

Article

3D ResESAM: A Sequential Attention Mechanism Enhanced 3D ResNet50 for Dog Emotion Recognition

Xiangyun Guo ¹, Jiashuo Feng ^{1,*}, Xiaoya Kong ¹, Chuiyu Kong ² and Yuxin Liu ¹

¹ College of Management Science and Engineering, Beijing Information Science & Technology University, Beijing 100192, China; guoxy@bistu.edu.cn (X.G.); 2024021050@bistu.edu.cn (X.K.); liuyuxin6409@163.com (Y.L.)

² College of Engineering, China Agricultural University, Beijing 100083, China; chuiyukong99@163.com

* Correspondence: fjshhxx@163.com; Tel.: +86-131-8125-7406

Abstract

An animal's emotion is regarded as a critical indicator for assessing its welfare. At present, most canine emotion recognition models rely on static images, which leads to insufficient modeling of spatiotemporal features and low accuracy. To address this limitation, this study proposed a 3D ResESAM model based on 3D ResNet50, which is a pre-trained model on the Kinetics dataset. Specifically, the Enhanced Sequential Attention Module (ESAM) is integrated into the backbone to improve the capability of extracting and enhancing spatiotemporal features, and YOLO (You Only Look Once) was employed to detect and crop canine facial regions to reduce interference from complex backgrounds (e.g., furniture, humans, or other animals). To evaluate the effectiveness of 3D ResESAM, a canine emotion dataset named Dog-Face was constructed, which contains four emotions: happiness, sadness, anger, and calmness. Experimental results demonstrate that the proposed 3D ResESAM achieves an accuracy of 0.9239, representing a 3.25% improvement over the baseline 3D ResNet50. Furthermore, compared with other models such as C3D and R(2 + 1)D, whose accuracies are just 0.5215 and 0.8570, respectively, from different attention mechanism perspectives, the proposed model outperforms not only the baseline model without ESAM but also models incorporating other attention modules such as ECA and CBAM, with average accuracies being 0.8972 and 0.9103, respectively. In addition, cross-domain experiments are conducted on a human emotion dataset. The results show that 3D ResESAM achieves an average accuracy improvement of 2.39% over 3D ResNet50, demonstrating its significantly superior performance, effectiveness and potential in real-world applications.

Keywords: canine emotion; animal welfare; 3D ResNet50; video recognition; attention mechanism



Academic Editor: Douglas O'Shaughnessy

Received: 2 June 2026

Revised: 7 July 2026

Accepted: 8 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Dogs are important companion animals to human beings, and their emotion directly influences their health, behavior, and the quality of human–dog interactions. Existing research has shown that canine facial expressions are closely associated with specific emotions [1] and can therefore serve as a simple and non-invasive indicator for assessing a dog's emotion. However, accurately recognizing canine emotions based on facial expressions remains a challenging scientific problem [2]. In recent years, with the rapid advancement of deep learning, substantial models represent critical temporal features. In contrast, 3D convolutional neural networks (3D CNNs) are capable of jointly modeling spatial and temporal features, enabling a more accurate representation of the continuous evolution of

emotional states—for example, the gradual transition of ears from drooping to erect, or variations in wrinkle depth around the nose and mouth.

3D convolution was first introduced by Ji et al. in 2012, who proposed 3D convolutional neural networks (3D CNNs) for video analysis, particularly in the context of human action recognition [3]. Since then, 3D CNNs have been widely adopted for various video classification and recognition tasks. For instance, Teixeira et al. extended pre-trained 2D CNNs into 3D CNNs to capture spatiotemporal features in video sequences, enabling continuous prediction of emotion in terms of valence and arousal for dogs [4]. Similarly, Anjali Sharma et al. proposed an attention-enhanced 3D CNN framework for real-time emotion recognition in movie scenes. By incorporating attention mechanisms to emphasize key frames and emotionally salient regions, the approaches improve recognition accuracy while reducing computational redundancy [5]. These studies collectively demonstrate that 3D CNNs exhibit significant potential in video classification and emotion recognition tasks, mainly due to their superior capability in modeling spatiotemporal features.

Deep learning demonstrates great advantages in modeling spatiotemporal features and progress has been achieved in the field of human facial emotion recognition [6]. Nevertheless, compared with human emotion recognition, canine emotion recognition still faces special challenges, such as limited data and insufficient domain knowledge. Moreover, emotions are not instantaneous states but dynamic processes that evolve gradually in response to external stimuli and internal physiological changes; thus, emotion recognition can be essentially regarded as a temporal modeling problem [7]. Traditional 2D convolutional neural networks (2D CNNs) cannot capture temporal features from static images and therefore struggle to effectively exhibit considerable potential for emotion recognition. Studies specifically on the spatiotemporal feature modeling of canine facial expressions remain limited. Therefore, in order to achieve accurate recognition of dog facial emotions, this study developed a canine facial emotion recognition model based on 3D convolutional neural networks. The main contributions are summarized as follows:

- (1) To address the scarcity of publicly available datasets in the field of canine emotion recognition, this study constructed a canine emotion video dataset. Videos were collected from publicly licensed video platforms and then selected according to predefined criteria. And then we used the Dog Facial Action Coding System (DogFACS) as an objective annotation standard [8]. Finally, we obtained a dog emotion dataset comprising 2057 video clips with a total duration of 6171 s.
- (2) To improve the accuracy of emotion recognition, this study adopts dynamic image sequences of dogs as input features, with a specific focus on the facial region. Experimental results demonstrate that utilizing dynamic facial image sequences enables more effective capture of the temporal evolution and subtle motion patterns of facial expressions, therefore facilitating the development of more robust emotion recognition models.
- (3) To mitigate the loss of temporal features in existing approaches, the Enhanced Sequential Attention Module (ESAM) is designed to strengthen the model's capability to capture and discriminate dynamic spatiotemporal emotional features.

2. Related Work

With the continuous advancement of deep learning technologies, research on canine emotion recognition has explored a variety of technical approaches. Existing methods can be broadly categorized into a number of classes based on different perspectives. The first category comprises recognition approaches based on static images, which currently dominate this field. These methods typically employ pre-trained convolutional neural networks (CNNs) on large-scale image datasets and apply transfer learning to recognize

canine emotions by static images. For example, Chauhan et al. fine-tuned models such as EfficientNetB0 on a dataset of over 13,000 images following data augmentation, achieving an accuracy of 93.0% in classifying three emotions: anger, sadness, and happiness [9]. Tribethran et al. utilized a modified VGG16 trained on a dataset of 4000 canine emotion images, obtaining an average F1-score of 84.30% on the test set [10]. Kowalczyk et al. proposed an enhanced mVGG16 model based on VGG16 and VGG19 and performed transfer learning for fine-tuning. Using a dataset of 360 images, their model achieved accuracy rates of 92% and 58% on test sets consisting of dogs from the same breed and different breeds, respectively [11]. Li et al. proposed a localized facial feature recognition framework based on YOLOv8, which detects key anatomical regions, including the eyes, ears, and muzzle, and integrates their predictions using a confidence-weighted decision fusion strategy [12].

