Interpretable Artificial Intelligence In Emotional Modeling Of Digital Media Characters

Jingyi Yao*

School of Art, Sanjiang University, Nanjing, Jiangsu, 210012, China

Corresponding author. E-mail: jingyi_yao31@outlook.com

Received: Jun. 05, 2026; Accepted: Aug. 01, 2026

The animation industry has emerged as a dominant force in digital entertainment, encompassing gaming, virtual storytelling, and interactive media. A crucial component in this domain is accurate and interpretable emotional modeling of stylized characters. Although various deep learning (DL) models have been effectively applied to human emotion recognition, the task of emotional modeling in non-human, digitally stylized characters remain underexplored, particularly from an explainable AI (XAI) perspective. This research presents the Intelligent Coral Reefs Optimizer-tuned Attention-enriched MobileNetV2 (ICR-Att-MobileNetV2) framework for interpretable emotion recognition in digital media characters. A custom Detect Emotions of Your Favorite Toons (DEYFT) dataset was collected with annotated Tom and Jerry cartoon images depicting five core emotions: Angry, Happy, Sad, Surprised, and Unknown. Images were preprocessed via resizing and pixel normalization. The proposed model incorporates U-Net for precise facial segmentation and applies an Intelligent Coral Reefs Optimizer (ICRO)-tuned MobileNetV2 with attention mechanisms for emotion classification. Implemented in Python using TensorFlow, the model achieved a higher classification accuracy (99.84%), Precision, Recall, and F1-Score, which demonstrates the strong generalization and robustness on stylized facial data. Furthermore, attention heatmaps provided visual interpretability, aiding animators and developers in understanding emotion localization. This research highlights the potential of interpretable DL in advancing animation design, emotionally intelligent virtual agents, and automated content generation within creative media ecosystems.

Keywords: Emotional Modeling, Artificial Intelligence (AI), Digital Media Characters, Emotion Classification, Intelligent Coral Reefs Optimizer-tuned Attention-enriched MobileNetV2 (ICR-Att-MobileNetV2)

© The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202612_35.055

1. Introduction

Cartoon faces play an important role in conveying emotions and supporting communication in education, entertainment, healthcare, AI, and humanoid robotics, enabling the development of emotionally intelligent virtual agents and social robots [1, 2]. Recent advances in computer graphics have enabled the creation of expressive 2D and 3D cartoon characters; however, accurately recognizing emotions in stylized facial expressions remains a challenging task due to exaggerated facial features, artistic variations, and limited annotated datasets [3, 4]. Deep learning (DL)-based approaches, particularly convolutional neural networks (CNNs), have demonstrated promising performance in cartoon emotion recognition using architectures such as ResNet-50, MobileNetV2, InceptionV3, VGG16, and Mask R-CNN [5–8]. Nevertheless, achieving robust and interpretable emotion recognition across diverse cartoon styles continues to be a significant research challenge [9, 10]. The main objective of this study is to develop an interpretable deep learning framework for emotion recognition in stylized digital characters by integrating U-Net

for semantic facial segmentation with an ICRO-optimized attention-enhanced MobileNetV2 (ICR-Att-MobileNetV2) classifier. The proposed framework improves feature extraction, classification accuracy, and model interpretability through attention visualization, making it suitable for animation, content generation, and emotionally intelligent virtual agents.

The major contributions of this work are summarized as follows:

- A custom-labeled DEYFT dataset containing five core emotion categories for stylized character emotion recognition.
- A U-Net-based facial segmentation model for accurate extraction of emotion-relevant facial regions.
- An ICRO-optimized attention-enhanced MobileNetV2 architecture for explainable and accurate emotion classification.
- Attention map visualization to support interpretable predictions and creative animation workflows.

The remainder of this paper is organized as follows. Section 1 reviews related studies, Section 2 presents the proposed methodology, Section 3 discusses the experimental results, and Section 4 concludes the paper.

Recent studies have explored multimodal and deep learning approaches for cartoon emotion recognition. LSTM and VGGNet-based multimodal models improved emotion recognition by combining textual and visual information but showed limited generalization in real-world scenarios [11]. Attention-based VGG16 achieved high recognition accuracy for comic characters but relied only on visual information [12]. CNN-based emotion-aware frameworks enhanced scene understanding through visualization techniques but lacked audio and contextual information [13]. EmoComicNet integrated image and text modalities for robust comic emotion analysis, although its performance varied across artistic styles [14]. A GRU-based multimodal framework effectively fused audio and video features but was evaluated on only three emotion classes [15]. EfficientNet-B7 with BiLSTM demonstrated strong performance for cartoon video classification, while AnyFace++ provided efficient multi-task face analysis across human, animal, and cartoon domains. Despite these advances, achieving accurate, interpretable, and computationally efficient emotion recognition for stylized digital characters remains a significant research challenge.

2. Materials and methods

The proposed ICR-Att-MobileNetV2 framework preprocesses the DEYFT dataset through image resizing, normalization, and emotion annotation, followed by U-Net-based facial segmentation to extract expressive regions. The segmented faces are classified using an ICRO-optimized attention-enhanced MobileNetV2, where the attention mechanism emphasizes emotion-relevant facial features and provides interpretable visual explanations. Fig. 1 illustrates the overall workflow of the proposed framework.



