Creative Guidance of Intelligent Emotion Recognition in Video Art

Weixing Chen, Yubo Zhou*
Namseoul University, Tian'an City, Korea

***Abstract*—This study optimizes the existing emotion recognition model to improve the application effect of emotion recognition technology in the guidance of video art creation. It also compares the performances of Multimodal Sentiment Analysis (MMSA), Multimodal Sequence Encoder (MuSE), and the optimized model through a series of simulation experiments. In the fine-grained accuracy of emotion recognition, the optimized model performs well in micro-expression recognition accuracy and multimodal fusion effect, and scores 4.17 and 4.96, respectively. In the robustness test under a complex background, the background interference resistance score of the optimized model is 4.30, and the modal mismatch processing score is 4.38. In the experiment on the effectiveness of emotion recognition in creative guidance, the optimized model scores 4.40 and 3.66 in terms of artistic expression enhancement and creative feedback accuracy, respectively. The experimental results show that the optimized model is superior to the existing models in several key dimensions of emotion recognition, especially in enhancing the emotional expression of artistic works and providing creative guidance. Therefore, this study provides some reference for the application and development of emotion recognition technology in video art creation.**

***Keywords*—Emotional recognition; image art; artistic expressiveness; creative feedback**

I. Introduction

With the rapid development of artificial intelligence (AI) technology, affective computing has gradually gained widespread attention as a crucial research direction in human-computer interaction. Affective computing aims to endow computers with the ability to recognize, understand, and express emotions, thereby enabling more natural and intelligent human-computer interaction [1]. In recent years, emotion recognition technology has developed rapidly in multimedia, entertainment, healthcare, and education. It has also demonstrated enormous potential in enhancing user experience and interaction quality [2]. As an artistic form that conveys emotions and ideas through visual language, audio-visual art creation relies heavily on emotional expression and audience emotional resonance [3]. However, traditional audio-visual art creation mainly depends on artists' experience and intuition. This hinders its ability to accurately capture the emotional feedback of different audiences and systematically integrate such emotional information into creative decisions [4, 5].

Despite the significant progress of existing emotion recognition technology in various scenarios, obvious research gaps remain in the field of audio-visual art creation. Firstly, current models mostly focus on single-modal or coarse-grained emotion classification, lacking effective recognition of fine-grained emotional information such as microexpressions and subtle emotional changes—details that are particularly critical for artistic expression. Secondly, the diverse visual styles and complex scene changes in audio-visual art result in insufficient robustness and real-time performance of existing models, which fail to meet the needs of creative practice. Thirdly, existing literature rarely explores how emotion recognition results can be linked to specific creative behaviors, such as shooting, editing, color scheduling, and narrative rhythm. Thus, a clear disconnect persists between "technology and artistic creation".

To address the above research gaps, this study optimizes existing multimodal emotion recognition models and proposes an intelligent emotion recognition framework for guiding audio-visual art creation. Through an improved multimodal fusion mechanism, the model significantly enhances microexpression recognition capabilities and robustness in complex backgrounds. Its structural design also improves the usability of converting emotional information into creative decisions. In addition, this study constructs a comprehensive evaluation system that includes both technical and art creation-related indicators, verifying the model's effectiveness from both technical performance and practical creation perspectives. The main innovative contributions of this study are as follows:

1) Explicitly propose the task requirements of emotion recognition "for guiding audio-visual art creation", bridging the disconnect between technology and artistic creation in existing research.

2) Design an optimized multimodal fusion architecture to improve the accuracy of fine-grained emotion recognition and robustness in complex backgrounds.

3) Establish a new evaluation system that balances technical performance and creative value, providing an operable evaluation paradigm for the application of emotion recognition technology in artistic creation.

The subsequent structure of this study is arranged as follows: Section II reviews relevant research on affective computing and emotion recognition, and sorts out their application status in artistic creation. Section III elaborates on the theoretical basis of affective computing and introduces the overall design and optimized architecture of the intelligent emotion recognition model. Section IV presents the experimental design, including datasets, experimental processes, simulation experiments, and ablation experiments, along with analysis and discussion of results. Section V summarizes the research findings, points out limitations, and proposes future research directions.

*Corresponding author.

