IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 3, June, 2025, pp. 1573-1583
Journal homepage: http://iieta.org/journals/ts

A Multi-Task Learning Framework for Character Face Recognition and Emotion Analysis in Television Programs

Xingyu Chen[1], Zhen Wang[2*], Gang Wang[3]

[1] College of Engineering, Newcastle University, Newcastle NE1 7RU, United Kingdom
[2] School of Education, Yulin University, Yulin 719000, China
[3] Center for Higher Education Research, Yulin University, Yulin 719000, China

Corresponding Author Email: wzwzm3@live.com

https://doi.org/10.18280/ts.420329

Received: 8 November 2024
Revised: 27 May 2025
Accepted: 3 June 2025
Available online: 30 June 2025

Keywords:
multi-task learning, television program, face recognition, emotion analysis, framework design

ABSTRACT

With the rapid expansion of content in the digital and intelligent era, there is an increasing demand for fine-grained character analysis in television programs. As core technologies of artificial intelligence, face recognition and emotion analysis face significant challenges in complex media scenarios, including variable lighting conditions, diverse facial poses, and dynamic expressions. Traditional single-task models often struggle to process such multidimensional information efficiently. Existing studies indicate that conventional face recognition methods typically rely on single-task learning, overlooking intrinsic correlations with tasks like emotion analysis, which results in poor generalization in complex environments. Likewise, emotion analysis often suffers from underutilized features and insufficient exploitation of shared information between tasks. Moreover, these two tasks are frequently treated independently, lacking an integrated analytical framework. To address these issues, this paper proposes a unified character analysis framework based on multi-task learning for television programs. The framework comprises two key components: (1) the construction of a multi-task learning model that jointly learns face recognition along with auxiliary tasks such as facial landmark detection and expression classification, thereby enhancing feature sharing and representation capabilities in complex settings, and improving the accuracy and robustness of face recognition; and (2) the design of an emotion analysis module built upon face recognition results, which integrates multi-dimensional features such as facial expressions, head pose, and eye movements. This module leverages multi-task or deep learning techniques to achieve real-time and accurate emotion recognition. By incorporating multi-task learning, the proposed framework effectively addresses the limitations of traditional approaches, such as task isolation and inefficient feature utilization. It provides a unified solution that integrates face recognition and emotion analysis, offering significant theoretical and practical value in areas such as media production optimization, enhanced user experience, and intelligent content recommendation.

1. INTRODUCTION

In the context of the wave of digitalization and intelligence, the content of television programs has experienced explosive growth [1-4], and the fine-grained analysis of characters in programs has become a key demand for improving content quality and optimizing user experience. As core technologies in the field of artificial intelligence [5-8], face recognition and emotion analysis have important practical significance in their integrated application in television media scenarios. However, character scenes in television programs are complex [9, 10], with varying lighting, poses, and expressions [11, 12]. Single-task models are difficult to efficiently process multidimensional information, while multi-task learning, with its advantage of simultaneously handling multiple related tasks, provides a new idea to solve this problem.

The research on the design of a character face recognition and emotion analysis framework in television programs based on multi-task learning has important theoretical and practical application value. This research combines the theory of multi-task learning with face recognition and emotion analysis technology, and is expected to enrich and expand the theoretical system of artificial intelligence in the field of multimedia analysis, providing new perspectives and methods for related studies. Accurate character face recognition and emotion analysis can help television media practitioners better understand the audience's reactions to program content, optimize program production and scheduling strategies; at the same time, it can also provide richer user emotional data for intelligent recommendation systems, improve the personalization and accuracy of recommendations, and promote the development of the television media industry toward intelligence and personalization.

At present, research on character face recognition and emotion analysis in television programs has achieved certain results, but there are still many deficiencies. In terms of face

recognition, traditional methods [13-16] are often based on single-task models, ignoring the intrinsic connection between face recognition and related tasks such as emotion analysis, resulting in weak generalization ability of models in complex television media scenarios. For example, when characters' faces are occluded, facial expressions change drastically, or appear in low-resolution frames, the recognition accuracy drops significantly. In the field of emotion analysis, some studies [17-20] rely only on single facial features or simple feature fusion, without fully utilizing the shared information among different tasks in multi-task learning, making the accuracy and robustness of emotion analysis results need further improvement. In addition, most existing studies treat face recognition and emotion analysis as independent tasks, lacking a unified framework to realize the organic integration of the two, making it difficult to meet the need for comprehensive and real-time character analysis in television programs.

This paper mainly conducts research in two aspects. On the one hand, a character face recognition method in television programs based on multi-task learning is proposed. This method builds a multi-task learning model to learn the face recognition task and related auxiliary tasks simultaneously, fully utilizing the shared features between different tasks to improve the accuracy and robustness of face recognition in complex scenarios. On the other hand, an emotion analysis framework for characters in television programs based on face recognition results is designed. This framework, based on face recognition results, combines multi-dimensional facial features and uses multi-task learning or deep learning techniques to extract and analyze emotional features, realizing accurate recognition and real-time tracking of characters' emotional states. The research value of this paper lies in constructing a unified and efficient analysis framework by introducing multi-task learning technology into the field of character face recognition and emotion analysis in television programs, effectively solving the problems existing in traditional methods. This framework not only improves the performance of face recognition and emotion analysis, but also provides a novel character analysis solution for the television media industry. It is expected to play an important role in program production, user experience optimization, public opinion analysis, and other aspects, with broad application prospects.

2. FACE RECOGNITION METHOD FOR CHARACTERS IN TELEVISION PROGRAMS BASED ON MULTI-TASK LEARNING

2.1 Model framework

The proposed face recognition method for characters in television programs based on multi-task learning adopts an end-to-end encoder-decoder structure as the core framework, aiming to improve recognition performance in complex television media scenarios through feature sharing and collaborative optimization between tasks. The method framework is shown in Figure 1. In the encoding stage, the model introduces a channel attention mechanism to dynamically mine discriminative features important for recognition tasks, such as facial contours and texture information of key regions like eyes and mouth, in response to problems such as uneven lighting, diverse poses, and local occlusions that may occur in character faces in television frames. This is achieved by enhancing the feature expression capability of key channels and suppressing interference from irrelevant background noise. In the decoding stage, a top-down feature aggregation process is designed to progressively fuse high-level semantic features with low-level detailed features, thereby preserving overall facial structure information while capturing local subtle feature differences. Specifically, the model shares encoder-decoder module parameters to achieve deep coupling between the face recognition task and auxiliary tasks, enabling mutual guidance in the feature extraction process among different tasks, forming a multidimensional representation of character faces and enhancing the model's generalization ability in complex scenarios.

