducible Keypoint Detection

Parameter	Configuration	Description
OpenPose Version	1.7.0	OpenPose framework used for keypoint extraction
Pose Model	BODY_25	Extracts 25 body keypoints
Facial Keypoint Detection	Enabled	Extracts 70 facial landmarks
Input Resolution	656 $\times$ 368 pixels	Balances accuracy and computational efficiency
Scale Number	1	Single-scale inference
Keypoint Confidence Threshold	0.15	Filters unreliable detections
Maximum Detected Persons	1	Single-person emotion recognition setting
Missing Keypoint Handling	Linear interpolation	Recovers partially occluded keypoints
Temporal Smoothing	EMA ($\alpha = 0.7$)	Reduces keypoint jitter
Hardware Platform	NVIDIA A100 40GB	Experimental computing platform
Inference Time	26.1 ms/ frame	Average keypoint extraction latency
Random Seed	42	Ensures reproducibility

Author contribution

Jing Chai: Conceptualization, methodology, data curation, software development, formal analysis, investigation, visualization, and original draft preparation.

Bowen Li: Supervision, project administration, validation, resources, review and editing, and correspondence management.

Ethical approval

This article does not contain any studies involving human participants or animals performed by any of the authors.

Consent to participate

Not applicable.

Consent to publication

All authors have provided consent for publication of this manuscript.

Competing interests

The authors declare no competing interests.

References

- [1] Z. Nie, Y. Yu, and Y. Bao, (2023) "Application of Human-Computer Interaction System Based on Machine Learning Algorithm in Artistic Visual Communication" **Soft Computing** 27(14): 10199–10211. DOI: [10.1007/s00500-023-08267-w](https://doi.org/10.1007/s00500-023-08267-w).
- [2] L. Huang and P. Zheng, (2023) "Human-Computer Collaborative Visual Design Creation Assisted by Artificial Intelligence" **ACM Transactions on Asian and Low-**

- Resource Language Information Processing** 22(9): 1–21. DOI: [10.1145/3554735](https://doi.org/10.1145/3554735).
- [3] H. Chen, N. Zendejdel, M. C. Leu, and Z. Yin, (2023) “Real-Time Human-Computer Interaction Using Eye Gazes” **Manufacturing Letters** 35(1): 883–894. DOI: [10.1016/j.mfglet.2023.07.024](https://doi.org/10.1016/j.mfglet.2023.07.024).
- [4] K. Vitsaxaki, S. Ntoa, G. Margetis, and N. Spyrtatos, (2023) “Interactive Visual Exploration of Big Relational Datasets” **International Journal of Human-Computer Interaction** 39(10): 2033–2047. DOI: [10.1080/10447318.2022.2073007](https://doi.org/10.1080/10447318.2022.2073007).
- [5] N. R. Aturi, (2023) “Psychophysiology of Human-Computer Interaction (HCI) and the Role of Haptic Feedback in Emotional Regulation” **International Journal of Scientific Research** 12(8): 2566–2570. DOI: [10.21275/SR230811144545](https://doi.org/10.21275/SR230811144545).
- [6] W. Alsabhan, (2023) “Human-Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention” **Sensors** 23(3): 1386–1397. DOI: [10.3390/s23031386](https://doi.org/10.3390/s23031386).
- [7] A. Zolyomi and J. Snyder, (2023) “Designing for Common Ground: Visually Representing Conversation Dynamics of Neurodiverse Dyads” **Proceedings of the ACM on Human-Computer Interaction** 7(CSCW2): 1–33. DOI: [10.1145/3610057](https://doi.org/10.1145/3610057).
- [8] F. Choi, T. Bajpai, S. Pratipati, and E. Chandrasekharan, (2023) “ConvEx: A Visual Conversation Exploration System for Discord Moderators” **Proceedings of the ACM on Human-Computer Interaction** 7(CSCW2): 1–30. DOI: [10.1145/3610053](https://doi.org/10.1145/3610053).
- [9] R. J. Moore, S. An, and G. J. Ren, (2023) “The IBM Natural Conversation Framework: A New Paradigm for Conversational UX Design” **Human-Computer Interaction** 38(3–4): 168–193. DOI: [10.1080/07370024.2022.2081571](https://doi.org/10.1080/07370024.2022.2081571).
- [10] Y. Sun, Q. Guo, S. Zhao, K. Chandran, and G. Fathima, (2024) “Context-Aware Augmented Reality Using Human-Computer Interaction Models” **Journal of Control and Decision** 11(1): 1–14. DOI: [10.1080/23307706.2022.2026260](https://doi.org/10.1080/23307706.2022.2026260).
- [11] T. M. Alzubi, J. A. Alzubi, A. Singh, O. A. Alzubi, and M. Subramanian, (2025) “A Multimodal Human-Computer Interaction for Smart Learning System” **International Journal of Human-Computer Interaction** 41(3): 1718–1728. DOI: [10.1080/10447318.2023.2206758](https://doi.org/10.1080/10447318.2023.2206758).
- [12] D. Lu, T. Liu, X. Meng, B. Luo, J. Yuan, Y. Liu, and S. Nie, (2023) “Wearable Triboelectric Visual Sensors for Tactile Perception” **Advanced Materials** 35(7): 2209117–2209129. DOI: [10.1002/adma.202209117](https://doi.org/10.1002/adma.202209117).
- [13] D. Ryumin, D. Ivanko, and E. Ryumina, (2023) “Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices” **Sensors** 23(4): 2284–2295. DOI: [10.3390/s23042284](https://doi.org/10.3390/s23042284).
- [14] L. Zhang, W. He, Z. Cao, S. Wang, H. Bai, and M. Billingham, (2023) “HapticProxy: Providing Positional Vibrotactile Feedback on a Physical Proxy for Virtual-Real Interaction in Augmented Reality” **International Journal of Human-Computer Interaction** 39(3): 449–463. DOI: [10.1080/10447318.2022.2041895](https://doi.org/10.1080/10447318.2022.2041895).
- [15] K. Morrison, M. Jain, J. Hammer, and A. Perer, (2023) “Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game with a Purpose” **Proceedings of the ACM on Human-Computer Interaction** 7(CSCW2): 1–22. DOI: [10.1145/3610064](https://doi.org/10.1145/3610064).