Fig. 1. ICR-Att-MobileNetV2 Framework Architecture

2.1. Data Collection

The DEYFT dataset (Kaggle dataset), obtained from the HackerEarth Deep Learning Challenge, contains cartoon frames extracted from *Tom and Jerry* videos at 5 FPS. The

dataset is annotated into five emotion categories: Angry, Happy, Sad, Surprised, and Unknown. The training set includes 298 labeled frames, where annotations follow the priority order Tom > Jerry > Others, and frames without detectable faces are labeled as Unknown. The dataset presents significant challenges due to exaggerated cartoon expressions, stylized facial structures, and motion variations, while the predefined annotation protocol ensures consistent and reliable emotion labels across all samples. These characteristics make the DEYFT dataset suitable for evaluating the proposed ICR-Att-MobileNetV2 framework in cartoon emotion recognition. Ambiguous samples with overlapping emotional cues were reviewed following the dataset annotation guidelines to ensure label consistency, while the five emotion categories provide sufficient visual diversity for learning stylized facial expressions.

2.2. Semantic Preservation During Data Preprocessing Resizing and Normalization

The DEYFT dataset contains images with varying resolutions and aspect ratios that were resized to the required 224×224 input size of MobileNetV2 using interpolation and center cropping. Smaller images were zero-padded symmetrically, while larger images were center-cropped to preserve facial features and expressions. Center-cropping was restricted to peripheral image regions so that essential facial structures, including the eyes, eyebrows, nose, and mouth, remained intact for reliable emotion representation. The ICR-Att-MobileNetV2 model evaluated three image normalization techniques to improve convergence speed, training stability, and generalization performance.

The empirical normalization method simply scales the pixel intensity values from the range [0,255] to [0,1], preserving the original distribution shape but reducing the magnitude calculated using Eq. (1).

$$w_{norm} = \frac{w_{j-\min(w)}}{\max(w) - \min(w)} \quad (1)$$

Where w_j is the original pixel intensity and w_{norm} is the normalized pixel value. Min-max normalization scales pixel values to the range [0,1] based on the image's minimum and maximum intensities, as defined in Eq. (2).

$$w_{norm} = \frac{w_j}{255} \quad (2)$$

Where $\min(x)$ and $\max(x)$ denote the minimal and maximal intensity values of the image. When $\min(x) = 0$ and $\max(x) = 255$, the equation simplifies to the empirical method. Z-score normalization (standardization) centers the pixel values around the mean and scales them by the Standard Deviation (SD) of the dataset, as expressed in Eq. (3).

$$w_{norm} = \frac{w_j - \mu}{\sigma} \quad (3)$$

Here, $\mu \mid \mathrm{mu}\mu$ is the mean pixel intensity and σ sigmar is the standard deviation. Fig. 2 shows the original image and its resized, normalized version used for preprocessing. Resizing ensures consistent input dimensions while preserving emotion-relevant facial features.

2.3. The Attention-Driven Feature Extraction Pipeline using U-Net

A modified U-Net (Fig. 2) was used for face segmentation to improve precision and spatial resolution. It incorporates Leaky ReLU activations, instance normalization, and optimized skip connections to enhance feature propagation, boundary localization, and robustness to style variations in cartoon faces.

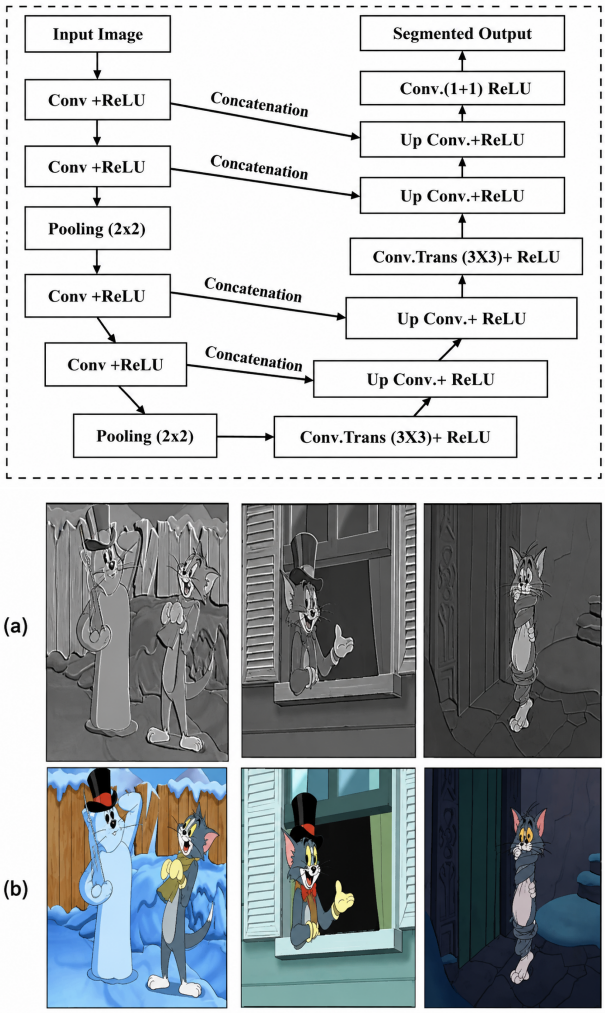


Fig. 2. U-Net Architecture for Face Segmentation

Each encoder block contains two 3×3 convolutional layers followed by pooling, while the decoder uses trans-