The primary advantage of these approaches lies in the use of transfer learning, which enables 2D CNNs pre-trained on large-scale datasets to enhance performance on new tasks, particularly in scenarios with limited labeled data. However, since these methods rely on static images, they are inherently incapable of modeling the temporal dynamics of emotions. Moreover, the relatively small scale of available datasets further constrains their generalization capability.

The second is video-driven methods based on sequence analysis, whose primary objective is to capture dynamic features. For example, Chen et al. proposed a multi-module framework that integrates YOLOv3, DeepDogTrack, and LDFDMN (Long Short-Term Deep Features of Dog Memory Networks) for dog detection, tracking, and emotion recognition in video data [13]. Specifically, the recognition module combines ResNet with a long short-term memory (LSTM) network to learn temporal dependencies from consecutive frames. The main advantage of this approach lies in its ability to infer emotion based on a dog's posture. However, its recognition accuracy (approximately 76–82%) does not significantly outperform that of state-of-the-art static image-based models. Franzoni et al. employed 2D CNNs such as VGG19 for emotion recognition; although videos were utilized, the sequences were decomposed into a large number of independent frames during processing, resulting in a maximum accuracy of only 60.5% [14]. In essence, this approach addresses a fundamentally dynamic problem using a static paradigm and fails to construct an effective spatiotemporal feature representation.

In addition, some studies have attempted to bypass complex image representations by leveraging keypoint-based features for pose analysis and emotion recognition. For example, Ferres et al. employed DeepLabCut to extract the coordinates of 24 key body points from dogs, which were subsequently input into neural networks or decision trees for classification [15]. The advantages of this approach are reduced data dimensionality, lower computational cost, and robustness to background variations. However, its recognition accuracy (up to 67.5%) remains relatively limited because of the inherent ambiguity of pose-based representations and the lack of fine-grained facial information.

Several studies have also explored canine emotion recognition from the perspectives of supervised, self-supervised, and unsupervised learning. For instance, Bhave et al. applied the MoCo framework to a self-constructed dataset comprising 2184 images, where the unsupervised model achieved an accuracy of only 43.2% [16]. Boneh-Shitrit et al. conducted a comparison between supervised and self-supervised learning for canine emotion recognition. Their results showed that self-supervised models based on the DINO (self-distillation with no labels) framework, including DINO-ResNet50 and DINO-ViT, generally outperformed supervised models with identical backbones on a dataset of 19,392 images extracted from 248 video clips. Among these, the DINO-ViT achieved an accuracy of 85.3% in frame-level emotion classification, significantly surpassing the 78.0% obtained by its

supervised counterpart [17]. The primary advantage of self-supervised methods lies in their ability to leverage large-scale unlabeled data. However, their performance in canine emotion recognition still falls behind that of supervised methods.

Overall, canine emotion recognition has gradually evolved from conventional CNN-based static image classification toward video-based dynamic analysis and fine-grained facial feature representation. In recent years, studies have extensively adopted techniques such as transfer learning, sequence modeling, localized facial feature extraction, pose estimation, feature fusion, and self-supervised representation learning to enhance emotion representation and recognition performance. With the continuous development of canine emotion recognition, integrating temporal information with facial representation has gradually emerged as a promising research direction. However, research on dynamic facial expression modeling remains limited, largely due to the scarcity of publicly available video datasets containing facial expressions.

3. Materials and Methods

3.1. Dataset Construction

3.1.1. Data Collection

The emotional videos of dogs used in this study were collected from mainstream platforms. The footage primarily consists of close-up facial recordings. A data collection tool was developed using the Python programming language to facilitate collection. The collection was restricted to publicly available and authorized content, conducted solely for non-commercial research purposes, and complied with the robots.txt protocol. The collected videos were selected according to predefined criteria. Specifically, only video clips in which the dog's major facial regions, including the eyes, ears, nose, and mouth, were clearly visible and free from severe occlusion were retained. All video selection was performed manually by the authors according to these predefined criteria. Following data cleaning and preprocessing, a dataset named Dog-Video was constructed, comprising 2057 valid video clips. Each clip has an average duration of approximately 3 s, with a total duration of 6171 s. Dog-Video encompasses a diverse range of real-world scenarios, dog breeds, and camera perspectives, thereby providing a robust data foundation for enhancing model generalization in complex environments.

3.1.2. Data Preprocessing

To simplify the input representation for subsequent analysis while preserving essential emotional information, a YOLO-based detection model was trained to identify dog faces using self-annotated video frames in conjunction with publicly available canine facial datasets. The trained model was applied in a frame-by-frame manner to detect and crop facial regions from each video frame. In the model configuration, the confidence threshold was set to 0.6. The detection results were further manually verified to remove false positives (e.g., human or cat faces). For video samples with insufficient numbers of valid frames, data padding or removal strategies were employed to ensure dataset quality. Subsequently, all cropped facial images were uniformly normalized and resized to a fixed resolution of 224×224 pixels, which forms the Dog-Face dataset, and it contains only facial regions. This design helps reduce background interference and enables the model to focus on discriminative facial expression features. The overall data preprocessing pipeline is illustrated in Figure 1.

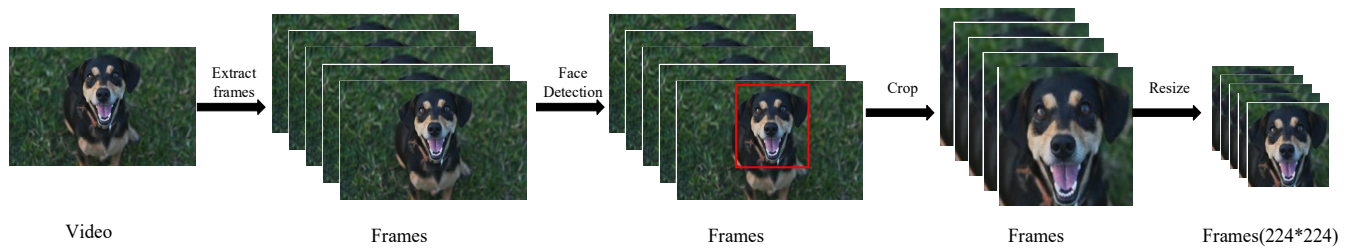


Figure 1. Data preprocessing workflow. Video frames are extracted, detected, cropped, and resized to 224×224 pixels. Red bounding boxes indicate the detected dog faces.