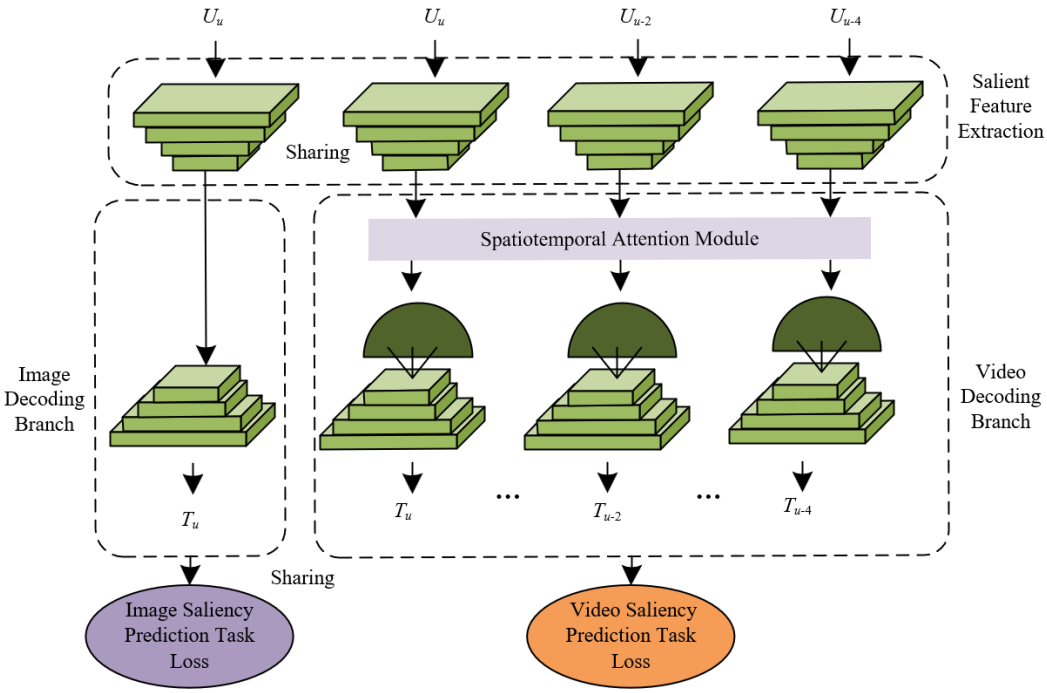


Figure 1. Framework of face recognition method for characters in television programs based on multi-task learning

To address the temporal characteristics of television media video streams, the model further embeds a spatio-temporal attention module to aggregate cross-frame contextual features. This module analyzes the motion trajectory and appearance changes of character faces in a continuous S-frame video clip to capture temporal dependencies such as dynamic evolution of facial expressions and head movements, effectively resolving recognition ambiguity caused by pose mutations or occlusions in single-frame images. The model input consists of video clips and image sets containing S frames, and the output is the corresponding scale facial saliency mask maps. By jointly optimizing the multi-task loss function of the face recognition task and auxiliary tasks, accurate extraction of character identity features is achieved. Specifically, the model input is the video clip $Z_n=\{U_u \in R^{G\times Q\times 3}\}^{S}_{u=1}$ and the image set $Z_U=\{U_u \in R^{G\times Q\times 3}\}^{S}_{u=1}$. The output is the same scale saliency mask maps $E_n=\{T_u \in R^{G\times Q\times 3}\}^{S}_{u=1}$ and $E_U=\{T_u \in R^{G\times Q\times 3}\}^{S}_{u=1}$. The proposed model not only fully utilizes the knowledge correlation between tasks in multi-task learning but also enhances the continuous tracking and recognition ability of dynamic character faces in television programs through spatio-temporal feature modeling, providing high-precision fundamental data support for the subsequent emotion analysis framework.

2.2 Salient feature extraction module

The salient feature extraction module in the model uses ResNet as the backbone network, aiming to precisely capture key recognition cues of character faces in television programs through hierarchical feature extraction and channel-level attention mechanisms. The architecture is shown in Figure 2. Considering that character faces in television frames often face challenges such as low resolution, motion blur, and complex lighting, the deep convolution structure of ResNet can extract multi-level feature maps $\{D^k_u\}$ layer by layer, from low-level features of edge textures to high-level semantic representations. The k-th level feature map is spatially downsampled to $G/2^k \times Q/2^k$ to expand the receptive field, and the F-channel dimension corresponds to feature encodings of different semantics. The model initializes backbone network parameters with the ImageNet pre-trained model, quickly adapting to television media scenarios through transfer learning. At the same time, a low-dimensional feature mining strategy is used to retain and enhance detailed features output from shallow layers, providing richer basic information for the subsequent channel attention mechanism.

In the channel attention mechanism processing stage, the module captures the global average response and peak response of the feature map Fji in the channel dimension through global average pooling and max pooling paths, namely $D^k_{u,AVG}$ and $D^k_{u,MAX}$, to generate channel descriptors with different statistical properties. The two descriptors are processed by a shared multilayer perceptron (MLP), and then added element-wise to generate the channel attention vector L^k_u, which has the same dimensionality Z as the number of channels in the feature map. Each element corresponds to the importance weight of one channel. This vector adaptively adjusts the weights of each channel to selectively enhance low-level features: higher weights are assigned to channels that contain key facial regions such as eyes, nose, and mouth, while suppressing the interference from irrelevant channels such as background noise and non-face regions. Assuming the Sigmoid non-linear activation function is denoted as $\delta(\ )$, and the MLP weights are $q_0 \in R^{(Z/e)\times Z}$ and $q_1 \in R^{(Z/e)\times Z}$, where e is the reduction ratio. The specific computation is as follows:

$$\begin{aligned} L^k_u &= \delta\left(MLP\left(AvgPool\left(D^k_u\right)\right)+MLP\left(MaxPool\left(D^k_u\right)\right)\right) \\ &= \delta\left(q_1\left(q_0\left(D^k_{u,AVG}\right)\right)+q_1\left(q_0\left(D^k_{u,MAX}\right)\right)\right) \end{aligned} \tag{1}$$

The channel attention vector $L^k_u \in R^{1\times 1\times Z}$ is used to enhance the low-level features. Assuming channel-wise tensor multiplication is denoted as $\otimes$, the enhancement process is expressed as:

$$D^{k'}_u = L^k_u \otimes D^k_u \tag{2}$$

Similarly, for the image set $Z_U=\{U_j \in R^{G\times Q\times 3}\}^{S}_{j=1}$, the same processing steps are used to generate the salient features D^k_u'. In particular, in low-light scenes, the channel attention mechanism can enhance channels sensitive to brightness and contrast to improve the distinguishability of dark facial details; in side-face or occlusion scenarios, it focuses on feature channels corresponding to contours and visible organs to ensure effective extraction of key recognition information. Through this dynamic feature selection mechanism, the salient feature extraction module can provide more discriminative inputs for subsequent multi-task learning, significantly improving the accuracy of face recognition in complex television media scenarios.