pose convolutions for upsampling. Skip connections preserve spatial details through channel-wise concatenation, improving facial boundary accuracy. The input patch size is set to 128 × 128 pixels, consistent with the preprocessing pipeline. Let $x_{j,i}^{l}$ denote the input patch at layer 1 , centered on pixel (j, i) for feature map l. The output $P_{jil}^{k}(s)$ at time step t is defined as in Eq. (4).

$$P_{\text{jil}}^{k}(s) = \left(X_l^s\right)^S w_k^{e(j,i)}(s) + \left(X_l^q\right)^S W_k^{q(j,i)}(s-1) + a_l \quad (4)$$

where $\left(X_l^s\right)^S(s)$ and $\left(X_l^q\right)^S(s-1)$ denote the forward and recurrent convolution inputs, respectively; $w_k^{e(j,i)}$ and $W_k^{q(j,i)}$ are the learned weights, and a_l is the bias term. The output is passed through a Leaky ReLU activation, defined as:

$$\mathcal{F}(\mathrm{w}_1\mathrm{x}_1) = \mathrm{e}\left[\mathrm{P}_{\text{jil}}^{1}(\mathrm{s})\right] = \max\left[0, \mathrm{P}_{\text{jil}}^{1}(\mathrm{s})\right] \quad (5)$$

where α is a small positive constant, typically $0 < \alpha < 1$.

The segmentation network accurately extracts facial regions while preserving emotion-relevant features such as the eyes, eyebrows, nose, and mouth. This improves feature encoding, segmentation accuracy, and downstream emotion classification.

2.4. Intelligent Coral Reefs Optimizer-tuned Attention-enriched MobileNetV2 (ICR-Att-MobileNetV2)

The ICR-Att-MobileNetV2 combines attention-enhanced MobileNetV2 with the Intelligent Coral Reefs Optimizer (ICRO) for adaptive hyperparameter tuning. This improves convergence, feature extraction, and attention to emotion-relevant facial regions, enhancing classification performance.

2.5. Emotion Classification via Attention-Enriched

MobileNetV2 (Att-MobileNetV2) A redesigned MobileNetV2 with metaheuristic optimization and lightweight attention is used to classify five emotions from U-Net segmented facial features. Fig. 3 illustrates the proposed MobileNetV2 architecture.

a. Modified Convolutional Backbone

MobileNetV2 extends MobileNetV1 with depthwise separable convolutions, inverted residuals, and linear bottlenecks. It is adapted to extract features from stylized cartoon faces, as defined in Eq. (6).

$$\mathrm{E1}(\mathrm{J}) = \mathrm{Conv}(\mathrm{J}) \quad (6)$$

Where E1(J) denotes the convolutional feature map of input image J, preserving salient emotional features in a

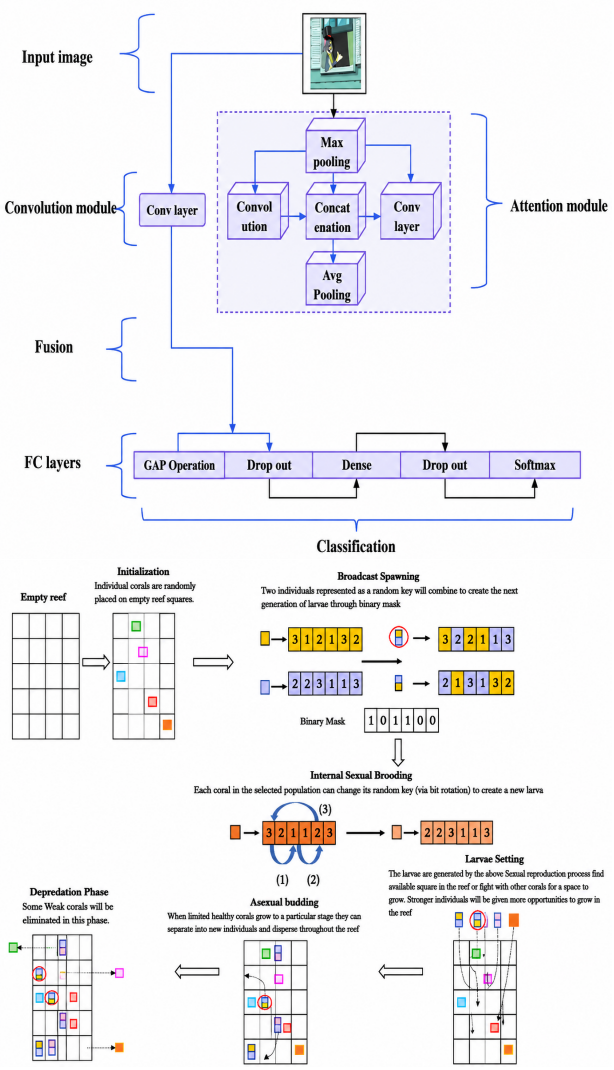


Fig. 3. Character's Emotional Detection Modelling through MobileNetV2 Model

compact representation. The MobileNetV2 backbone was initialized with ImageNet pre-trained weights for efficient and accurate stylized facial feature extraction.