3.1.3. Data Annotation

To ensure objective and reproducible annotation, the DogFACS (Dog Facial Action Coding System) was adopted as the fundamental framework for describing canine facial movements. DogFACS provides a standardized, anatomy-based coding system for observable facial actions, including action units (AUs) and action descriptors (ADs). However, DogFACS itself does not define mappings between specific AUs/ADs and discrete emotional states. Therefore, emotion labels for the video clips were assigned with reference to previous empirical studies investigating the associations between canine facial movements and emotion-eliciting contexts. Emotion annotation was performed by comprehensively considering the behavioral context, overall facial expression, and representative facial action patterns reported in the literature, rather than relying solely on any individual AU or AD. For emotion categories for which the current literature lacks consistent evidence supporting specific DogFACS patterns, annotations were based primarily on the dogs' overall behavioral manifestations and facial characteristics. The representative facial action patterns used as references for emotion annotation are summarized in Table 1.

Table 1. Representative DogFACS facial action patterns reported in previous studies for different emotion-related contexts.

Emotion	Representative AUs/ADs Reported in Previous Studies	Describe
Happy [18]	AU27	mouth stretch
Anger [19]	AU109 + 110 AD33	nose wrinkle; upper lip; raiser
Sad [18]	AU45	blink
Calm [18]	AU1	inner brow raiser

Based on these criteria, a canine facial dataset encompassing the four emotional categories was ultimately constructed. A subset of the annotated Dog-Face dataset is illustrated in Figure 2, which shows representative samples across different emotional categories. The train–test split was performed at the video level. Specifically, all 2057 valid video clips from the Dog-Video dataset were randomly partitioned into training and testing subsets with a ratio of approximately 8:2, resulting in 1646 clips for training and 411 clips for testing. All frames originating from the same video clip were assigned exclusively to either the training or testing set, ensuring that no data from the same video appears in both subsets. In terms of data volume, the four categories are relatively balanced in both the training and test sets. Although the “Happy” and “Calm” categories contain slightly more valid frames, while the “Anger” and “Sad” categories contain slightly fewer, the overall differences are not substantial. In the training set, the number of valid frames per category ranges from approximately 23,524 to 26,423, which helps mitigate potential category bias during model training. Similarly, the test set exhibits a comparable distribution, with valid

frame counts ranging from approximately 5829 to 6342 across categories. This balanced distribution ensures the reliability of the evaluation results and enhances the model's generalization capability.

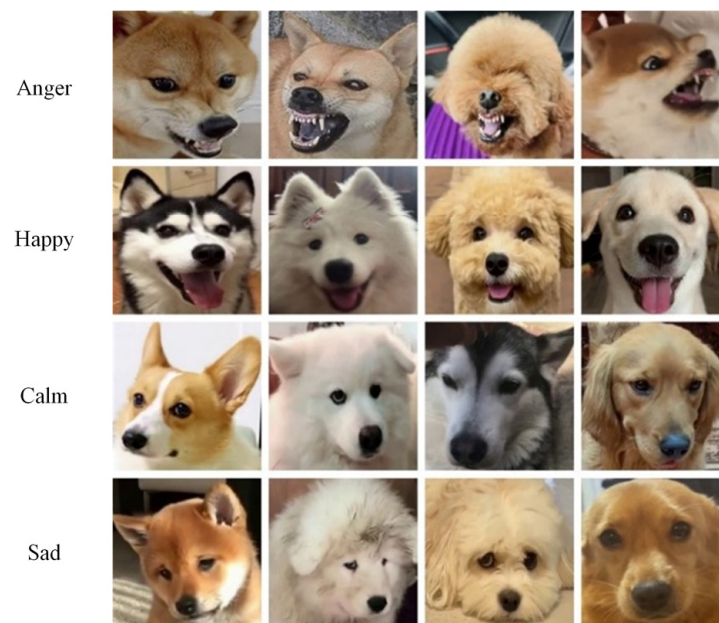


Figure 2. Sample examples from the dataset. Each image shows a representative dog face from one of the four expression categories.

3.2. 3D ResESAM Model

The proposed 3D ResESAM model, which adopts 3D ResNet50 as its backbone network to borrow the capability for spatiotemporal features extraction, introduces the Enhanced Sequential Attention Module (ESAM). This module can strengthen the ability of the model's frames to focus on salient channel features and critical spatial regions, mitigating information loss caused by dimensionality reduction and enhancing the modeling of spatiotemporal dependencies. Since canine emotional expressions are often subtle and mainly reflected through fine-grained facial movements and temporal behavioral changes, the incorporation of the attention mechanism can also enable the network to emphasize discriminative emotional cues while suppressing irrelevant background information.

3.2.1. 3D ResNet50 Model

3D ResNet50 typically consists of an input layer, multiple residual stages, and an output layer [20]. The input comprises a sequence of consecutive video frames, which first undergo preliminary feature extraction through a 3D convolutional layer followed by a pooling layer. During feature learning, the 3D convolutional kernels operate jointly across temporal and spatial dimensions, enabling the network to capture motion and temporal dependencies between frames (3D convolutions model spatiotemporal information directly). Subsequently, the extracted features are fed into a series of stages composed of stacked residual blocks. Each stage progressively increases the number of channels while reducing the spatiotemporal resolution, thereby facilitating multi-scale feature extraction. Finally, the classification task is performed using global average pooling followed by fully connected layers. The incorporation of residual connections effectively alleviates the vanishing gradient problem in deep networks and improves training stability. The overall network structure is illustrated in Figure 3.

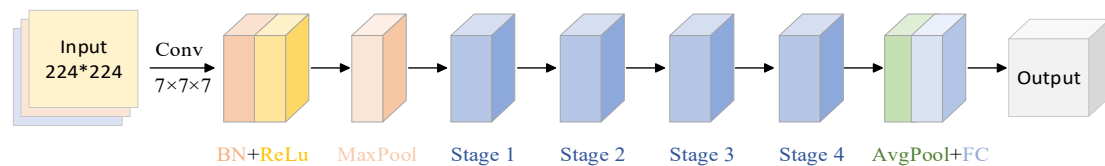


Figure 3. Overall architecture of the 3D ResNet50 network. It consists of convolutional layers, four stages, and a final fully connected layer.