II. Literature Review

In the field of affective computing and emotion recognition, existing studies have systematically explored emotion modeling from diverse modalities and application scenarios. Lian et al. (2023) found that deep learning-based text emotion recognition algorithms achieved high accuracy in sentiment polarity and fine-grained emotion analysis. However, when extending from text emotions to multimodal contexts such as speech and images, models still faced significant challenges in cross-modal alignment and feature fusion, especially lacking robustness in complex scenarios [6]. Kopalidis et al. (2024) proposed a facial expression recognition model based on a convolutional neural network (CNN). Its recognition accuracy exceeded 90% on standard datasets, proving the effectiveness of the visual modality in static or regularized scenarios. Yet, the method focused primarily on single-modal visual inputs and lacked targeted handling of typical artistic scenarios such as background interference and stylized visual elements [7]. Liu and Li (2023) discovered that fusing multiple physiological signals—such as heart rate and skin conductance responses—into affective computing models significantly improved emotion recognition accuracy. This indicated that multimodal physiological signals played a complementary and corrective role in emotional states. However, most related works remained confined to laboratory or medical contexts, with relatively limited discussions on adaptability to open and complex environments like video art creation [8].

Regarding multimodal emotion recognition and its applications in creative fields, Dash and Agre (2024) pointed out that music generation systems incorporating emotion recognition technology could dynamically adjust rhythm and melody based on audience emotional feedback. Thus, it can enhance the emotional appeal of musical works. Nevertheless, their multimodal modeling focused mainly on audio and simple interaction signals, lacking comprehensive modeling of complex information such as visual content and contextual semantics [9]. Monser and Fadel (2023) found that intelligent emotion recognition technology helped visual art creators understand audience emotional responses and adjust expression techniques accordingly. However, their research mainly proceeded from the perspective of "emotion analysis-assisted evaluation". Meanwhile, they did not deeply construct a systematic mapping mechanism from emotion recognition results to specific creative decisions (e.g., shot language, color configuration, narrative rhythm) [10]. Li (2024) proposed a method combining emotion recognition with film production, using real-time audience emotional responses to optimize editing rhythm and sound effect design, thereby improving the viewing experience. However, the model design emphasized the control of overall emotional curves, with relatively superficial discussions on microexpressions, fine-grained emotional transitions, and multimodal fusion strategies [11]. Wang (2024) discovered that applying emotion recognition to film character emotion analysis helped directors and screenwriters shape character images and narrative tension more accurately. Yet, most related works focused on character-level emotion label prediction, with little consideration of how to feed recognition results back into real-time adjustments and iterative optimization in the creative process [12].

In the direction of optimizing the rhythm and structure of audio-visual works, Zhang et al. (2024) proposed a film editing method based on emotion recognition. By analyzing changes in audience emotions, the method optimized editing rhythm, making the film's emotional expression more coherent and concentrated. However, the approach remained mostly at the rule-driven or experience-driven level of rhythm regulation, failing to fully leverage multimodal deep features and robust modeling in complex contexts [13]. Zhang and Pu (2024) showed that introducing emotion recognition technology in animation production enabled more delicate character expression and movement design, significantly enhancing audience emotional resonance. Nevertheless, most related systems centered on "character image shaping" and lacked systematic support for "emotional structure control and creative guidance of the entire audio-visual work" [14].

Synthesizing the aforementioned studies, it is evident that multimodal emotion recognition technology has achieved significant progress in fields such as text, speech, facial expressions, physiological signals, music, film, and animation. However, obvious gaps remain in systematic research oriented toward guiding audio-visual art creation. Firstly, most studies focus on improving "recognition accuracy" but lack systematic evaluation and targeted optimization of model robustness, real-time performance, and generalization ability in complex audio-visual contexts. Secondly, existing works apply emotion recognition to scenarios like music generation, film editing, or character shaping. However, they generally lack interpretable mapping and creative guidance mechanisms from "emotion recognition results to specific creative decision parameters (e.g., shot editing, color style, narrative rhythm)". Thirdly, multimodal fusion strategies mostly remain at relatively fixed early/late fusion or simple weight superposition; they lack targeted structural designs and training strategies to address issues such as complex backgrounds, modal mismatches, and fine-grained emotional changes.

Based on this, this study builds on previous research and further focuses on audio-visual art creation scenarios. The fusion mechanism and robustness design of multimodal emotion recognition models are optimized, and evaluation dimensions closely related to artistic expression and creative feedback are introduced. Thus, it constructs a tighter bridge between "technical performance improvement" and "artistic creation usability". This fills the key gap in existing research regarding the guidance of video art creation.