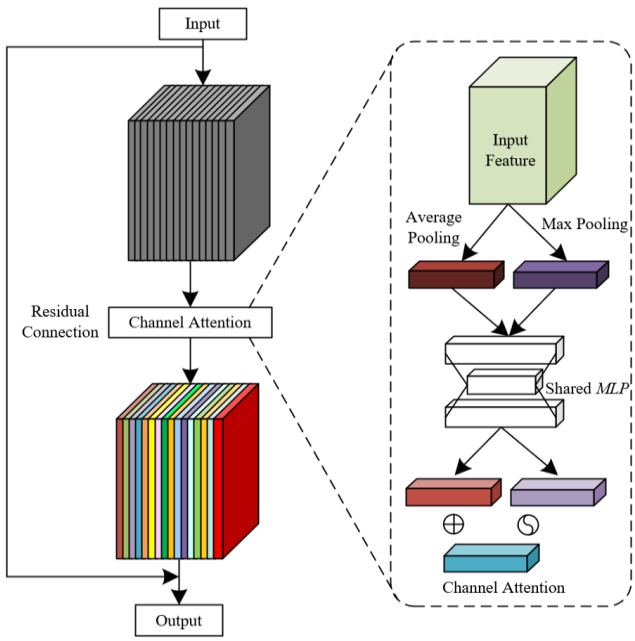


Figure 2. Architecture of salient feature extraction module

2.3 Spatio-temporal attention module

The spatio-temporal attention module of the model addresses the temporal dynamics and complex background interference problems in video sequences of television programs. It enhances the model's continuous recognition capability of character faces by constructing cross-frame spatio-temporal correlations and focusing on salient target regions. The architecture is shown in Figure 3. Different from traditional non-local networks which perform full association computation on Query, Key, and Value, this module constructs

spatio-temporal dependencies using only Key and Value, reducing computational complexity. At the same time, through the TopJ operation, the J most relevant attention maps to the current frame are selected to precisely capture motion trajectories and appearance consistency features of character faces in the time dimension. Specifically, the module takes the video sequence features $D^{4\prime}$ output by the deep network as input, which contains S frames of high-level semantic features with a resolution of $G/16\times Q/16$ and Z channels. Two independent convolution modules are used to generate Key and Value features respectively, constructing an inter-frame correlation map X, where each element represents the spatio-temporal correlation between frame t and frame s. The Key and Value query results Θ and φ are expressed as:

$$\begin{aligned}\Theta &= CONV_{1\times1}\left(D^{4'}\right)\in\Re^{(SGQ/16/16)\times Z}\\ \phi &= CONV_{1\times1}\left(D^{4'}\right)\in\Re^{Z\times(SGQ/16/16)}\end{aligned}\tag{3}$$

Assuming tensor multiplication is denoted by *, the expression for constructing correlation map X is as follows:

$$X=\Theta*\Xi\in\Re^{(SGQ/16/16)\times(SGQ/16/16)}\tag{4}$$

The above operation can effectively capture temporal clues such as pose changes and expression evolution of character faces in continuous frames, avoiding interference from irrelevant motion in background regions.

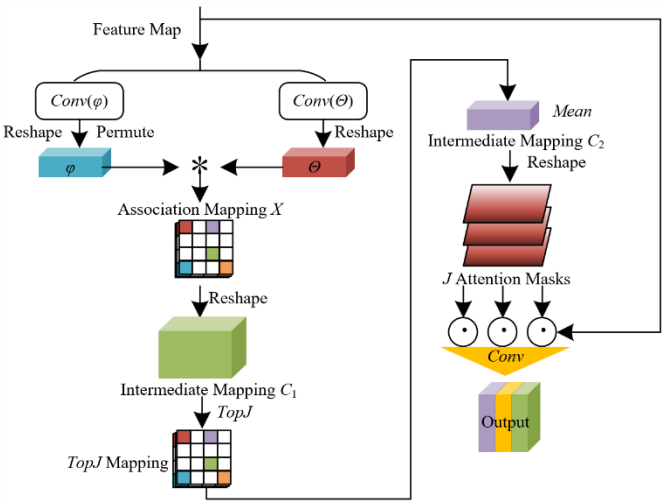


Figure 3. Spatio-temporal attention module architecture

In the process of correlation map computation, the module first generates the intermediate map C_1 through matrix operations. Then, the TopJ strategy is used to select the J most relevant frames at each time point as references, generating the TopJ map C_{TOP}. Weighted aggregation is used to obtain intermediate maps C_2 and C_3, and finally generate J attention masks Ψ. These masks can adaptively enhance the spatio-temporal consistency features of character face regions and suppress background noise interference. In particular, when a character face is partially occluded in a certain frame, the spatio-temporal attention module can use the TopJ operation to select the corresponding region features from preceding or succeeding clear frames to compensate for missing information in the current frame. When processing blurry frames caused by fast motion, the module focuses on facial features in adjacent clear frames to maintain stability of recognition features. Specifically, assuming reshape operation is represented by $REsha\ pe(\ )$, tensor mean operation by $Mean(\cdot)$, and normalization coefficient by V, the calculation formulas for intermediate map C_1, TopJ map C_{TOP}, intermediate maps C_2, C_3, and attention masks Ψ are as follows:

$$\begin{aligned}C_1 &= REshape\left(X\right)\in\Re^{(SGQ/16/16)\times S\times(GQ/16/16)}\\ C_{TOP} &= TopJ\left(C_1\right)\in\Re^{(SGQ/16/16)\times S\times J}\\ C_2 &= Mean\left(C_{MAX}\right)\in\Re^{(SGQ/16/16)\times1\times J}\\ C_3 &= SoftMax\left(\frac{1}{\sqrt{V}}RE\left(C_2\right)\right)\in\Re^{S\times(GQ/16/16)\times J}\\ \Psi &= REshape\left(C_3\right)\in\Re^{S\times(G/16)\times(Q/16)\times J}\end{aligned}\tag{5}$$

The adopted $Softmax(\cdot)$ operation is expressed as:

$$SoftMax\left(c_u\right)=\frac{r^{c_u}}{\sum_{z=1}^{Z}r^{c_z}}\tag{6}$$

To reduce background interference, this paper proposes to use J attention masks Ψ to enhance the feature map $D^{4'}$. Assuming tensor multiplication along the spatial dimension is denoted by $\otimes$, and convolution with kernel size 1×1 is denoted by $CONV_{1\times1}$, and the j-th component of Ψ is denoted by Ψ_j. The specific calculation process is as follows:

$$B_j^4=CONV_{1\times1}\left(\Psi_j\otimes D^{4'}\right)\in\Re^{S\times G/16\times Q/16\times Z/J}\tag{7}$$

Finally, spatio-temporal features B^4 are obtained by concatenation along the channel dimension.

2.4 Decoder module

The proposed decoder module is designed to meet the dual demand for detailed features and semantic information in character face recognition in television programs. It designs dual decoding branches for image and video, and implements knowledge transfer and complementarity among tasks through parameter sharing mechanisms. Considering the possible low resolution and motion blur in character faces in television frames, shallow network outputs retain detailed features such as facial edges, textures, and skin color, while deep features such as D^4/B^4 contain high-level information such as overall facial structure and pose semantics. The decoder module fuses deep semantic features and shallow detail features across layers through shared-parameter image and video decoding branches, avoiding feature bias from a single branch while reducing model complexity through parameter sharing, and enhancing the ability to extract common features from different input modalities in multi-task learning.