3.2.2. Construction of the 3D ResESAM Model

This paper proposes 3D ResESAM, which integrates ESAM formed by Efficient Channel Attention (ECA) [21] and the Convolutional Block Attention Module (CBAM) [22] based on 3D ResNet50. The overall structure of the proposed model is illustrated in Figure 4. At the input stage, the model takes a video sequence as input. The input data X are defined as shown in Equation (1).

$$X \in \mathbb{R}^{C \times T \times H \times W} \quad (1)$$

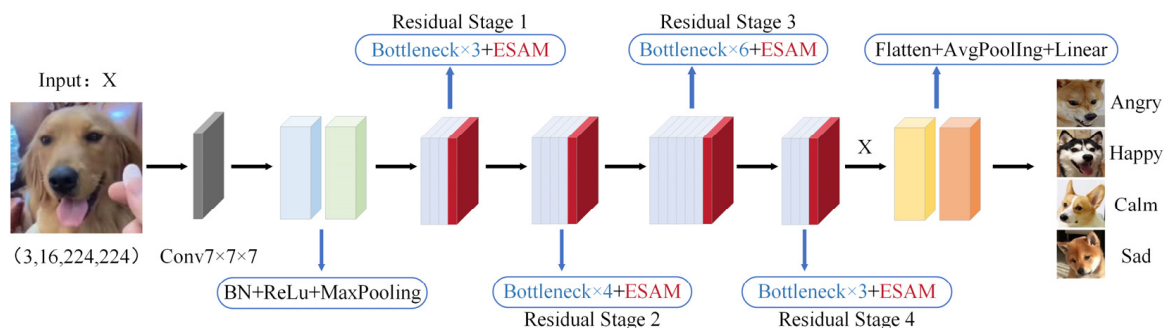


Figure 4. Overall architecture of the 3D ResESAM network. The model integrates 3D residual learning with the ESAM attention mechanism for canine emotion recognition from facial video sequences.

Here, C denotes the number of input channels, corresponding to the 3 RGB color channels. T represents the temporal length, defined as a sequence of 16 consecutive video frames for each input clip. H and W denote the spatial dimensions, i.e., the height and width of each video frame, respectively.

Preliminary feature extraction is first performed using a 3D convolutional layer, which operates jointly across temporal and spatial dimensions, thereby enabling effective modeling of spatiotemporal information in video data. During the deep feature learning stage, the model adopts four residual stages as the backbone network. Each stage consists of multiple bottleneck residual blocks, with an Enhanced Sequential Attention Module (ESAM) inserted after each stage to refine feature representations. Finally, the enhanced features are mapped to the target emotion categories. A 3D global average pooling layer is applied to compress both spatial and temporal dimensions, producing a fixed-length feature vector. The formulation of 3D global average pooling is given in Equation (2).

$$X_c = \frac{1}{T \times H \times W} \sum_{t=1}^T \sum_{h=1}^H \sum_{w=1}^W F_c(t, h, w) \quad (2)$$

Here, X_c denotes the output of the c -th channel after 3D global average pooling. T , H , and W represent the temporal length, height, and width, respectively. $F_c(t, h, w)$ denotes the feature value at position (t, h, w) in the c -th channel.

The extracted features are subsequently flattened into a one-dimensional representation. A fully connected layer is then applied to perform linear mapping, producing prediction scores for each emotion category. The predicted scores are defined in Equation (3).

$$z = Wx + b \quad (3)$$

Here, z denotes the output vector, W represents the weight matrix, x denotes the input feature vector, and b represents the bias vector.

The class probabilities are computed using the Softmax function, as defined in Equation (4).

$$p_i = \frac{\exp^{z_i}}{\sum_{j=1}^C \exp^{z_j}} \quad (4)$$

Here, P denotes the predicted probability, i indexes the i -th class, and C represents the total number of classes.

3.2.3. ESAM

ESAM is the key improvement of the proposed model, aiming to enable the network to focus on emotion-relevant spatiotemporal regions. The structure of the ESAM is illustrated in Figure 5. Specifically, the ESAM replaces the channel attention submodule in the original Convolutional Block Attention Module with an Efficient Channel Attention module (Figure 5a) and sequentially integrates it with the spatial attention submodule of CBAM (Figure 5b). First, the ECA module is applied immediately after the global average pooling layer. A 1×1 convolutional layer is employed to replace the fully connected layer in the original channel attention mechanism, thereby avoiding dimensionality reduction while enabling efficient cross-channel interaction. This design enhances informative feature channels while suppressing less relevant ones. To further facilitate richer cross-channel interactions for layers with higher channel dimensions, the convolutional kernel size k is adaptively determined by a mapping function, as defined in Equation (5).

$$k = \varphi(C) = \left\lceil \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rceil \quad (5)$$

Here, $\gamma = 2$, $b = 1$, and C denotes the number of input channels. Through the mapping function ϕ , higher-dimensional channels are associated with broader cross-channel interaction ranges, whereas lower-dimensional channels correspond to more localized interactions due to the nonlinear mapping. The enhanced features are subsequently fed into the spatial attention submodule. Within this module, max pooling and average pooling are applied along the channel dimension to generate two $1 \times H \times W$ feature maps. These maps are then concatenated and passed through a 7×7 convolutional layer to produce a single-channel feature map F .

The resulting feature map F is further processed by a sigmoid activation function to generate the spatial attention map. Finally, this attention map is multiplied by the input feature map to restore the original $C \times H \times W$ dimensionality. The spatial attention feature map is achieved with Equation (6).

$$Ms(F) = \sigma \left(f^{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)]) \right) \quad (6)$$

Here, F denotes the input feature map with dimensions $C \times H \times W$; $Ms(F)$ denotes the final generated spatial attention feature map; AvgPool and MaxPool denote average pooling and max pooling on the feature map along the channel dimension, respectively; f

denotes a 7×7 convolution; and $[\text{AvgPool}(F); \text{MaxPool}(F)]$ denotes concatenating average pooling and max pooling along the channel axis.