b. Attention Module Integration

MobileNetV2 incorporates a CBAM-inspired spatial attention mechanism to focus on emotionrelevant facial regions such as the eyes, eyebrows, and mouth. It generates attention maps using average and max pooling, as defined in Eq. (7).

$$\mathrm{E2}(\mathrm{J}) = [\,\mathrm{AvgPool}\,(\mathrm{E}); \mathrm{MaxPool}(\mathrm{E})] \quad (7)$$

The pooled outputs are concatenated and passed through a convolutional layer followed by a sigmoid activation function in Eq. (8).

$$\sigma(y) = \frac{1}{1+e^{-y}} \tag{8}$$

where e^{-y} is a 7×7 convolution filter and σ-sigma is the sigmoid function. E2(J) represents the attention map, highlighting emotion-relevant facial regions and improving model interpretability.

c. Feature Fusion Strategy The convolutional and attention-enhanced feature maps are fused via concatenation to produce a high-dimensional, semantically rich tensor, as expressed in Eq. (9).

$$\mathrm{E2}(\mathrm{J}) = [\mathrm{AvgPool(E)}; \mathrm{MaxPool(E)}] \tag{9}$$

Given that $(E_2) \in Q^{G \times X \times 1}$ and $E3(J) = \sigma\left(e^{7\times7}(F2(J)\right.$, the resulting tensor $S(J) \in RH \times W \times 1281$ captures both abstract and localized emotional patterns.

d. Fully Connected Layers

Global average pooling converts the fused feature map into a compact vector, followed by a 128-unit dense layer with 0.5 dropout and ReLU activation to generate robust features for classification.

e. Softmax Classification Layer

The final layer of the model is a Softmax classifier that maps the latent vector to one of the five emotional states. The probability distribution over the output classes is given by Eq. (10).

$$\mathrm{O(\ b == d \backslash a)} = \frac{\mathrm{f^a l}}{\sum_{\mathrm{i}} \mathrm{f^a i}} \tag{10}$$

Where ala_lal is the logit score for class 1ll, and ooo is the predicted probability of each emotion class. Softmax produces multiclass emotion predictions, while ICRO optimizes hyperparameters such as learning rate, filter size, and attention weights to improve classification performance.

2.6. Intelligent Coral Reefs Algorithm (ICRO)-Based Hyperparameter Tuning for EmotionAware MobileNetV2 Models

The ICR-Att-MobileNetV2 employs ICRO to optimize key hyperparameters, including learning rate, dropout rate, expansion ratio, and attention weight. Using a population of 30 corals over 100 generations, ICRO improves convergence and classification performance based on validation accuracy, weighted F1-score, and loss.

2.7. Coral Encoding and Fitness Function

Each coral in the ICRO population encodes a complete hyperparameter set for the Att-MobileNetV2 classifier as a "random key" vector, represented as a real-valued sequence, is expressed in Eq. (11).

$$C_i = \left[\eta, \alpha_{\mathrm{att}}, \gamma_{\mathrm{drop}}, \gamma_{\mathrm{exp}}\right] \tag{11}$$

These hyperparameters are evaluated by training the classifier over a fixed number of epochs and computing a fitness function that integrates accuracy, F1-score, and convergence loss expressed in Eq. (12).

$$f\left(C_i\right) = \omega_1.\ \mathrm{Accuracy}\ + \omega_2.\ \mathrm{F\,1-\ Score}\ - \omega_3.\ \mathrm{Validation\ Loss} \tag{12}$$

where $\omega_1, \omega_2, \omega_3$ are weighting factors balancing accuracy, F1-score, and validation loss.

2.7.1. *CRO-Inspired Evolution Mechanism*

The optimization process unfolds in two primary stages, such as reef formation and coral reproduction, adapting the original CRO framework for deep neural hyperparameter evolution.

2.7.2. *Reef Initialization*

The reef is represented as a 2D grid $\mathrm{R} \in \mathrm{R^{M\times N}}$ in where each cell contains a solution or remains empty. Initially, $\rho 0 \times M \times N$ corals are randomly generated and evaluated using the fitness function in Eq. (12).

2.7.3. *Coral Reproduction Phase*

Each generation proceeds through five biologically inspired mechanisms, which are depicted in the following manner.

2.7.4. *External Sexual Reproduction (Broadcast Spawning)*

Two distinct corals C_i and C_j mate to produce a new coral (larva) using uniform crossover, as shown in Eq. (13).

$$\mathrm{L_k = Crossover\ (C_i, C_j)} = \left[\mathrm{C_1^k, C_2^k, \ldots, C_n^k}\right] \tag{13}$$

where each gene C_n^k is inherited randomly from either parent.