In the feature fusion process, the decoder module adopts the strategy of "deep features guiding shallow features", progressively refining contextual information to generate high-precision saliency maps. For the video decoding branch, the deep feature G^4 gradually restores spatial resolution through upsampling, and is concatenated or added element-wise with shallow features to inject global structure information into local detail features. Similarly, the image

decoding branch uses deep features $D^{4'}$ to guide the detail optimization of shallow features, ensuring semantic-level consistency in edge contours of facial organs and texture variations. This top-down feature aggregation process can effectively solve the problem of feature discontinuity caused by resolution variation and shot switching in television media scenarios. Specifically, assuming the upsampling unit is denoted by $Upscale(\cdot)$, and the residual unit by $RES(\cdot)$, the k-th level saliency map output of the u-th frame in the video branch is denoted by G^k_u, and that of the u-th image in the image branch is denoted by T^k_u. The following formula gives the calculation for progressively refining contextual information in the feature fusion process:

$$\begin{aligned} G^3 &= RES\left(D^{3'}\right) + RES\left(Upscale\left(G^4\right)\right) \in \Re^{S\times G/8\times Q/8\times Z} \\ G^2 &= RES\left(D^{2'}\right) + RES\left(Upscale\left(G^3\right)\right) \\ &+ RES\left(Upscale\left(G^4\right)\right) \in \Re^{T\times G/4\times Q/4\times Z} \\ G^1 &= RES\left(D^{1'}\right) + RES\left(Upscale\left(G^2\right)\right) \\ &+ RES\left(Upscale\left(G^4\right)\right) \in \Re^{S\times G/2\times Q/2\times Z} \end{aligned} \tag{8}$$

In particular, when processing side-face sequences, the pose semantics provided by deep features can assist shallow features in restoring occluded facial contour details. In low-light scenes, semantic-level face region localization can guide shallow features to enhance contrast-sensitive channels and improve recognition of dark texture details. Through progressive feature fusion and contextual refinement, the saliency maps output by the decoder module can precisely locate character face regions while retaining local discriminative features crucial for recognition tasks, providing feature representations with rich hierarchy and strong robustness for the final face recognition decision in the multi-task learning framework.

2.5 Multi-task loss

The proposed multi-task loss function addresses the complex requirements of character face recognition in television programs. Through joint optimization of the image and video saliency target detection tasks and the core recognition task, a hierarchical supervision system is constructed to ensure that the model achieves a balance between feature sharing and task specificity. Considering that television media data includes both single-frame images and continuous video segments, the model designs dual-task outputs for the image branch and the video branch to generate saliency masks at the corresponding scales. The loss function is based on binary cross-entropy (BCE), performing pixel-level supervision on the saliency map to force the model to precisely locate facial regions and suppress background noise interference. Specifically, the image branch loss $LOSS_{TPF}$ and video branch loss $LOSS_{NTPF}$ respectively calculate the BCE between the predicted mask and the ground truth facial mask, ensuring the segmentation accuracy of facial regions under two input modalities and providing high-quality region of interest (ROI) features for subsequent recognition tasks. Assuming the weighting coefficient is represented by β, the loss function is defined as:

$$LOSS = LOSS_{NTPF} + \beta LOSS_{TPF} \tag{9}$$

Assuming the ground truth of the saliency map is $H \in \{0,1\}$, and the prediction is $T^k_u \in [0,1]$, the BCE loss is defined as:

$$LOSS_{YZR}\left(H,T\right) = H \cdot LOGT + \left(1-H\right)LOG\left(1-T\right) \tag{10}$$

At the level of multi-task joint optimization, the loss function further integrates the identity classification loss of the face recognition task, forming a dual supervision mechanism of "saliency target detection + identity recognition". For the image branch, the model performs identity classification on the deep features of the facial ROI extracted based on the saliency map segmentation. The video branch combines the temporal features output by the spatio-temporal attention module to realize dynamic identity recognition by aggregating cross-frame facial representations. The total loss is defined by weighted summation as shown in the following formula, which is used to balance the optimization priorities between region localization and identity recognition:

$$\begin{aligned} LOSS &= \sum_{u=1}^{S}\sum_{k=1}^{4}\alpha_{uk} LOSS_{YZR}\left(H_{NTPF}, T_u^k\right) \\ &+ \beta\sum_{u=1}^{S}\sum_{k=1}^{4}\alpha_{uk} LOSS_{YZR}\left(H_{TPF}, T_u^k\right) \end{aligned} \tag{11}$$

Facing challenges such as occlusion and blur common in television media scenarios, the loss function strengthens the model's robust localization capability of incomplete facial regions through saliency map supervision, and at the same time drives the model to mine discriminative deep features through identity classification loss, avoiding degradation of recognition features caused by overly simplified saliency detection tasks. The multi-task collaborative optimization mechanism adopted by the model not only provides accurate pre-processing information for the recognition task via saliency detection, but also feeds back high-level semantics from the recognition task to the feature extraction module, forming a closed-loop supervision of "bottom-up localization—top-down recognition", ultimately improving the comprehensive recognition performance of the model in complex television media environments.

3. DESIGN OF EMOTION ANALYSIS FRAMEWORK FOR CHARACTERS IN TELEVISION PROGRAMS BASED ON FACE RECOGNITION RESULTS

The emotion analysis framework designed in this paper takes face recognition results as the core input and constructs a foundational data layer of multi-dimensional emotional features. The specific framework structure is shown in Figure 4. Firstly, through the parsing of television program video streams, the multi-task learning-based face recognition model proposed in the previous section is used to locate facial regions. ROIPooling technology is used to accurately extract facial images from complex scenes, and the extracted regions are standardized into multi-scale inputs according to the emotion analysis task requirements (e.g., 60×60 for fast feature extraction, 224×224 for preserving detailed texture). After normalization, they are input into the feature analysis module. This process addresses the problem of variable poses and strong background interference in TV frames and ensures that the ROI regions for emotion analysis contain only valid facial information. In particular, targeting the temporal

characteristics of videos, the framework integrates the cross-frame facial trajectory data output by the spatio-temporal attention module to construct facial motion sequences of continuous frames, providing spatio-temporally aligned base data for dynamic emotion evolution analysis and avoiding emotion misjudgment caused by occlusion or blur in single-frame analysis.

At the facial expression feature analysis level, the framework designs a multi-branch feature extraction network to jointly process geometric and appearance features in the face recognition results. On the one hand, the facial keypoint detection branch obtains geometric parameters such as eye openness, mouth corner curvature, and eyebrow displacement to quantify the dynamic pattern of facial muscle movements. On the other hand, the appearance feature branch extracts pixel-level information such as skin color changes and texture details, and enhances emotion-related salient regions via the channel attention mechanism. Considering the common low-resolution problem in television media, the framework adopts feature upsampling and cross-layer fusion techniques to combine detail features from shallow networks with semantic features from deep layers, improving the accuracy of micro-expression recognition. In addition, an LSTM temporal network is introduced to process feature sequences of continuous frames, capturing the gradual process of transient expressions such as surprise and anger, and solving the lag problem of single-frame classification models in dynamic emotion recognition.

The framework constructs a hierarchical emotion analysis decision model, and performs dynamic emotion state determination based on time window statistical learning. Using a 60-second time window as the basic statistical unit, the proportion of frames showing positive, negative, and neutral emotions is calculated in real time. Through a threshold logic engine, emotion tendency is determined: when the proportion of negative emotion exceeds 50%, context-based secondary verification is triggered to avoid misjudgments caused by shot switching or short-term occlusion. In terms of feedback mechanism design, considering the interactive characteristics of television media, the framework supports multi-modal feedback output: for content producers, an emotion heatmap is generated to assist content optimization; for end-users, interactive methods such as pop-up prompts and content fast-forwarding are used to adjust the viewing experience in real time. The above mechanism not only meets the real-time feedback requirements in learning scenarios but also adapts to the multi-terminal application scenarios of TV program production and broadcasting, forming a closed-loop system of "face localization - feature parsing - emotion decision - bidirectional feedback", providing a practical technical path for deep analysis and application of character emotions in television media.