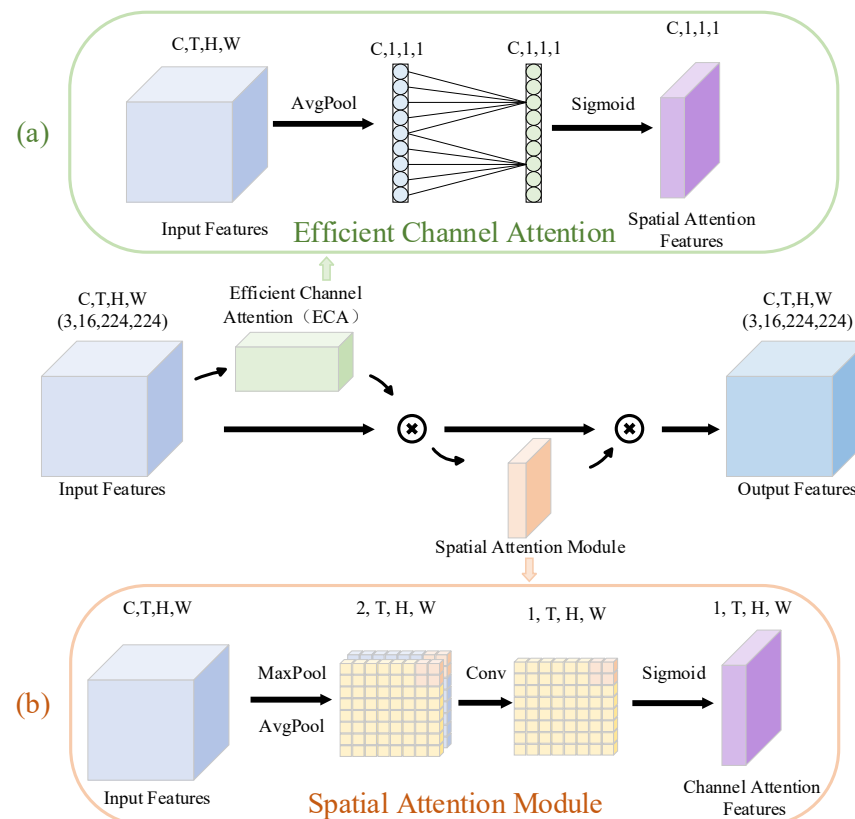


Figure 5. Structure diagram of the ESAM. The module integrates channel and spatial attention mechanisms for enhanced feature extraction. (a) Structure of the channel attention module; (b) structure of the spatial attention module.

3.3. Model Evaluation Metrics

This study evaluates the performance of the proposed model using standard metrics, including accuracy, precision, recall, and F1-score. Accuracy is defined as the proportion of correctly predicted samples to the total number of samples and serves as a fundamental indicator for assessing overall performance. The formulation of accuracy is given in Equation (7).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (7)$$

In this context, TP, FP, FN, and TN denote the numbers of true positives, false positives, false negatives, and true negatives, respectively.

Precision is defined as the proportion of true positive samples among all samples predicted as positive by the model, reflecting the accuracy of positive predictions. The formulation of precision is given in Equation (8).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

Recall is defined as the proportion of actual positive samples that are correctly identified by the model, reflecting its ability to capture positive instances. The formulation of recall is given in Equation (9).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

The F1-score is defined as the harmonic mean of precision and recall, and is widely used to provide a balanced evaluation of model performance, particularly under class-imbalanced conditions. The formulation of the F1-score is given in Equation (10).

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

4. Results

4.1. Experimental Settings

The experiments were conducted on the Ubuntu 22.04 operating system using Python 3.10, the PyTorch 2.1.0 deep learning framework, and an NVIDIA GeForce RTX 3090 (24 GB) GPU. During training, the Adam optimizer was employed for parameter optimization, with an initial learning rate of 0.001. The batch size was set to 64 to balance training efficiency and computational resources. L2 regularization (weight decay) and the Dropout mechanism were applied to mitigate overfitting. The activation function was ReLU, and the cross-entropy loss function was used as the optimization objective, for it is well-suited for multi-class classification tasks. The model was trained for 100 epochs and initialized with Kinetics-700 pre-trained weights. Each input sample consisted of 16 consecutive RGB frames resized to 112×112 pixels. Random cropping and horizontal flipping were adopted for data augmentation during training. In addition, the learning rate was adaptively adjusted using the ReduceLROnPlateau scheduling strategy.

4.2. Model Performance

Figure 6 presents the results for each emotion category on the test set. Overall, our proposed model demonstrates better results and relatively balanced recognition performance across the four emotion categories. Specifically, the Sad category achieves the best overall performance, with an accuracy of 0.9406, indicating strong discriminative capability for Sad emotion. The Happy category ranks second with an accuracy of 0.9293. In contrast, the Anger category exhibits slightly lower performance, with an accuracy of 0.9191, indicating a certain degree of misclassification. The Calm category achieves the lowest accuracy (0.9083) among the four categories. Overall, the model performs well in recognizing Sad and Happy categories, while its performance on Anger and Calm is comparatively weaker.

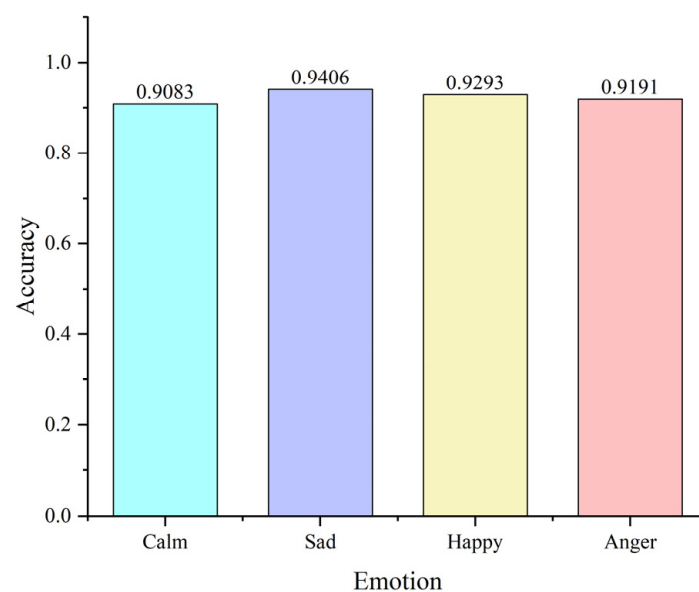


Figure 6. Test set accuracy by emotion category. The chart shows accuracy values for Calm, Sad, Happy, and Anger.

To quantitatively evaluate the model's fine-grained recognition performance, a Sankey diagram of misclassification patterns on the test set is presented in Figure 7. In the diagram, nodes on the left represent the ground-truth emotion, while those on the right denote the categories predicted by the model. The width of the connecting flows indicates the frequency of each misclassification, thereby revealing the dominant error patterns. The Calm category exhibits a relatively high number of misclassifications, with 15 samples incorrectly predicted as Happy, 15 as Anger, and 7 as Sad. For the Sad category, the primary misclassification direction is toward Anger (13 instances), followed by Happy (6 instances), while confusion with Calm is relatively rare (2 instances). Misclassifications for Happy are more evenly distributed, with five samples each incorrectly classified as Anger, Sad, and Calm. Notably, significant bidirectional confusion is observed between Anger and Calm.