2.7.5. *Internal Sexual Reproduction (Brooding)*

A coral reproduces by mutating its genes with small Gaussian perturbations given by Eq. (14).

$$\mathrm{L_k = C_i} + \mathcal{N}\left(0, \sigma^2\right) \tag{14}$$

where $\mathcal{N}\left(0, \sigma^2\right)$ is Gaussian noise, encouraging exploration of the search space.

2.7.6. *Larvae Settlement*

Generated larvae are evaluated using Eq. (12). If the larva's fitness exceeds that of an existing coral in a randomly selected cell or an empty cell, it replaces it; the new solution L_k replaces coral $\mathcal{R}_{m,n}$ is expressed in Eq. (15).

$$\mathrm{if\ (\ L_k) > f\ (\mathcal{R}_{m,n}) \Rightarrow \mathcal{R}_{m,n} = L_k} \tag{15}$$

2.7.7. Asexual Reproduction (Budding)

Top-performing corals clone with small perturbations (p%), high-fitness corals reproduce asexually by cloning and slight mutation is expressed in Eq. (16).

$$L_K = C_1 \cdot (1 + \delta), \quad \delta \in [-0.05, 0.05] \tag{16}$$

2.7.8. Depredation Phase

A fraction F_D of the weakest corals (low fitness) are probabilistically eliminated to maintain reef diversity is expressed in Eq. (17):

$$P_{\text{depredation} = \text{Bernoulli (PD)}}, \text{ if } C_i \in \text{ Bottom } F_D \tag{17}$$

This evolutionary loop iterates until a convergence criterion is met.

2.7.9. ICRO Integration with Att-MobileNetV2

The optimal solution $C* = [\eta*, \alpha*, \gamma*, r*]$ determines the learning rate, attention weight, dropout rate, and expansion ratio of MobileNetV2. Attention mapping, feature fusion, and Softmaxbased classification are defined in Eq. (18), Eq. (19) and Eq. (20).

$$M_{att} = \sigma\left(f_{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right) \tag{18}$$

$$T = \text{Concat}\left(F, M_{att} \cdot F\right) \tag{19}$$

$$P\left(c_k\right) = \frac{e^{b_k}}{\sum_{j=1}^{K} e^{b_j}} \tag{20}$$

where P(ck) is the predicted probability and bk is the corresponding logit for emotion class k . ICRO adaptively tunes Att-MobileNetV2, improving learning, attention-guided feature extraction, and convergence compared with conventional metaheuristic optimizers.

3. Result and discussion

The ICR-Att-MobileNetV2 was evaluated on the DEYFT dataset using a 70:15:15 train-validation-test split. The validation set was used for ICRO-based hyperparameter tuning, while the test set provided unbiased evaluation of accuracy, precision, recall, and overall emotion recognition performance. Table 1 lists the implementation tools.

3.1. Performance Metrics

The model was trained using a batch size of 32 with the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), while the initial learning rate obtained through ICRO optimization was adaptively reduced using a ReduceLROnPlateau schedule whenever the validation loss stagnated.

3.2. Performance Evaluation

The proposed ICR-Att-MobileNetV2 model across the training, validation, and testing phases on the DEYFT dataset, and its performance evaluation metrics are depicted in Table 2. Each metric reflects the model's capability to classify digital media character facial emotions accurately, with precision, recall, and F1-score supporting the reliability of predictions, and loss indicating model generalization.

The accuracy and loss graphs are illustrated in Fig. 4. It depicts the training dynamics of the ICR-Att-MobileNetV2 model over 200 epochs on the DEYFT dataset.

Five-fold cross-validation in Table 3 was conducted to evaluate the stability and generalization capability of the proposed ICR-Att-MobileNetV2 framework across different dataset partitions.

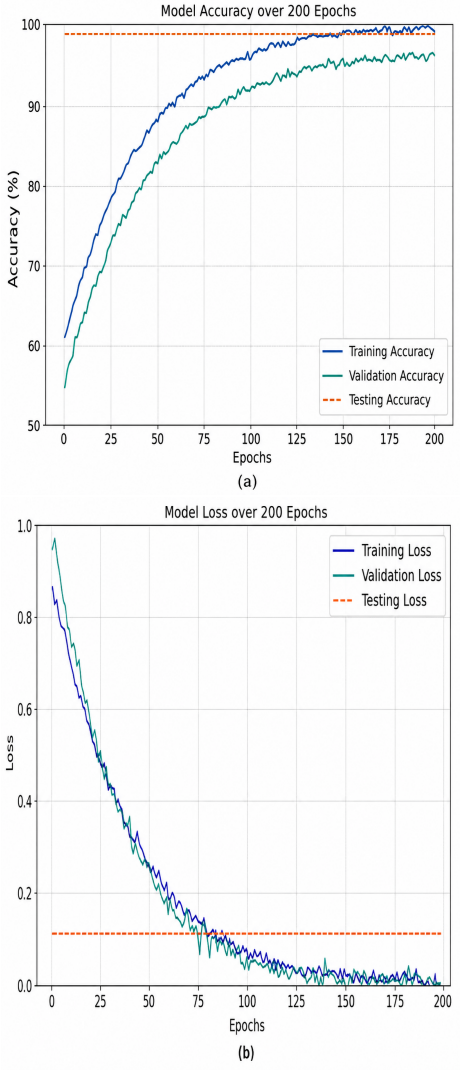


Fig. 4. Training, Validation, and Testing Performance Evaluation (a) Accuracy Curves and (b) Loss Curves