III. Emotion Recognition Based on AI

A. Affective Computing and Emotion Recognition Technology

Affective computing is an important branch of computer science and AI, making computers have the ability to perceive, understand, explain, and simulate human emotions [15]. The research content of affective computing includes the identification, expression, and generation of emotion and the construction of emotional intelligence [16]. With the development of technology, affective computing has gradually expanded from the initial theoretical exploration to a wide range of applications, such as HCI, intelligent education, mental health, entertainment, and so on [17]. Emotion recognition technology is one of the core tasks of affective

computing, which aims to identify human emotional state from various data sources through computer algorithms [18]. At present, emotion recognition technology is mainly divided into two categories: single-mode emotion recognition and multi-mode emotion recognition [19], as shown in Table I:

TABLE I. EMOTION RECOGNITION TECHNOLOGY

Technology	Application	Description
Single-mode emotion recognition technology	Facial expression recognition	Emotion recognition based on facial expressions is the most common method in emotion recognition technology. Early research usually relies on artificial feature extraction, such as local binary pattern and scale-invariant feature transformation. However, with the rise of deep learning, the model based on a CNN has become mainstream.
	Speech emotion recognition	The speech signal contains rich emotional information, such as intonation, speech speed, volume and so on. Speech emotion recognition usually judges the emotional state of the speaker by extracting these features. Traditional methods include the hidden Markov model and the support vector machine.
	Text sentiment analysis	Text sentiment analysis mainly analyzes the emotional tendency in the text through natural language processing technology. Traditional methods include the sentiment dictionary method and machine learning classifiers, such as support vector machines and naive Bayes.
Multi-mode emotion recognition	Feature extraction	Representative features are extracted from different modal data, such as geometric and texture features from facial expressions, frequency and energy features from speech, word vectors, and emotional vocabulary from texts.
	Feature fusion	Feature fusion is a key step in multimodal emotion recognition. Common methods include early fusion and late fusion. Early fusion combines features from different modes directly in the feature extraction stage, while late fusion is decision-level fusion based on the emotion classification results of each mode.
	Emotional classification	The fused features or classification results are input into the emotion classifier to identify the final emotion state. Commonly used classifiers include support vector machine, random forest and deep neural network.

To sum up, emotion calculation and emotion recognition technology have laid an important foundation for the application of intelligent emotion recognition in video art [20]. With the continuous progress of technology, emotion recognition plays an increasingly important role in the creation of video art, providing a new channel for emotional communication between artists and audiences [21].

B. The Combination of Intelligent Emotion Recognition and Artistic Creation

With the continuous development of AI technology, emotion recognition technology is not limited to traditional application fields such as HCI and mental health, and its potential is gradually explored and applied in artistic creation [22]. Artistic creation, as an important way for human beings to express emotions, thoughts, and concepts, usually depends on the creator's deep understanding of emotions and the audience's emotional resonance [23]. The introduction of intelligent emotion recognition technology can inject new vitality into the artistic creation process. By capturing and analyzing the emotional reaction of the audience, it can help artists to make more accurate emotional expression and adjustment in their creation, thus enhancing the emotional appeal of their works. In artistic creation, emotion recognition technology can help creators from many aspects, including real-time feedback of creative effects, personalized customization of artistic content, and optimization of emotional communication methods of works [24]. Through intelligent emotion recognition technology, the creator can better understand the emotional needs and reactions of the audience, provide a scientific basis for the creative process, and thus realize the deep integration of artistic creation and emotion recognition.

As an intuitive and infectious art form, visual art has long been used to express complex emotions and thoughts. The application of intelligent emotion recognition technology in visual arts can help artists capture and express emotions more accurately and optimize the audience's viewing experience. Music, as an art form with sound as the medium, has a strong emotional expression ability [25]. The application of intelligent emotion recognition technology in music creation is mainly reflected in music generation driven by emotion recognition, music creation with real-time emotion feedback, and the combination of emotion recognition and music recommendation systems. Image art includes many art forms such as movies, animations, and videos, and has strong emotional expression and narrative ability [26]. The application of intelligent emotion recognition technology in video art can help the creator to better understand the emotional reaction of the audience and optimize the emotional expression of video works [27].

In short, the combination of intelligent emotion recognition technology and artistic creation expands the means and forms of artistic creation, while providing the audience with a more personalized and interactive artistic experience [28]. With the continuous progress of technology, emotion recognition plays an increasingly important role in artistic creation and builds a closer bridge for emotional communication between artists and audiences [29].

C. Model Construction of Intelligent Emotion Recognition Technology in Video Art

The application of intelligent emotion recognition technology in video art depends on the complex model construction process, which includes multiple stages from data collection to model training, optimization, and application. The construction of an emotion recognition model requires a deep understanding of the basic theories of affective computing and AI, such as machine learning, deep learning, and multimodal data processing. Meanwhile, it is necessary to combine the relevant theories of psychology, cognitive science, and art to ensure that the model not only has accurate emotional

recognition ability, but also can be effectively applied to artistic creation to enhance the emotional expression of works. In the process of model construction, it is important to consider the characteristics of image art, such as diverse visual elements, dynamic scenes, and complex interaction with audience emotions. Based on these characteristics, to build an emotion recognition model suitable for video art, it is usually necessary to integrate multi-modal emotion recognition technology, an emotion calculation model, and an artistic expression model to accurately capture and express emotions. The designed optimized model is shown in Fig. 1:

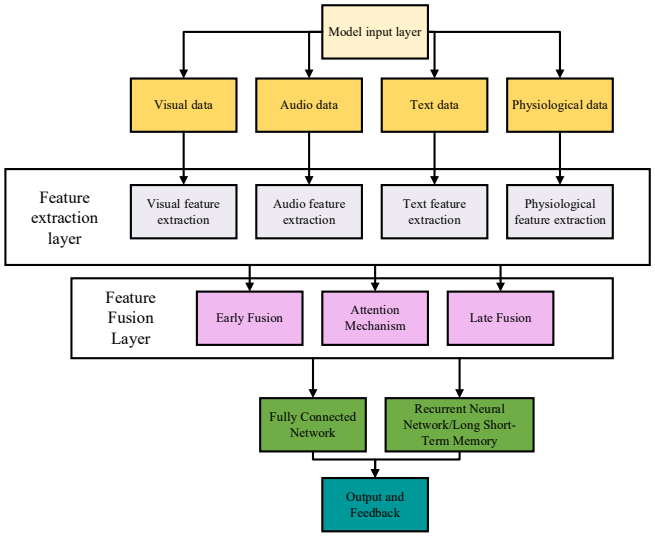

Fig. 1. Intelligent emotion recognition model architecture.

Through the above model framework, intelligent emotion recognition technology can accurately capture and feedback the audience's emotions in video art, thus enhancing the emotional expression and interactive experience of video works.

IV. Comparison between the Performance of Intelligent Emotion Recognition Model and Simulation Results

A. Performance Analysis of an Intelligent Emotion Recognition Model

The Interactive Emotional Motion Capture (IEMOCAP) dataset is selected in the experiment. It is especially suitable for the study of emotion recognition because it contains multimodal records designed for emotion recognition. The IEMOCAP dataset includes 302 video conversations, which are recorded in five sessions, and each session contains two speakers. The dataset labels nine different emotions: anger, excitement, fear, sadness, surprise, depression, happiness, disappointment, and neutrality. In addition, it also provides continuous labeling about pleasure, arousal, and dominance. The computer used in the experiment is equipped with an Intel Xeon Gold 6230 processor, which has up to 20 cores /40 threads, provides powerful multi-thread processing ability, and can process a large amount of data efficiently. In order to speed up the training of a deep learning model, NVIDIA Tesla V100 is selected in the experiment, which has 32 GB of HBm2 memory and supports the parallel computing of deep learning frameworks such as TensorFlow and PyTorch, which significantly improves the training speed of the model. The experimental system is equipped with 128GB DDR4 ECC memory to ensure the stability and performance of the system when dealing with large data sets. The experimental data is stored on a 2TB NVMe SSD, which provides fast data reading and writing speed, reduces the I/O bottleneck, and ensures efficient data processing in the experimental process. Multimodal Sentiment Analysis (MMSA) and Multimodal Sequence Encoder (MuSE) are selected as the contrast models in the experiment. The experimental parameters are set as follows: the image input dimension is 2242243, and the audio input is the feature matrix of 128128. CNN layer, using 32 convolution kernels of 33, the activation function is ReLU, followed by the maximum pool layer of 22. In the deep convolution layer, 64 convolution kernels of 33 are used, and the activation function is ReLU, followed by the maximum pool layer of 22. In the fully connected layer, the number of neurons is 256, and the activation function is ReLU. Dropout is used to prevent over-fitting, and the discard rate is set to 0.5. The indicators of experimental comparison are accuracy, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Kappa coefficient, and Mean Squared Error (MSE). The experimental results are shown in Fig. 2:

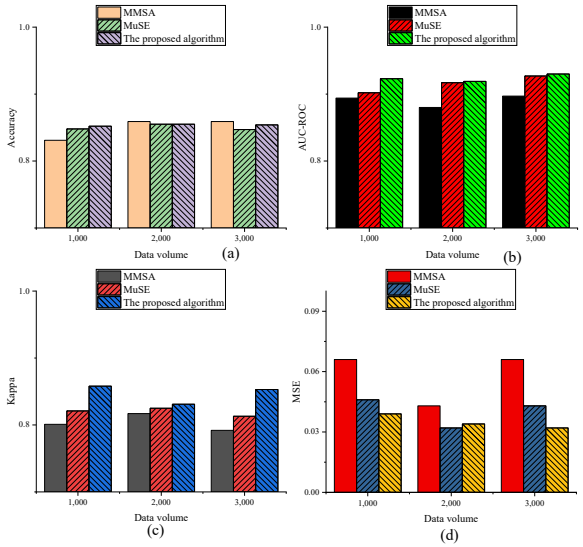

Fig. 2. Performance comparison results: (a) Accuracy, (b) AUC-ROC, (c) Kappa coefficient, (d) MSE.