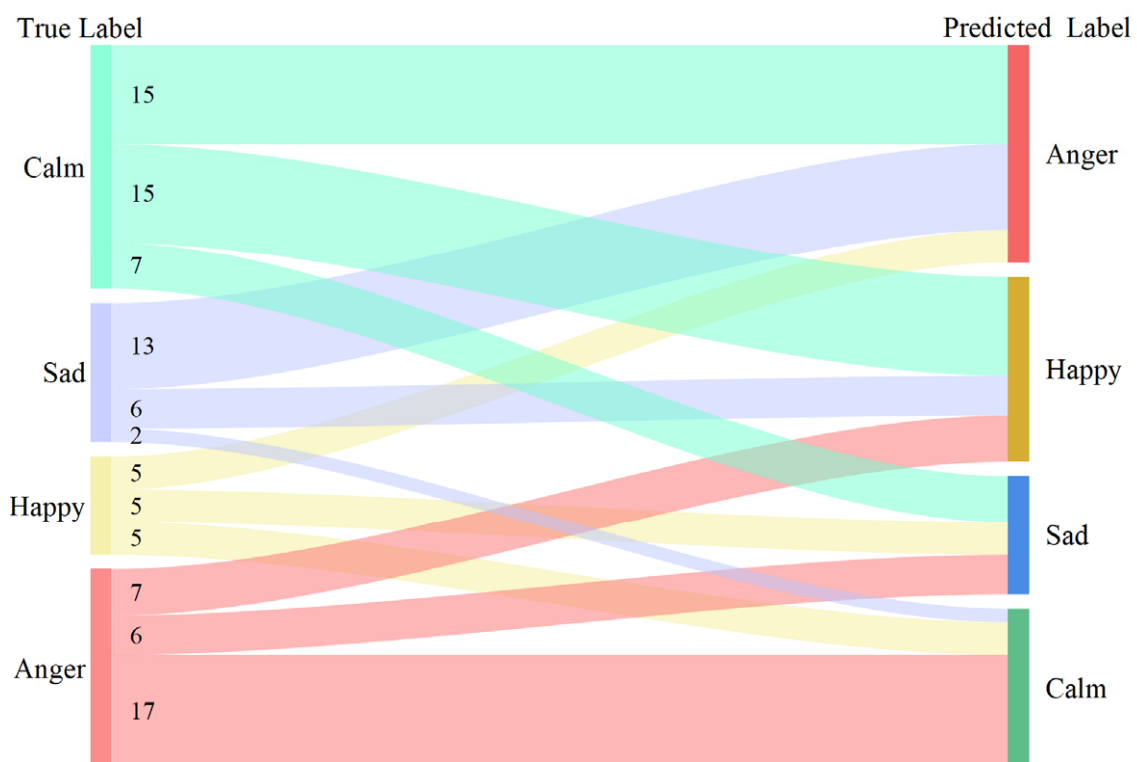


Figure 7. Distribution of misclassification cases of the model on the test set. The figure shows the distribution of true labels versus predicted labels across the four emotion categories.

4.3. Comparative Experimental Analysis

To validate the effectiveness of the proposed 3D ResESAM model, several representative 3D convolutional neural network models, including C3D [23], I3D [24], and R(2 + 1)D [25], were selected as benchmark models to conduct comparative experiments on our Dog-Face dataset. All experiments were conducted under identical parameter settings and training protocols to ensure a fair comparison. Figure 8 presents the comparison of recognition accuracy for each model on the test set. Overall, the proposed 3D ResESAM model achieves the best performance among all compared methods, attaining an accuracy of 0.9239. In contrast, the conventional 3D ResNet50 achieves an accuracy of 0.8948; although its performance is competitive, it remains notably inferior to 3D ResESAM. The R(2 + 1)D model achieves an accuracy of 0.8570, whereas the C3D model attains a maximum accuracy of only 0.5215.

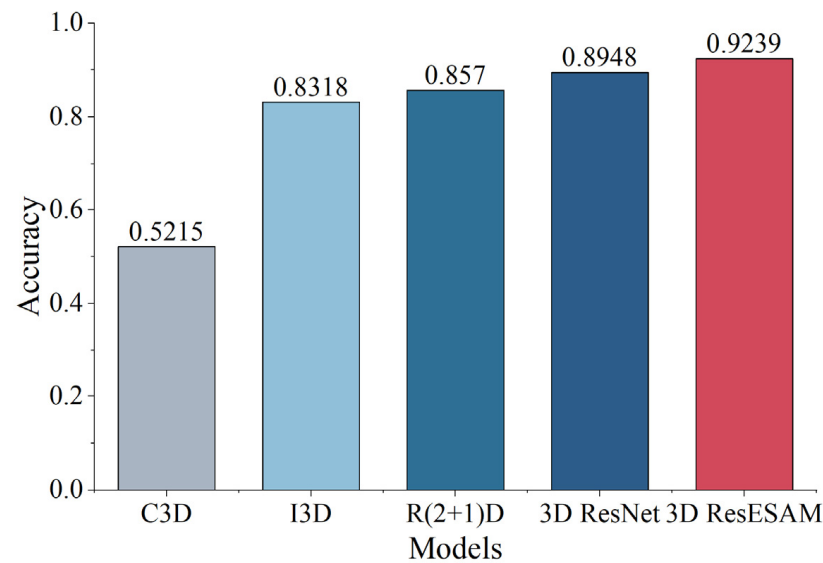


Figure 8. Accuracy comparison of different 3D CNN models. The figure compares the accuracy of C3D, I3D, R(2 + 1)D, 3D ResNet50, and the proposed 3D ResESAM.

4.4. Comparative Study of Attention Modules

To systematically evaluate the contributions of each component and the overall effectiveness of the proposed 3D ResESAM model, 3D ResNet50 was adopted as the baseline. The ECA channel attention module and the CBAM hybrid attention module were then sequentially incorporated, acquiring two variants: 3D ResNet50 + ECA and 3D ResNet50 + CBAM. These models were further compared with the complete 3D ResESAM model, which integrates an enhanced attention mechanism. The result is presented in Figure 9, which indicates that the introduction of attention mechanisms consistently improves model performance. Specifically, the baseline 3D ResNet50 model achieves an accuracy of 0.8948, and it increases to 0.8972 and 0.9103, respectively, after incorporating ECA modules and CBAMs. The proposed 3D ResESAM model achieves the highest accuracy of 0.9239, corresponding to an absolute improvement of 0.0291 and a relative improvement of 3.25% over the baseline, significantly outperforming other variants.

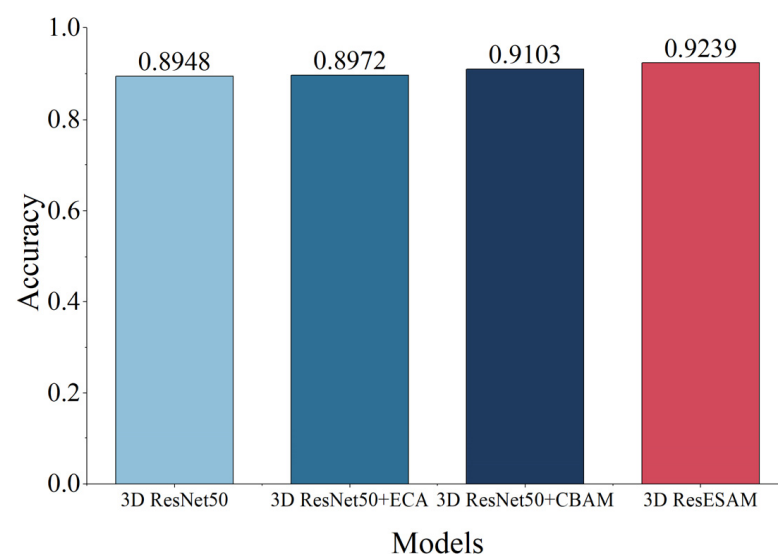


Figure 9. Accuracy comparison of different attention mechanism modules. The figure compares the accuracy of 3D ResNet50, 3D ResNet50 + ECA, 3D ResNet50 + CBAM, and 3D ResESAM.