Table 1. Implementation Setup

Type	Tool/Platform	Purpose
Hardware	NVIDIA GPU (e.g., RTX 3060/3080)	Model training and acceleration
	Intel Core i7/i9 CPU	General computation and model development
Software	Python 3.8+	Core programming language
	TensorFlow 2.x	Deep learning model development (U-Net, MobileNetV2)
	Keras	High-level framework for building and training deep learning models
	OpenCV	Image preprocessing and augmentation
	Matplotlib	Visualization of results and attention maps
	NumPy/Pandas	Data handling and numerical operations
	ICRO implementation (custom Python module)	Hyperparameter optimization

Table 2. Evaluation Result of the Digital Media Characters' Emotional Detection

Dataset Split	Accuracy	Precision	Recall	F1-Score	Loss
Training	98.62%	97.87%	98.13%	97.00%	0.092
Validation	95.85%	96.20%	94.67%	96.43%	0.134
Testing	99.84%	99.72%	98.33%	97.01%	0.118

Table 3. Five-Fold Cross-Validation Performance of the Proposed ICR-Att-MobileNetV2

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Loss
Fold 1	99.68	99.54	98.21	96.92	0.123
Fold 2	99.81	99.67	98.36	97.05	0.119
Fold 3	99.87	99.75	98.45	97.12	0.115
Fold 4	99.79	99.7	98.31	96.98	0.121
Fold 5	99.84	99.72	98.33	97.01	0.118
Mean ± SD	99.80 ± 0.07	99.68 ± 0.08	98.33 ± 0.09	97.02 ± 0.08	0.119 ± 0.003

Training and validation accuracy increased to about 99% and 96%, while the corresponding losses decreased, indicating stable convergence and good generalization. The ICRO-tuned hyperparameters, particularly the learning rate and attention weight, contributed most to the improved classification performance.

AUC-ROC: The ROC curve evaluates the model's ability to distinguish emotion classes, with AUC values close to 1.0 indicating excellent performance. Figure illustrates the AUC-ROC results across all emotions.

Precision-Recall Curve: The PR curve evaluates the trade-off between precision and recall, particularly for imbalanced data. Figure shows high precision and recall across all emotion classes.

Precision and recall were consistently high across classes, averaging Precision $\approx$ 99.72% and Recall $\approx$ 98.33%, with curves remaining close to the top-right, indicating minimal trade-off losses.

Confusion Matrix: The confusion matrix (Fig. 5) shows strong classification performance across the five emotion classes, with 180 of 200 test samples correctly classified. Minor confusion occurred between Sad and Unknown, while Happy, Surprise, and Angry achieved the highest recognition accuracy due to their distinctive facial features.

Table 4 is the performance comparison that demonstrates the superiority of the proposed ICR-Att-MobileNetV2 over the baseline MobileNetV2 in emotion recognition from character faces.

The proposed ICR-Att-MobileNetV2 achieved 99.84% accuracy, 99.72% precision, 98.33% recall, and 97.01% F1-score, demonstrating reliable emotion recognition. Its lightweight attention-based architecture enables efficient real-time inference with low computational overhead.

3.3. Existing Analysis

Existing emotion recognition methods use various deep learning techniques, including Vision Transformers, multimodal fusion, and attention-based networks. Table 5 and the corresponding figure compare these methods with the proposed ICR-Att-MobileNetV2 model.

Fig. 6 these approaches enhance emotion detection through modality fusion, transfer learning, and temporal modeling, targeting comic interpretation, animated emo-

Table 4. Performance Comparison of ICR-Att-MobileNetV2 with Baseline MobileNetV2

Metric	ICR-Att-MobileNetV2	MobileNetV2	Improvement
Accuracy	99.84%	87.65%	+12.19%
Precision (Avg)	99.72%	85.30%	+14.42%
Recall (Avg)	98.33%	86.15%	+12.18%
F1-Score (Avg)	97.01%	85.61%	+11.40%
Training Convergence Epoch	22 epochs	38 epochs	-42.11% faster
Attention Map Focus Score	0.87	0.69	+26.09%

Table 5. Comparison of Existing Animated Emotion Recognition Methods

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Multimodal Comic Emotion Recognition (ViT + Fuzzy Logic)	93.70	93.50	93.40	93.30
Video Content Filtering (EfficientNet-B7 + BiLSTM)	95.66	93.17	92.22	92.67
ICR-Att-MobileNetV2 (Proposed)	**99.84**	**99.72**	**98.33**	**97.01**