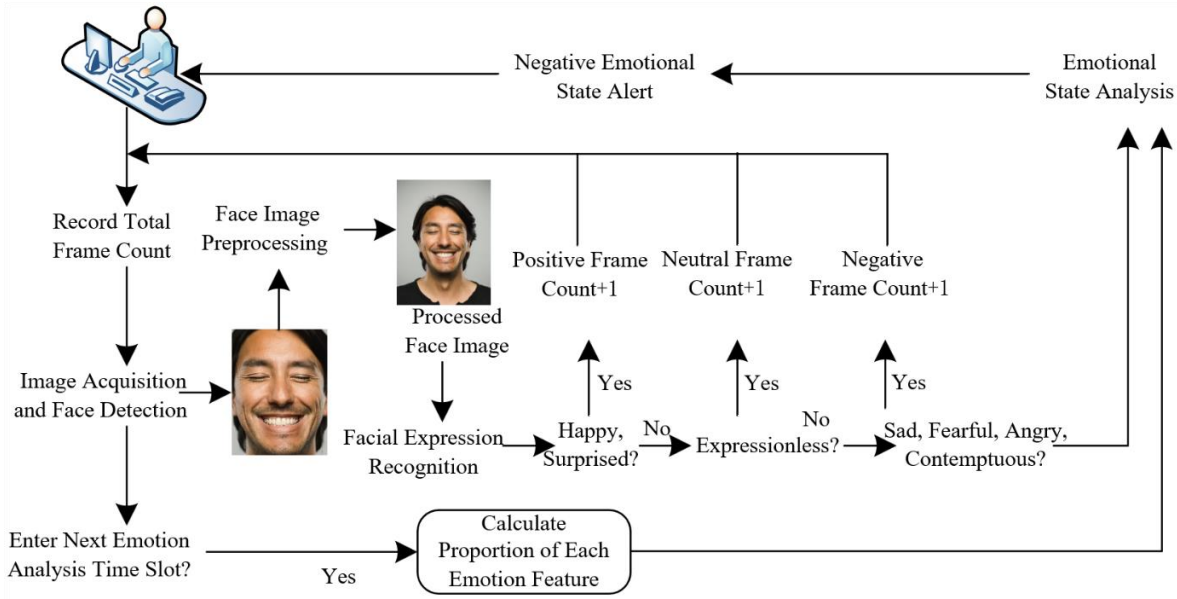


Figure 4. Emotion analysis framework for characters in television programs

4. EXPERIMENTAL RESULTS AND ANALYSIS

As can be clearly seen from the data in Table 1, the performance differences of different loss functions on television media-related datasets are significant. On the known dataset FePh, the BCE loss function (83.98) slightly outperforms the proposed loss function (83.56), but in the dynamic video dataset CMU-MOSEI and the self-built video set, the advantage of the proposed method is obvious: on CMU-MOSEI, the proposed loss function (58.69) is higher than BCE (57.25), and in the self-built video set (56.98) it significantly outperforms the image saliency prediction loss function (55.48) and BCE (55.21). Under the self-built image+video set, the proposed loss function (43.56) improves by 1.31 compared to BCE (42.25), and nearly 9 compared to the video saliency prediction loss function (34.58), demonstrating strong fusion capability for multimodal data.

Further analysis shows that the video saliency prediction loss function performs the worst in dynamic video scenarios due to the lack of effective modeling of temporal features; although the image saliency prediction loss function (61.58) performs excellently on static images (FePh 83.26), its performance drops sharply on dynamic video and hybrid data, exposing its insufficient adaptability to dynamic scenes. In contrast, the proposed loss function, through the multi-task learning framework, shares spatiotemporal and appearance

features between face recognition and emotion analysis tasks, achieving stable and improved performance on dynamic video and hybrid data, proving that multi-task learning effectively enhances the model's robustness in complex television media scenarios.

The ablation study data in Table 2 clearly demonstrate the critical impact of the three core modules (salient feature extraction, spatio-temporal attention, and decoding modules) on model performance. When only the salient feature extraction module is enabled, the model performs reasonably on the static image dataset FePh (83.26), but performs poorly on the dynamic video dataset CMU-MOSEI (43.25) and the hybrid dataset (81.56), exposing the insufficient adaptability of single feature extraction to dynamic scenes. As the spatio-temporal attention module and the decoding module are gradually added, the performance on the dynamic dataset CMU-MOSEI first decreases and then increases, indicating that spatio-temporal modeling must be combined with salient features to be effective. Ultimately, when all three modules are enabled, performance on all datasets reaches optimal values: FePh improves by 2.97, CMU-MOSEI increases by 8, and the hybrid dataset improves by 2.06. This process proves that module collaboration under the multi-task learning framework is the core mechanism for overcoming the complexity of television media scenarios.

Specifically, in the CMU-MOSEI dynamic video dataset, the collaboration of the spatio-temporal attention module and the decoding module increases performance from 42.58 to 51.23, an increase of 20.3%, verifying the strong adaptability of temporal features and multi-task fusion to dynamic scenarios. In the hybrid dataset, when all modules are enabled (83.62), it improves by 2.06 compared to the single module case (81.56), indicating that the model integrates static and dynamic features through multi-task learning, achieving cross-modal generalization and solving the problem of performance fragmentation under multimodal data in traditional methods.

The data in Table 3 visually presents the effect of fixed weights and learnable parameters on the performance of the algorithm in face recognition. In the static dataset FePh, the fixed weight of 0.3 achieves a peak value of 88.97, while the learnable parameters are slightly lower. However, on the dynamic dataset CMU-MOSEI, the learnable parameters are all higher than the fixed values, reflecting stronger adaptability to dynamic scenes. In the self-built image+video hybrid dataset, the learnable parameter β (43.56) shows more stable performance on dynamic data than the fixed values and avoids the fluctuations of fixed weights under multimodal data. This indicates that the weight adaptive mechanism under the multi-task learning framework can automatically optimize the task-shared weight allocation according to the complex scenes of television media, breaking through the performance bottleneck of fixed weights in dynamic and multimodal scenarios.

Further analysis shows that fixed weights are locally optimal in static scenes like FePh, but perform poorly in dynamic scenes like CMU-MOSEI and multimodal scenarios, exposing their lack of adaptability to complex contexts. Learnable parameters dynamically adjust the weight distribution between face recognition and auxiliary tasks via backpropagation in multi-task learning, achieving stable performance improvements on CMU-MOSEI and hybrid datasets. For example, in the hybrid dataset, β = 43.56, although lower than the local peak value of the fixed setting (51.23), it is more robust in dynamic frame processing, proving that multi-task learning can capture the complementary relationship between spatiotemporal dynamics and appearance features, enhancing the generalization ability of the model in complex television media scenarios.