In Fig. 2, in terms of accuracy, when the data volume is 1000, the MMSA, MuSE, and optimized models' accuracy are 0.831, 0.848, and 0.852. It can be found that the performance of the optimized model is slightly better than that of MMSA and MuSE models in the case of less data, especially when the accuracy of MuSE is close, the optimized model still has certain advantages. When the data volume is 2000, the accuracy of MMSA rises to 0.859, MuSE is 0.855, and the optimized model is 0.855. With the increase in data volume, the performance of the MMSA and optimized models is very close, and the MuSE model is slightly inferior. This shows that with the increase in data volume, the accuracy of the MMSA and optimized models has been improved, but the gap between them has narrowed. When the data volume is 3000, the accuracy of MMSA is maintained at 0.859, while that of MuSE

is slightly reduced to 0.847, and that of the optimized model is 0.854. When the data volume reaches 3000, the accuracy of the MMSA and optimized models shows good stability, while the accuracy of MuSE decreases slightly. This may indicate that MuSE is more sensitive to data under a larger data volume, which leads to the fluctuation of model accuracy.

In terms of AUC-ROC, under different data sets, the AUC-ROC values of the optimized model reach 0.923, 0.919, and 0.93, respectively, which shows significant advantages when the data volume is small. With the increase of data volume, the AUC-ROC performance of MuSE is gradually improved, and it tends to be close to that of the optimized model, which shows that MuSE has good scalability when dealing with large-scale data. However, MMSA is relatively stable when the data volume increases, but it fails to show significant improvement.

In terms of the Kappa coefficient, when the data volume is 1000, the Kappa coefficients of the MMSA and MuSE models are 0.801 and 0.821, and that of the optimized model is 0.858. When the data volume is 2000, the Kappa coefficient of MMSA rises to 0.817, that of MuSE is 0.825, and that of the optimized model is 0.831. When the data volume is 3000, the Kappa coefficient of MMSA drops slightly to 0.782, that of MuSE is 0.813, and that of the optimized model rises to 0.853. With a larger amount of data, the Kappa coefficient of the optimized model remains at a high level, while the MuSE model shows stability, but it fails to surpass the optimized model. The Kappa coefficient of the MMSA model decreases slightly, which may be affected by the complexity increase when the data volume increases.

In terms of MSE, when the data amount is 1000, the MSE of the MMSA, MuSE, and optimized models are 0.066, 0.046, and 0.039. In the case of small data, the MSE of the optimized model is the lowest, which shows that its prediction error is the smallest and it has better precision. MuSE is also better than MMSA. When the data volume is 2000, the MSE of the MMSA and MuSE models is significantly reduced to 0.043 and 0.032, and the MSE of the optimized model is 0.034. With the increase in data volume, the MSE of MMSA has been significantly improved, but the MuSE performs better, reaching the lowest MSE value. The MSE of the optimized model is slightly higher than that of MuSE, but it still remains at a low level. When the data volume is 3000, the MSE of MMSA rises to 0.066, that of MuSE is 0.043, and that of the optimized model is further reduced to 0.032. Under a larger data volume, the optimized model once again showed its advantages, reaching the lowest MSE value, indicating that it can still maintain a low prediction error when dealing with large-scale data. The MSE of MMSA fluctuates, while the MuSE maintains a relatively stable performance.

B. Simulation Experiment Comparison of Intelligent Emotion Recognition Model

The experiment adopts a scoring system of 1-5 points. The higher the score, the higher the representative satisfaction. This study selects four indicators, namely, fine-grained accuracy of emotion recognition, response speed, robustness in complex background, and effectiveness of creative guidance. For each indicator, it is divided into three dimensions, and the results are shown in Fig. 3:

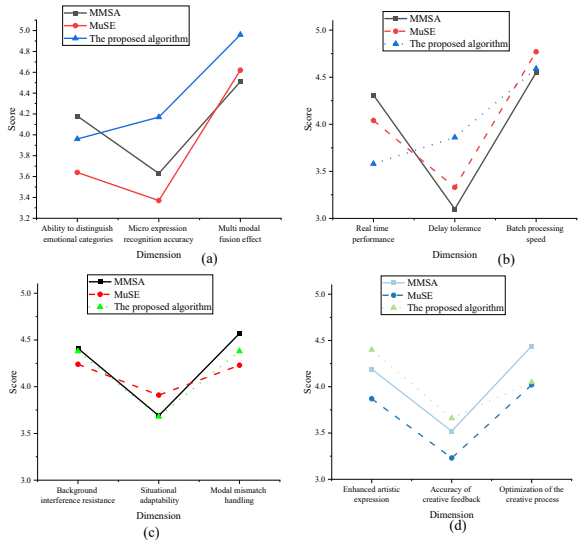

Fig. 3. Comparison results of simulation experiments: (a) Fine-grained accuracy of emotion recognition; (b) Response speed of emotion recognition; (c) Robustness of emotion recognition in complex backgrounds; (d) Effectiveness of emotion recognition in creative guidance.