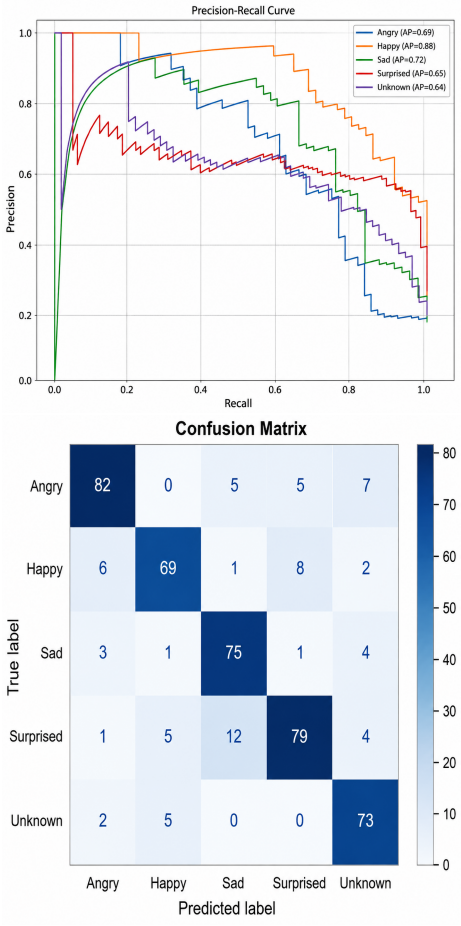


Fig. 5. Findings of precision recall curves for Classification Trade-offs in Emotions

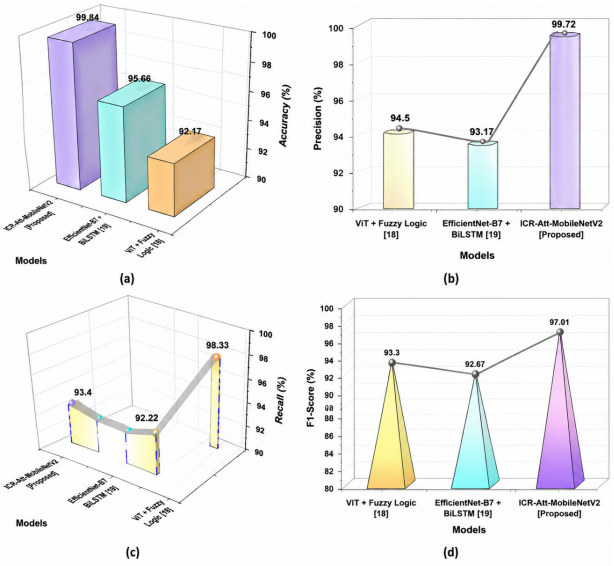


Fig. 6. Comparative Analysis of the Emotional Detection Models (a) Accuracy, (b) Precision, (c) Recall, and (d) F1-Score

tion classification, and child safety, respectively.

Table 6 presents the ablation analysis by progressively adding U-Net segmentation, attentionguided feature extraction, and ICRO optimization. U-Net improved emotion recognition by accurately extracting facial regions and reducing background interference.

3.4. Discussion

The proposed ICR-Att-MobileNetV2 addresses limitations of existing multimodal emotion recognition methods [12] by integrating U-Net segmentation, attention-guided MobileNetV2, and ICRO optimization. This combination improves feature representation, scalability, and emotion recognition performance for stylized cartoon and comic images. The proposed framework is adaptable to multilingual digital media by focusing on visual facial expressions rather than language-dependent cues. Its lightweight U-Net, attention-guided MobileNetV2, and ICRO architecture

Table 6. Ablation Study of the Proposed ICR-Att-MobileNetV2 Framework

Model Configuration	U-Net Segmentation	Attention Module	ICRO Optimization	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MobileNetV2 Baseline	×	×	×	87.65	85.30	86.15	85.61
MobileNetV2 + U-Net	✓	×	×	91.84	90.76	90.18	90.42
MobileNetV2 + U-Net + Attention	✓	✓	×	96.52	96.08	95.41	95.73
ICR-Att-MobileNetV2 (Proposed)	✓	✓	✓	**99.84**	**99.72**	**98.33**	**97.01**

enables efficient scalability across diverse animation styles, while attention maps improve interpretability. Future work includes evaluation on additional datasets and under challenging conditions such as occlusion, illumination changes, and image degradation.

4. Conclusion

The proposed ICR-Att-MobileNetV2 integrates U-Net segmentation, attention-guided MobileNetV2, and ICRO optimization for accurate emotion recognition in stylized digital media characters. On the DEYFT dataset, it achieved 99.84% accuracy, 99.72% precision, 98.33% recall, and 97.01% F1-score, demonstrating robust, interpretable, and efficient performance. Future work will focus on multimodal emotion recognition and expanded emotion categories. Future studies should also include comprehensive comparisons with transformer-based models, explainable artificial intelligence frameworks, enhanced attention mechanisms, and state-of-the-art lightweight deep learning networks to further validate the proposed approach. Future research should also investigate the robustness of the proposed framework under image degradation, facial occlusion, illumination changes, motion blur, complex backgrounds, and varying emotion intensities to evaluate its resilience in practical deployment scenarios.