Table 2 shows the fine-grained emotion recognition results. It indicates that the baseline 3D ResNet50 model exhibits relatively balanced performance across all emotion categories; however, its overall accuracy remains limited. Its best performance is observed in the Sad emotion, achieving a precision of 0.9466. After incorporating the ECA module, the performance improves, particularly in terms of recall for the Happy and Anger categories, which reach 0.9667 and 0.9196, respectively. The CBAM-enhanced model achieves strong performance based on accuracy for Sad and Calm recognition. In contrast, the proposed 3D ResESAM model demonstrates the most consistent performance across all four emotions and achieves the best results, with accuracies of 0.9293 (Happy), 0.9406 (Sad), 0.9191 (Anger), and 0.9083 (Calm).

Table 2. Accuracy of different attention mechanisms across dog emotion categories. The best results in each column are shown in bold.

Model		Happy	Sad	Anger	Calm
3D ResNet50	Accuracy	0.9225	0.8778	0.8867	0.8623
	Precision	0.9072	0.9466	0.8726	0.8318
	Recall	0.9123	0.8717	0.8833	0.8615
	F1-Score	0.9097	0.9076	0.8781	0.8462
3D ResNet50 + ECA	Accuracy	0.9259	0.8877	0.8932	0.8746
	Precision	0.9136	0.9340	0.8492	0.8881
	Recall	0.9667	0.8683	0.9196	0.8984
	F1-Score	0.9396	0.9000	0.8831	0.8933
3D ResNet50 +CBAM	Accuracy	0.9288	0.9299	0.8978	0.9023
	Precision	0.9136	0.9302	0.8257	0.8492
	Recall	0.8737	0.8523	0.8605	0.8645
	F1-Score	0.8933	0.8895	0.8428	0.8568
3D ResESAM	Accuracy	0.9293	0.9406	0.9191	0.9083
	Precision	0.9471	0.9313	0.9019	0.8911
	Recall	0.9057	0.9406	0.8932	0.9266
	F1-Score	0.9259	0.9359	0.8975	0.9085

4.5. Evaluation of Preprocessing Effectiveness

To investigate the importance of facial regions in emotion recognition, a controlled comparative experiment was designed. Using an identical model (3D ResESAM) and training configuration, experiments were conducted on two datasets, Dog-Video and Dog-Face. These datasets share identical video and emotion labels, and their key difference is input spatial information: Dog-Video retains the original background and full scene, while Dog-Face focuses on the facial regions. Experimental results are presented in Table 3. It can be seen that recognition performance across all emotions improves consistently on the Dog-Face dataset. Specifically, for Happy, precision and recall increased by 0.0049 and 0.0320, respectively, accompanied by a slight improvement in accuracy. For the Sad emotion, the most notable gains were observed, with recall improving by 0.0883, although accuracy showed a slight decline. The result of the Anger emotion demonstrates consistent improvement across all metrics, particularly with a 0.0441 increase in precision. For the Calm category, recall improves significantly and accuracy increases marginally, while precision remains relatively stable.

Table 3. Results of dog emotion recognition with different datasets. The best results in each column are shown in bold.

	Dog-Video				Dog-Face			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Happy	0.9024	0.9422	0.8737	0.9070	0.9293	0.9471	0.9057	0.9259
Sad	0.8913	0.9618	0.8523	0.9040	0.9406	0.9313	0.9406	0.9359
Anger	0.9167	0.8578	0.8605	0.8594	0.9191	0.9019	0.8932	0.8975
Calm	0.9003	0.8920	0.8645	0.8782	0.9083	0.8911	0.9266	0.9085

4.6. Evaluation of 3D ResESAM Generalization

Further, to evaluate the generalization of the proposed 3D ResESAM model, we constructed the SCM human emotion dataset by integrating three public datasets: SMIC [26], CAS(ME)² [27], and MMEW [28]. To ensure class balance, emotion categories with limited samples, such as Fear and Sad, were excluded, resulting in a dataset comprising four emotions—Happy, Disgust, Anger, and Surprise. The detailed sample composition of the resulting SCM dataset is summarized in Table 4. Data augmentation techniques were applied during training to further maintain class balance. The performance of the 3D ResNet50 and 3D ResESAM models on the SCM dataset is shown in Table 5.

Table 4. Sample count distribution of different human emotion categories across datasets.

	Disgust	Fear	Happy	Sad	Surprise	Anger	Total
SMIC	10	16	17	8	20	0	71
CAS(ME) ²	88	0	155	0	0	102	345
MMEW	72	16	36	13	89	8	234
SCM	170	0	208	0	109	110	597

Table 5. Accuracy of different models for each human emotion category on the SCM dataset. The best results in each column are shown in bold.

	3D ResNet50				3D ResESAM			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Anger	0.9691	0.9728	0.9479	0.9600	0.9729	0.9767	0.9518	0.9642
Disgust	0.9661	0.9448	0.9527	0.9488	0.9887	0.9392	0.9753	0.9569
Happy	0.9214	0.9450	0.9080	0.9259	0.9142	0.9696	0.9009	0.9341
Surprise	0.9868	0.9803	0.9752	0.9775	0.9967	0.9902	0.9883	0.9893

It can be seen that 3D ResESAM outperforms 3D ResNet50 in most metrics. Our model demonstrates superior performance, especially for Anger, Disgust, and Surprise. For Anger, accuracy improves by 0.0038 and the F1-score improves by 0.0042. For Disgust, accuracy and recall both increase by 0.0226. For Surprise, all metrics approach optimal levels, with accuracy and F1-score improving by 0.0099 and 0.0118, respectively. For Happy, although accuracy decreases slightly by 0.0072, precision increases by 0.0246 and the F1-score increases by 0.0082.

5. Discussion

Based on the results in Section 4.2, the relatively lower accuracy of the Calm category suggests that distinguishing Calm emotion remains relatively challenging for the model. In addition, the high number of misclassifications for Calm suggests that the model may have limited stability in recognizing emotions with low intensity or subtle expressions.