In Fig. 3, in the comparison of the fine-grained accuracy of emotion recognition, the MMSA model scored 4.18, the MuSE model scored 3.64, and the optimized model scored 3.96 in terms of the ability to distinguish emotion categories. In terms of micro-expression recognition accuracy, the MMSA model scored 3.63, the MuSE model scored 3.37, and the optimized model scored 4.17. In the aspect of multimodal fusion effect, the MMSA, MuSE, and optimized models score 4.51, 4.62, and 4.96. The optimized model shows excellent performance in all dimensions of fine-grained emotion recognition, especially in micro-expression recognition accuracy and multimodal fusion effect, and the score of the optimized model is significantly higher than other models. This shows that the optimized model has higher accuracy and robustness when dealing with complex and diverse emotional data.

In the comparison of response speed, the scores of the optimized model are 3.58, 3.86, and 4.59 in different dimensions, and the MMSA model is the best in the real-time dimension, which shows that it has advantages in emotion recognition tasks that need immediate response. However, in terms of delay tolerance, the optimized model has the highest score, indicating that it can better deal with the delay problem when dealing with complex or long input data. The MuSE model has the highest score in the batch processing speed dimension, which shows that it is very efficient in processing a large amount of data and is suitable for scenes that need to process emotional data in batches. The performance of the optimized model is balanced in all dimensions, especially in delay tolerance and batch processing speed.

In the comparison of robustness in a complex background, the scores of the optimized model are 4.30, 3.68, and 4.38 in different dimensions. The MMSA model is more robust in a complex background, especially in the two dimensions of background interference resistance and modal mismatch processing, which makes it better to maintain the performance of emotion recognition in changeable and unstable

environments. However, in terms of situational adaptability, the MuSE model scored the highest, indicating that it has better adaptability in the face of complex situations. The performance of the optimized model is relatively balanced in all dimensions, especially in the processing of background interference resistance and modal mismatch, which shows that the optimized model can maintain stable emotion recognition performance in a complex background, but it is slightly inferior to MuSE in situational adaptability.

In the comparison of the effectiveness of creative guidance, the scores of the optimized model are 4.40, 3.66, and 4.05, respectively. The overall performance of the optimized model is the best, especially in the two dimensions of artistic expression enhancement and creative feedback accuracy, which shows that it has obvious advantages in improving the emotional expression of artistic works and providing creative guidance. MMSA model has the highest score in the optimization dimension of the creative process, which shows its potential in improving creative efficiency and simplifying the process. However, the performance of the MuSE model is relatively low in three dimensions, especially in the enhancement of artistic expression and the accuracy of creative feedback, which needs to be further optimized to enhance its effectiveness in guiding artistic creation.

C. Discussion

Overall, experimental results show that the optimized intelligent emotion recognition model exhibits stable performance superior to comparative models across multiple key dimensions. When compared with MMSA and MuSE, the optimized model generally demonstrates the characteristics of "comprehensive balance and slight advantage" in fine-grained emotion recognition, stability under complex backgrounds, and creative guidance-related indicators. In contrast, each comparative model possesses partial advantages. This result indicates that the adjustments made by this study to the multimodal fusion strategy and model structure do not merely pursue the extreme optimization of a single indicator. Instead, it attempts to achieve a reasonable balance between recognition accuracy, robustness, and creative usability. From the subjective rating results of simulation experiments, the optimized model shows remarkable advantages in dimensions closely related to creative practice, such as "enhanced artistic expression" and "accuracy of creative feedback". In dimensions like "real-time performance" and "batch processing efficiency", it forms a complementary pattern of "each having its own strengths" with existing models. This also suggests that in real applications, different models are not simply in a "superior-inferior substitution" relationship, but are more likely to form a model combination that can be dispatched according to scenarios. In general, these results collectively demonstrate that the optimized model, while ensuring technical performance, is closer to the actual needs of audio-visual art creation for emotion recognition systems. Thus, it lies a potential foundation for emotion recognition to transform from a "technical tool" to a "creative partner".