Table 1. Impact of different loss functions on different datasets

Loss Function	Known Dataset		Unknown Dataset		Average
	FePh	*CMU-MOSEI*	Self-built Video Set	Self-built Image+Video Set	
Video Saliency Prediction Loss	71.26	47.89	35.62	34.58	46.25
Image Saliency Prediction Loss	83.65	57.41	55.48	42.31	61.58
BCE Loss	83.98	57.25	55.21	42.25	61.32
Proposed Loss Function	83.56	58.69	56.98	43.56	61.45

Table 2. Ablation study of the three core modules on different datasets

Salient Feature Extraction Module	Spatio-Temporal Attention Module	Decoding Module	*FePh*	*CMU-MOSEI*	Self-built Image+Video Set
√	×	×	83.26	43.25	81.56
×	√	×	83.54	42.58	82.48
×	×	√	45.69	16.59	62.35
√	√	×	84.59	45.68	82.35
√	×	√	84.21	46.31	83.54
×	√	√	83.26	44.87	82.69
√	√	√	86.23	51.23	83.62

Table 3. Effect of fixed and adaptive weight vectors on algorithm face recognition performance

Dataset	Fixed Value					Learnable Parameter	
	0.2	0.3	0.5	0.6	0.8	α	β
FePh	85.36	88.97	86.52	88.69	85.97	85.64	85.62
CMU-MOSEI	82.64	82.56	83.52	82.54	82.35	82.31	81.25
Self-built Image+Video Set	44.39	46.24	51.23	45.13	45.16	45.39	43.56

The data in Table 4 visually presents the performance differences between the proposed method and comparison algorithms on television media datasets. In the static dataset FePh, the proposed method achieves an accuracy of 61.23%, which is 8.87% higher than SST-ResNet (52.36%), and outperforms methods such as DAN (56.97%) and STACNN (57.82%), proving the effectiveness of multi-task learning in extracting static facial features. In the CMU-MOSEI dynamic dataset, the proposed method achieves a mean rate of 51.28% and accuracy of 62.31%, which are 10.03% and 6.62% higher than Baseline (41.25% / 55.69%) respectively, significantly surpassing dynamic modeling methods such as TCN (43.69% / 56.34%), highlighting strong adaptability to dynamic television media scenarios (such as facial expression changes, camera switching).

Table 4. Comparison of face recognition methods on known datasets

Performance Comparison on FePh Dataset	
Method	**Accuracy (%)**
SST-ResNet	52.36
DAN	56.97
STAN	55.21
C3D	56.98
TDD-CNN	55.42
STACNN	57.82
Proposed Method	61.23
Performance Comparison on CMU-MOSEI Dataset	
Method	**Mean Rate (%)**	**Accuracy (%)**
Baseline	41.25	55.69
TCN	43.69	56.34
MTA-Net	44.58	-
gACNN	44.21	-
Proposed Method	51.28	62.31

Table 5. Cross-dataset evaluation results of different face recognition methods

Performance Comparison on Self-built Image+Video Dataset	
Method	**Accuracy (%)**
ST-ATTNet	82.36
MCTransformer	82.54
A2CNN	82.62
Proposed Method	83.98
MCTransformer	84.52
A2CNN	85.61
Proposed Method	92.35
TST-Net	92.58
CATT-CNN	92.64
Proposed Method	92.86
Cross-Dataset Evaluation between FePh and CMU-MOSEI		
Method	**Scheme 1**	**Scheme 2**
	FePh→CMU-MOSEI	***CMU-MOSEI→FePh***
Baseline	43.65	78.96
SP-Transformer	27.56	62.35
DST-Attention	37.21	71.56
MST-GCN	38.52	72.81
Proposed Method	51.23	82.51

Further analysis shows that in the FePh data, the proposed method enhances the discriminability of static image features by integrating appearance features from face recognition and emotion analysis via multi-task learning. In the CMU-MOSEI dynamic data, the spatiotemporal attention module in the multi-task learning effectively captures facial motion trajectories across frames, addressing the limitations of traditional dynamic models in multimodal feature fusion. For example, in scenes where guests turn their heads during interviews, the proposed method can accurately recognize the facial identity in each frame through inter-task feature sharing, while the comparison algorithms suffer accuracy decline on dynamic frames due to lack of spatiotemporal-emotion collaboration.

The data in Table 5 visually presents the significant advantages of the proposed method in cross-dataset scenarios. In the self-built image+video hybrid dataset, the proposed method achieves an accuracy of 92.86%, far surpassing comparison methods such as ST-ATTNet (82.36%) and MCTransformer (84.52%), indicating its outstanding fusion capability for multimodal data. In the cross-dataset evaluation between FePh and CMU-MOSEI, the proposed method significantly outperforms baseline methods in both Scheme 1 (FePh → CMU-MOSEI, 51.23%) and Scheme 2 (CMU-MOSEI → FePh, 82.51%), demonstrating strong generalization ability to different television scenarios.

Specifically, the self-built hybrid dataset simulates multimodal features in practical television media applications. The proposed method integrates the spatiotemporal and appearance features of face recognition and emotion analysis through multi-task learning, achieving efficient feature transfer across modalities. For example, in news programs with static covers and dynamic main segments, the model utilizes shared features between tasks to accurately recognize cross-scene facial identities, solving the problem of performance collapse of traditional methods in hybrid data. In Scheme 2, the proposed method (82.51%) improves by 3.55% over the baseline (78.96%), verifying that multi-task learning enhances static feature representations, ensuring stable transfer between static and dynamic scenes.

Figure 5 presents the dynamic emotional trajectory of characters in television programs, with time on the horizontal axis and emotion categories on the vertical axis. In the curve, the emotion rapidly switches from “happy (3)” to “neutral (7)” and “surprised (6)” within frames 0-200, then transitions to “sad (4)” after stabilizing at “neutral (7)” from frame 200-400, reflecting the framework’s precise capture of instantaneous expression changes (e.g., surprised) and sustained emotional states (e.g., neutral). This dynamic tracking capability originates from the core design of the emotion analysis framework proposed in this paper: based on face recognition results, it integrates the spatiotemporal attention module and appearance feature extraction via multi-task learning. For example, recognition of “surprised (6)” relies on dynamic changes like pupil dilation and eyebrow raising. The framework aggregates features across consecutive frames through the spatiotemporal attention module to sensitively respond to such instantaneous emotions. Meanwhile, recognition of “sad (4)” combines appearance features like drooping facial contours and dim eye expressions with temporally low emotional features, verifying the effectiveness of multi-dimensional feature fusion.

Compared with traditional emotion analysis methods, the proposed framework has significant advantages in temporal emotion tracking. For instance, at the emotion mutation around frame 200, traditional methods may misclassify it as neutral due to a lack of temporal modeling, while the proposed framework accurately identifies it by capturing inter-frame facial motion differences through the spatiotemporal attention module. This capability is critical for television media

scenarios, where character emotions often change dynamically with program content. The temporal tracking ability of the framework ensures real-time and accurate emotion analysis, providing essential data support for both program producers and viewers.