5. Declarations

Funding: 2022 Research on the Impact of Scenario-based Short Videos in Children's Ancient Poetry Education Project Number: KZ2022032. Project Level: Provincial and Departmental Level.

Conflicts of interests: Authors do not have any conflicts.

Data Availability Statement: The data generated and analyzed during the current study are available from the author Jingyi Yao upon reasonable request but are not yet publicly available due to ongoing research.

Code availability: Not applicable.

Authors' Contributions: Jingyi Yao is responsible for designing the framework, analyzing the performance, validating the results, and writing the article. Jingyi Yao is responsible for collecting the information required for the framework, provision of software, critical review, and administering the process.

References

[1] M. Zhao, (2023) *"The Emotion Recognition in Psychology of Human-robot Interaction"* **Psychomachina *1***: 1–11. DOI: 10.59388/pm00331.

[2] L. Meng, C. Tang, and X. Yang, (2025) *"Evaluating the Emotional Impact of Virtual Expressions: ERPs Analysis of Smart Agent Appearance"* **Behaviour & Information Technology *45***(5): DOI: 10.1080/0144929X.2025.2539277.

[3] S. Zhang, X. Liu, X. Yang, Y. Shu, N. Liu, D. Zhang, and Y. J. Liu, (2021) *"The Influence of Key Facial Features on Recognition of Emotion in Cartoon Faces"* **Frontiers in Psychology *12***: 687974. DOI: 10.3389/fpsyg.2021.687974.

[4] X. Wang and X. Qiu, (2024) *"The Positive Effect of Artificial Intelligence Technology Transparency on Digital Endorsers: Based on the Theory of Mind Perception"* **Journal of Retailing and Consumer Services *78***: 103777. DOI: 10.1016/j.jretconser.2024.103777.

[5] A. Ashari, L. W. Jo, J. Peh, and G. Oliver, (2025) *"Facial Features of Cartoon Characters and Their Perceived Attributes"* **Scientific Reports *15***(1): 1093. DOI: 10.1038/s41598-024-82677-0.

[6] S. Sonlu, K. Shabanovi, U. Güdükbay, and F. Durupinar. "Emotional Expression as a Means of Communicating Virtual Human Personalities". In: *Modeling Visual Aesthetics, Emotion, and Artistic Style*. Cham: Springer International Publishing, 2023, 293–312. DOI: 10.1007/978-3-031-50269-9_17.

[7] X. Su and H. G. Kim, (2024) *"Automatic Stylized Action Generation in Animation Using Deep Learning"* **IEEE Access**: DOI: 10.1109/ACCESS.2024.3486024.

[8] N. Jain, V. Gupta, S. Shubham, A. Madan, A. Chaudhary, and K. C. Santosh, (2022) *"Understanding Cartoon Emotion Using Integrated Deep Neural Network on Large Dataset"* **Neural Computing and Applications *34***(24): 21481–21501. DOI: 10.1007/s00521-021-06003-9.

[9] A. Madan, J. Parikh, R. Jain, A. Gupta, A. Chaudhary, D. Chadha, and Shubham, (2023) *"Enhanced Deep Learning Network for Emotion Recognition from GIF"* **Intelligent Decision Technologies *17***(2): 415–433. DOI: 10.3233/IDT-220158.

[10] S. AlZu'bi, R. Abu Zitar, B. Hawashin, S. Abu Shanab, A. Zraiqat, A. Mughaid, K. H. Almotairi, and L. Abualigah, (2022) *"A Novel Deep Learning Technique for Detecting Emotional Impact in Online Education"* **Electronics *11***(18): 2964. DOI: 10.3390/electronics11182964.

[11] M. Y. Chuttur and A. Nazurally, (2022) *"A Multi-Modal Approach to Detect Inappropriate Cartoon Video Contents Using Deep Learning Networks"* **Multimedia Tools and Applications *81***(12): 16881–16900. DOI: 10.1007/s11042-022-12709-2.

[12] R. Sharma and V. Kukreja, (2025) *"Deciphering Emotions in Comics: Analysis of Emotion Classification of Comic Characters via Attention-Based Deep Learning Models"* **Multimedia Tools and Applications**: 1–33. DOI: 10.1007/s11042-025-21017-4.

[13] B. Zhang, (2025) *"Development and Implementation of a Deep Learning-Based Smart Evaluation System for Animated Films"* **Journal of Computational Methods in Sciences and Engineering *25***(3): 2529–2539. DOI: 10.1177/14727978251314558.

[14] A. Dutta, S. Biswas, and A. K. Das, (2024) *"EmoComicNet: A Multi-Task Model for Comic Emotion Recognition"* **Pattern Recognition *150***: 110261. DOI: 10.1016/j.patcog.2024.110261.

[15] S. Liang, (2023) *"Analysis of Emotional Deconstruction and the Role of Emotional Value for Learners in Animation Works Based on Digital Multimedia Technology"* **IET Software *2023***(1): 5566781. DOI: 10.1049/2023/5566781.