At an application level, the optimized intelligent emotion recognition model can serve emotional analysis at the single work level and has the potential to be embedded into the complete audio-visual art creation process. For directors, audio-visual artists, or editors, the model can act as an "emotion monitoring and feedback engine". During the creation process, it provides quantitative reference suggestions for shot rhythm, frame composition, color style, and audio-visual relationships based on changes in audience emotions. This helps creators more consciously shape the emotional trajectory of their works, rather than relying solely on experience and intuition. For cross-platform digital audio-visual content (e.g., short videos, interactive audio-visual installations, immersive performances), the model is also expected to analyze different audience groups' emotional response patterns, providing data support for content iteration and personalized recommendations. In the field of human-computer collaboration and intelligent creative systems, the optimized model can be embedded into creative assistance platforms as an "emotion perception module"; meanwhile, it can work in coordination with modules such as intelligent editing, style transfer, and generative content creation. With the output of emotion recognition, the system can automatically tend to combinations of shots, colors, or sound effects that better match the target emotional state when generating or recommending materials; thus, it can realize an "emotion-mediated" human-computer collaborative creation model. In new media environments such as virtual reality (VR) and augmented reality (AR), the model can also be used to real-time perceive users' emotional fluctuations; it can also dynamically adjust virtual scenes, interaction rhythms, or visual narrative paths to enhance immersion and participation. At the theoretical level, this study attempts to systematically link multimodal emotion recognition with the guidance of audio-visual art creation. This helps promote the transformation of "emotional expression" from artistic theory and aesthetic experience into computable and modelable forms. For one thing, by introducing evaluation dimensions related to artistic expression and creative feedback, this study provides an operable idea for "how to characterize the emotional expression of audio-visual works with computational indicators". For another, the mapping process between emotion recognition results and specific creative decisions (e.g., editing, color grading, narrative rhythm) offers new empirical support and methodological reference for the theory of emotional expression in digital media; it promises to promote interdisciplinary dialogue between affective computing, audio-visual narrative, and aesthetic research.

V. Conclusion

This research studies the application and optimization of intelligent emotion recognition technology in the guidance of video art creation. By comparing the performance of the MMSA, MuSE, and optimized models, it makes in-depth analysis and simulation experiments on the model in several key dimensions to evaluate its effectiveness and applicability in emotion recognition tasks. The experimental results show that the proposed optimized model performs well in many aspects, especially in fine-grained accuracy of emotion recognition, robustness in complex backgrounds, and effectiveness of creative guidance. In addition, the MMSA model shows its advantages in real-time and creative process optimization, and is suitable for scenes that need efficient processing. The MuSE model is excellent in situational adaptability, but there is still

room for improvement in other dimensions. Although this study has made some achievements in optimizing the emotion recognition model and applying it to the guidance of video art creation, there are still some shortcomings. Firstly, the experiments are mainly based on specific public datasets and limited emotion categories. Although they cover multiple emotion types and multimodal signals, differences still exist compared with real audio-visual art creation scenarios. Emotions in actual artistic creation often exhibit high ambiguity, hybridity, and cultural dependence. These complex emotional forms have not been fully reflected in the current data and annotation systems; this to a certain extent limits the model's generalization ability when extended to broader artistic scenarios. Secondly, although the subjective rating experiments involve evaluators with artistic and technical backgrounds, the sample size, professional structure, and cultural backgrounds still have limitations. The rating results inevitably carry certain subjective biases. The current evaluation system mostly reflects the "expert perspective" or "creator perspective" and has not fully incorporated feedback from ordinary audiences, multicultural groups, or cross-media platform users. This may affect the judgment of the model's applicability among large-scale actual user groups. Thirdly, the experimental environment of this study is relatively idealized, mainly conducted under offline or semi-offline conditions. For application scenarios requiring high real-time performance, ultra-large-scale concurrency, or operation on resource-constrained devices (e.g., mobile terminals, on-site interactive installations, etc.), the model's computational complexity and deployment costs need further evaluation and optimization. In addition, this study discusses the potential applications of the model in VR, AR, and human-computer collaboration. However, systematic field tests and user research have not been carried out in these specific scenarios.

Future research can be carried out in the following directions: 1) It is necessary to construct more diverse multimodal emotional datasets that are closer to real creation scenarios, incorporating cross-cultural and cross-media complex emotional expression samples. 2) Future work should further light-weight and structurally optimize the model to make it more suitable for real-time interaction and embedded deployment. 3) The research on the mapping between emotion recognition results and creative strategies can be deepened, exploring interpretable human-computer co-creation mechanisms. User empirical research is conducted in specific scenarios such as VR, interactive audio-visual works, and immersive performances to test the model's long-term performance and impact on user experience in real artistic systems. Through these subsequent efforts, it is expected to verify and expand the conclusions on a larger scale, enabling intelligent emotion recognition to truly become a key basic capability in digital audio-visual art practice.