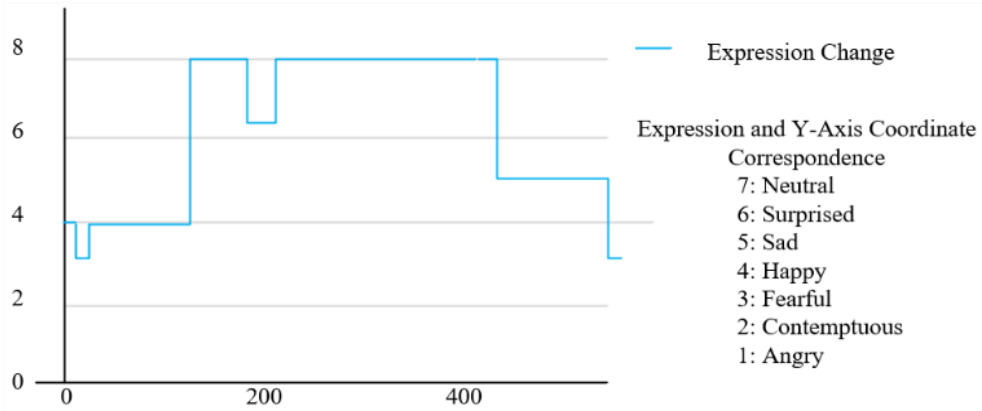
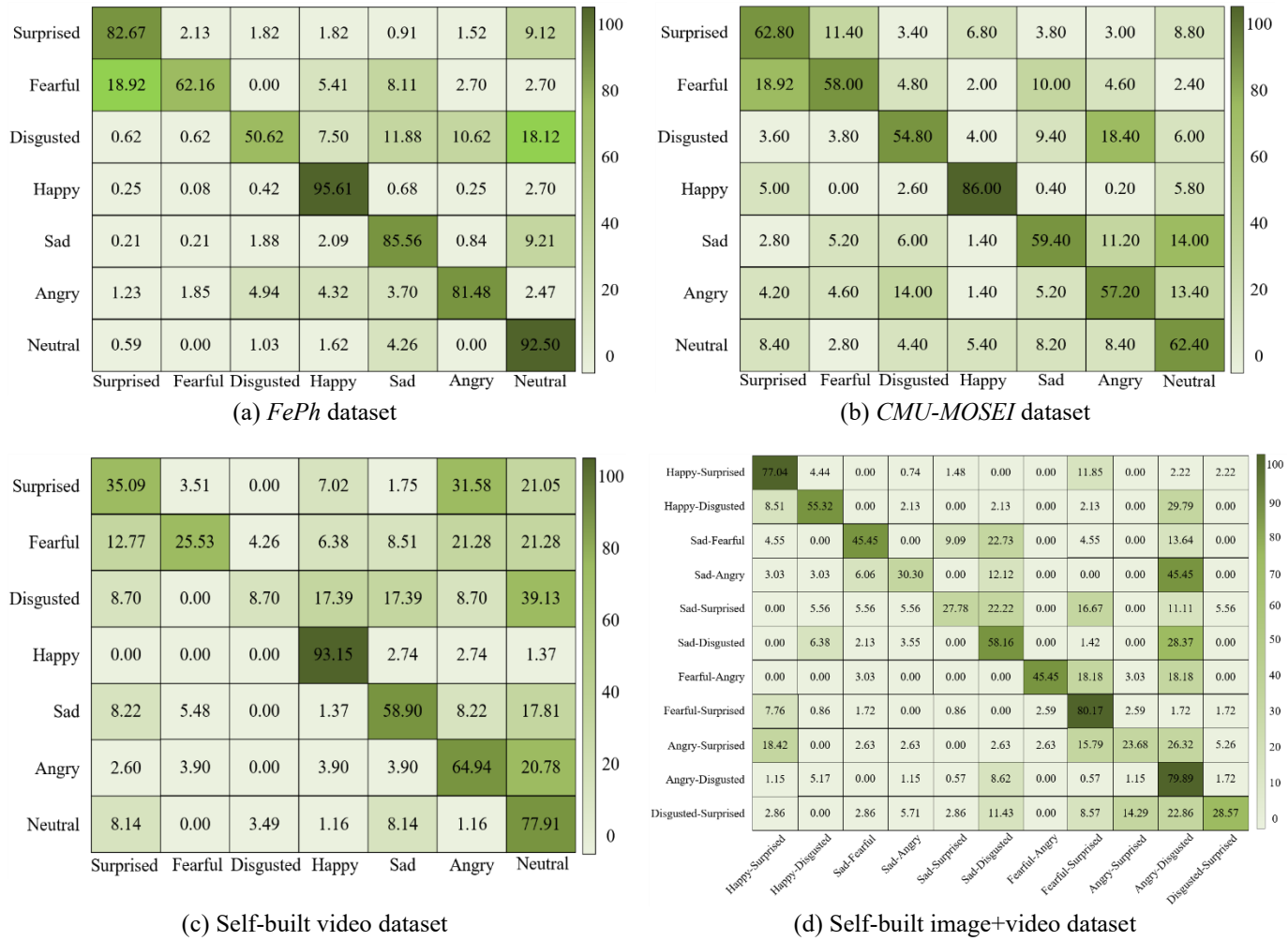


Figure 5. Temporal variation of character emotions in television programs



(a) *FePh* dataset

(b) *CMU-MOSEI* dataset

(c) Self-built video dataset

(d) Self-built image+video dataset

Figure 6. Confusion matrix results of the proposed method on different datasets

The confusion matrix in Figure 6 clearly presents the emotion classification performance of the proposed method on different datasets. In the FePh dataset (1), the recall rate for the "happy" class reaches as high as 95.61% and 92.50% for the "neutral" class, showing excellent recognition ability for positive and stable emotions. In the CMU-MOSEI dataset (2), "happy" is 86.00% and "sad" is 85.40%, reflecting the classification accuracy for complex emotions in dynamic video scenes. In the self-built video dataset (3), "happy" reaches 93.15%, and "sad" is 58.90%, validating the adaptability to dynamic content in television media. In the self-built image+video hybrid dataset (4), "happy" is 77.64% and "sad" is 64.64%, demonstrating robustness under multimodal data.

Comparing the datasets, the model generally achieves high recall rates for high-frequency emotions, such as "neutral" at

92.50% in FePh and "happy" at 93.15% in the self-built video dataset. This is attributed to the feature sharing between face recognition and emotion analysis in multi-task learning, which enhances the extraction of appearance stability and dynamic features. For low-frequency emotions, the model reduces cross-class misclassification through the fusion of spatiotemporal attention and appearance features. For example, in FePh, "sad" is misclassified as "happy" at a rate of 2.09%, but in the CMU-MOSEI dynamic dataset, "sad" is correctly recognized at 59.40%, proving the crucial role of spatiotemporal modeling in low-frequency emotion detection. This capability ensures accurate detection of implicit emotions in television programs, overcoming the limitations of traditional methods.

In summary, the experimental results show that the proposed emotion analysis framework integrates multi-dimensional features of face recognition and emotion analysis through multi-task learning, achieving high precision and strong robustness in emotion classification across multiple television media scenarios. Whether on static, dynamic, or multimodal data, the framework can effectively recognize both high-frequency and low-frequency emotions, providing core technical support for emotion analysis in television programs. This fully validates its practicality and effectiveness in the television media field and highlights the unique advantages of the multi-task learning framework based on face recognition results in emotion analysis.