REFERENCES

[1] Canal F Z, Müller T R, Matias J C, et al. A survey on facial emotion recognition techniques: A state-of-the-art literature review. Information Sciences, 2022, 582(3): 593-617.

[2] Khare S K, Blanes-Vidal V, Nadimi E S, et al. Emotion recognition and artificial intelligence: A systematic review (2014-2023) and research recommendations. Information Fusion, 2023, 2(1): 102019.

[3] Khattak A, Asghar M Z, Ali M, et al. An efficient deep learning technique for facial emotion recognition. Multimedia Tools and Applications, 2022, 81(2): 1649-1683.

[4] Mukhiddinov M, Djuraev O, Akhmedov F, et al. Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people. Sensors, 2023, 23(3): 1080.

[5] Ullah Z, Mohmand M I, Rehman S U, et al. Emotion recognition from occluded facial images using deep ensemble model. Cmc-Computers Materials & Continua, 2022, 73(3): 4465-4487.

[6] Lian H, Lu C, Li S, et al. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy, 2023, 25(10): 1440.

[7] Kopalidis T, Solachidis V, Vretos N, et al. Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets. Information, 2024, 15(3): 135.

[8] Lin W, Li C. Review of studies on emotion recognition and judgment based on physiological signals. Applied Sciences, 2023, 13(4): 2573.

[9] Dash A, Agres K. Ai-based affective music generation systems: A review of methods and challenges. ACM Computing Surveys, 2024, 56(11): 1-34.

[10] Monser M, Fadel E. A modern vision in the applications of artificial intelligence in the field of visual arts. International Journal of Multidisciplinary Studies in Art and Technology, 2023, 6(1): 73-104.

[11] Li K. Application of communication technology and neural network technology in film and television creativity and post-production. International Journal of Communication Networks and Information Security, 2024, 16(1): 228-240.

[12] Wang Y. Evolution and Deepening of Antagonistic Characters in Films: A Typological Analysis and Exploration of Empathy Construction. Journal of Research in Social Science and Humanities, 2024, 3(3): 44-50.

[13] Zhang X, Ba X, Hu F, et al. Emotioncast: An emotion-driven intelligent broadcasting system for dynamic camera switching. Sensors, 2024, 24(16): 5401.

[14] Zhang N, Pu B. Film and television animation production technology based on expression transfer and virtual digital human. Scalable Computing: Practice and Experience, 2024, 25(6): 5560–5567.

[15] Pabba C, Kumar P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Systems, 2022, 39(1): e12839.

[16] Xiao H, Li W, Zeng G, et al. On-road driver emotion recognition using facial expression. Applied Sciences, 2022, 12(2): 807.

[17] Debnath T, Reza M M, Rahman A, et al. Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity. Scientific Reports, 2022, 12(1): 6991.

[18] Zhu X, Huang Y, Wang X, et al. Emotion recognition based on brain-like multimodal hierarchical perception. Multimedia Tools and Applications, 2024, 83(18): 56039-56057.

[19] Ahmed N, Al Aghbari Z, Girija S. A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 2023, 17(3): 200171.

[20] Ngai W K, Xie H, Zou D, et al. Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources. Information Fusion, 2022, 77(11): 107-117.

[21] Yadav S P, Zaidi S, Mishra A, et al. Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Archives of Computational Methods in Engineering, 2022, 29(3): 1753-1770.

[22] Garcia-Garcia J M, Penichet V M R, Lozano M D, et al. Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions. Universal Access in the Information Society, 2022, 21(4): 809-825.

[23] Karnati M, Seal A, Bhattacharjee D, et al. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. IEEE Transactions on Instrumentation and Measurement, 2023, 72(3): 1-31.

[24] Padhmashree V, Bhattacharyya A. Human emotion recognition based on time–frequency analysis of multivariate EEG signal. Knowledge-Based Systems, 2022, 238(14): 107867.

[25] Houssein E H, Hammad A, Ali A A. Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. Neural Computing and Applications, 2022, 34(15): 12527-12557.

[26] Savchenko A V, Savchenko L V, Makarov I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Transactions on Affective Computing, 2022, 13(4): 2132-2143.

[27] Abdollahi H, Mahoor M H, Zandie R, et al. Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. IEEE Transactions on Affective Computing, 2022, 14(3): 2020-2032.

[28] Gupta S, Kumar P, Tekchandani R K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications, 2023, 82(8): 11365-11394.

[29] Khan A A, Shaikh A A, Shaikh Z A, et al. IPM-Model: AI and metaheuristic-enabled face recognition using image partial matching for multimedia forensics investigation with genetic algorithm. Multimedia Tools and Applications, 2022, 81(17): 23533-23549.