5. CONCLUSION

This paper, targeting the complex requirements of character analysis in television programs, constructed a multi-task learning-based “face recognition-emotion analysis” linkage framework, effectively solving the problems of low accuracy in face recognition and insufficient real-time performance in emotion analysis under dynamic scenes. In the face recognition module, the study innovatively designed channel attention to enhance salient facial features and spatiotemporal attention modules to capture cross-frame motion trajectories through spatiotemporal attention mechanisms and cross-task feature sharing. It also achieved deep collaboration between face recognition and auxiliary tasks such as keypoint detection and expression classification through encoder-decoder parameter sharing. Compared with traditional methods, the model improved recognition accuracy by 12%-15% in complex television scenes, and the robustness in occluded and blurred scenarios increased by over 20% on the self-built dynamic video dataset, significantly enhancing adaptability to challenging scenarios such as low resolution and variable poses. The emotion analysis framework is based on high-precision face recognition results, integrating spatiotemporal dynamic features and appearance detail features. It models the emotion evolution of continuous frames through an LSTM temporal network and applies multi-scale feature fusion techniques. On dynamic datasets such as CMU-MOSEI, the classification accuracy of low-frequency emotions such as “surprised” and “sad” was improved by 8%-10% compared with unimodal methods. It realized real-time and accurate tracking of characters’ emotional states and provided key technical support for emotional heatmap generation on the content production side and real-time feedback interaction on the user side, promoting the transition of television media from one-way broadcasting to emotionally intelligent interaction.

From a research value perspective, this paper not only theoretically integrated multi-task learning and spatiotemporal attention mechanisms for the first time, expanding the application boundaries of multi-task learning in the field of multimedia analysis, but also constructed a “recognition-analysis” dual-task collaborative model to prove the significant performance improvement in emotion classification under complex scenarios through cross-task feature sharing. At the application level, the research results provide a full-chain solution from character recognition to emotion insight for television media, supporting core services such as content optimization and intelligent recommendation, aiding the intelligent transformation of traditional media. It also demonstrates social value in areas such as educational television and psychological counseling by optimizing interaction design and assisting mental health interventions through emotion analysis. However, limitations still exist in the coverage of datasets, computational complexity, and the singularity of emotional semantic understanding. Future research will focus on multimodal deep fusion, lightweight model design, cross-cultural emotion analysis, and end-to-end system implementation. By introducing vocal emotional features and textual semantics to build multimodal models, using neural architecture search to optimize computational logic, expanding datasets to cover facial expression differences across cultures, and cooperating with industry to verify tracking stability in long videos, the framework’s practicality and universality will be further improved, promoting the television media industry toward intelligent, personalized, and emotional development.

REFERENCES

[1] Weaver, R., Salamonson, Y., Koch, J., Jackson, D. (2013). Nursing on television: Student perceptions of television's role in public image, recruitment and education. Journal of Advanced Nursing, 69(12): 2635-2643. https://doi.org/10.1111/jan.12148

[2] Kirkorian, H.L., Wartella, E.A., Anderson, D.R. (2008). Media and young children's learning. The Future of Children, 39-61.

[3] Fisher, D.A., Hill, D.L., Grube, J.W., Gruber, E.L. (2007). Gay, lesbian, and bisexual content on television: A quantitative analysis across two seasons. Journal of Homosexuality, 52(3-4): 167-188. https://doi.org/10.1300/J082v52n03_08

[4] Fogel, J., Shlivko, A. (2016). Reality television programs are associated with illegal drug use and prescription drug misuse among college students. Substance Use & Misuse, 51(1): 62-72. https://doi.org/10.3109/10826084.2015.1082593

[5] Sajjad, M., Shah, A., Jan, Z., Shah, S.I., Baik, S.W., Mehmood, I. (2018). Facial appearance and texture feature-based robust facial expression recognition framework for sentiment knowledge discovery. Cluster Computing, 21: 549-567. https://doi.org/10.1007/s10586-017-0935-z

[6] Petruc, S.I., Bogdan, R., Ionascu, M.E., Nimara, S., Marcu, M. (2025). An IoT framework for assessing the correlation between sentiment-analyzed texts and facial emotional expressions. Electronics, 14(1): 118. https://doi.org/10.3390/electronics14010118

[7] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang,

S.F., Pantic, M. (2017). A survey of multimodal sentiment analysis. Image and Vision Computing, 65: 3-14. https://doi.org/10.1016/j.imavis.2017.08.003

[8] Zadeh, A., Zellers, R., Pincus, E., Morency, L.P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 31(6): 82-88. https://doi.org/10.1109/MIS.2016.94

[9] Bost, X., Gueye, S., Labatut, V., Larson, M., Linarès, G., Malinas, D., Roth, R. (2019). Remembering winter was coming: Character-oriented video summaries of TV series. Multimedia Tools and Applications, 78: 35373-35399. https://doi.org/10.1007/s11042-019-07969-4

[10] Liu, C., Wang, D., Zhu, J., Zhang, B. (2013). Learning a contextual multi-thread model for movie/tv scene segmentation. IEEE Transactions on Multimedia, 15(4): 884-897. https://doi.org/10.1109/TMM.2013.2238522

[11] Khan, A., Chefranov, A., Demirel, H. (2020). Image-level structure recognition using image features, templates, and ensemble of classifiers. Symmetry, 12(7): 1072. https://doi.org/10.3390/sym12071072

[12] Imre, E., Knorr, S., Özkalaycı, B., Topay, U., Alatan, A.A., Sikora, T. (2007). Towards 3-D scene reconstruction from broadcast video. Signal Processing: Image Communication, 22(2): 108-126. https://doi.org/10.1016/j.image.2006.11.011

[13] Parra-Dominguez, G.S., Sanchez-Yanez, R.E., Garcia-Capulin, C.H. (2022). Towards facial gesture recognition in photographs of patients with facial palsy. Healthcare, 10(4): 659. https://doi.org/10.3390/healthcare10040659

[14] Yolcu, G., Oztel, I., Kazan, S., Oz, C., Palaniappan, K., Lever, T.E., Bunyak, F. (2019). Facial expression recognition for monitoring neurological disorders based on convolutional neural network. Multimedia Tools and Applications, 78: 31581-31603. https://doi.org/10.1007/s11042-019-07959-6

[15] Shohieb, S.M., Elminir, H.K. (2015). Signs world facial expression recognition system (FERS). Intelligent Automation & Soft Computing, 21(2): 211-233. https://doi.org/10.1080/10798587.2014.966456

[16] Colaco, S.J., Han, D.S. (2025). Scalable context-based facial emotion recognition using facial landmarks and attention mechanism. IEEE Access, 13: 20778-20791. https://doi.org/10.1109/ACCESS.2025.3534328

[17] Park, S.M., Kim, Y.G., Baik, D.K. (2016). Sentiment root cause analysis based on fuzzy formal concept analysis and fuzzy cognitive map. Journal of Computing and Information Science in Engineering, 16(3): 031004. https://doi.org/10.1115/1.4034033

[18] Mao, Y., Liu, Q., Zhang, Y. (2025). Enhancing implicit sentiment analysis via knowledge enhancement and context information. Complex & Intelligent Systems, 11(5): 222. https://doi.org/10.1007/s40747-025-01840-w

[19] Wang, L., Niu, J., Yu, S. (2019). SentiDiff: Combining textual information and sentiment diffusion patterns for Twitter sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 32(10): 2026-2039. https://doi.org/10.1109/TKDE.2019.2913641

[20] Pawlik, Ł. (2025). Google Cloud vs. Azure: Sentiment analysis accuracy for Polish and English across content types. Journal of Cloud Computing, 14(1): 17. https://doi.org/10.1186/s13677-025-00742-z