For the Sad category, the model tends to confuse the Sad category with more intense negative emotional states. Meanwhile, the relatively even distribution of misclassifications in the Happy category suggests that the decision boundary for this category is relatively ambiguous. Furthermore, the significant bidirectional confusion observed between the Anger and Calm categories indicates that certain negative and neutral emotional states may share similar visual characteristics across different dog breeds, which poses a challenge for accurate discrimination. This phenomenon may be attributed to inter-breed variability. For example, breeds such as Bulldogs naturally exhibit pronounced wrinkles around the muzzle; consequently, their facial appearance in a Calm state may resemble the muscle tension associated with Anger, leading to misclassification. However, this hypothesis was not quantitatively evaluated in the present study and requires further investigation using breed-specific analyses.

Regarding the comparative experimental results in Section 4.3, the C3D model shows the lowest recognition accuracy among all compared methods. This performance gap may be attributed to the relatively shallow structure of C3D, which limits its ability to capture fine-grained facial expression features.

In Section 4.4, the experimental results demonstrate the effectiveness of attention mechanisms in guiding the model to focus on informative features in emotion recognition. After incorporating the ECA module, the improvement in recall for the Happy and Anger categories indicates enhanced sensitivity in capturing emotional features with a very noticeable appearance. However, the gains in precision and F1-score are not very large, possibly due to the lack of spatial attention constraints, resulting in a less focused attention distribution. Although the CBAM-enhanced model achieves strong performance in Sad and Calm recognition, its channel attention mechanism relies on fully connected layers, which increases the number of parameters and may introduce redundant features, thereby affecting overall model stability. In contrast, the proposed 3D ResESAM model demonstrates superior stability and generalization ability across all four emotion categories.

The results in Section 4.5 suggest that the model benefits from focusing on facial regions, which contain critical cues for emotion recognition, such as subtle texture variations (e.g., around the eyes and mouth) and static postural features (e.g., ear orientation). In contrast, videos containing background information, such as substantial body movement, full-body interactions, or features heavily dependent on body language (e.g., tail wagging during motion or posture during jumping), may introduce redundant or irrelevant information, thereby negatively affecting recognition performance.

Regarding the generalization evaluation in Section 4.6, it indicates that the proposed model still maintains an advantage in discriminative capability for the Happy category despite the slight decrease in accuracy. These results further validate the strong generalization ability of the proposed 3D ResESAM. A possible reason for the superior performance of the proposed model on the human emotion dataset is that human facial expressions generally exhibit more explicit and standardized emotional patterns compared with canine emotional expressions. In addition, public human emotion datasets are typically collected under relatively controlled conditions, with more consistent illumination, viewpoints, and annotation standards, thereby reducing data variability and facilitating model learning.

Although DogFACS and other studies provide an objective and widely accepted framework for canine facial action annotation, the interpretation of animal emotional states remains inherently complex. The emotion labels adopted in this study were assigned according to observable facial action units rather than subjective human perception, and facial expressions may not always correspond uniquely to an animal's internal emotional state. Therefore, the reported recognition performance should be interpreted as the model's

ability to classify DogFACS-based emotion categories instead of directly inferring the true emotional state of dogs.

6. Conclusions

This study proposed an innovative dog emotion recognition model, named 3D ResESAM, which is built upon 3D ResNet50 and integrates ESAM. After localizing and cropping facial regions by a YOLO-based object detection module, irrelevant background information was effectively removed and interference was reduced to preserve key regions for emotion recognition. The designed ESAM enhanced 3D ResESAM's ability to focus on salient facial regions while it captures temporal dynamics of facial expression, thereby enabling more accurate extraction of emotion-related spatiotemporal features. Experimental results demonstrate that, compared with benchmarking methods, the proposed 3D ResESAM achieves balanced and robust performance across all emotions, with accuracies reaching 0.9293 for Happy, 0.9406 for Sad, 0.9191 for Anger, and 0.9083 for Calm. Furthermore, cross-domain experiments on human emotion datasets validate its strong generalization. The proposed framework has promising potential for practical applications in animal welfare monitoring, including veterinary clinics, animal shelters, pet hospitals, and companion animal health management. By automatically recognizing canine emotional states from facial video sequences, it may assist veterinarians and caregivers in monitoring stress, anxiety, pain, and other behavioral conditions in a non-invasive manner.

The complexity of canine emotion recognition is largely due to its multi-modal characteristics, involving voice, body posture, olfactory, etc. Moreover, similar facial expression patterns may convey different meanings under varying contexts, and individual differences further increase the difficulty of accurate interpretation. Therefore, future research should explore the integration of multi-source and multi-modal data, such as behavioral indicators and physiological signals, to improve emotion recognition accuracy.

Author Contributions: Writing—review and editing, X.G.; formal analysis, X.G. and J.F.; funding acquisition, X.G.; investigation, X.G.; project administration, X.G.; resources, X.G.; data curation, J.F.; methodology, J.F. and C.K.; writing—original draft, J.F.; visualization, J.F. and Y.L.; validation, X.K.; conceptualization, C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. Ethical review and approval were waived for this study because only publicly available online videos were used, and no human or animal subjects were directly involved.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is not publicly available. The video data were collected from publicly available online sources and further processed by the authors through detection, clipping, cropping, resizing, and annotation procedures to construct the final dataset used in this research. Researchers interested in the dataset may contact the corresponding author for reasonable academic requests.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Descovich, K.A.; Wathan, J.; Leach, M.C.; Buchanan-Smith, H.M.; Flecknell, P.; Farningham, D.; Vick, S.-J. Facial expression: An under-utilised tool for the assessment of welfare in mammals. *Altex* **2017**, *34*, 409–429. [[CrossRef](#)] [[PubMed](#)]
2. Sun, N.; Zhang, Y.; Xie, L.W.; Xu, S.; Huang, H.; Xu, X.; Sun, Y.; Tao, S. Research progress on the application of dog facial expressions in emotion assessment. *Acta Lab. Anim. Sci. Sin.* **2025**, *33*, 567–580. (In Chinese)

3. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [\[CrossRef\]](#)
4. Teixeira, T.; Granger, É.; Lameiras Koerich, A. Continuous Emotion Recognition with Spatiotemporal Convolutional Neural Networks. *Appl. Sci.* **2021**, *11*, 11738. [\[CrossRef\]](#)
5. Sharma, A.; Aurid, T.; Katekar, A.R.; Muthu, S.; Manzoor, S.; Thanganadar, H.T. Real-Time Emotion Recognition in Film Scenes Using Attention-Based 3D CNN Models. In Proceedings of the 2025 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 5–7 August 2025; pp. 461–467. [\[CrossRef\]](#)
6. Xie, H.X.; Lo, L.; Shuai, H.H.; Cheng, W.H. An overview of facial micro-expression analysis: Data, methodology and challenge. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1857–1875. [\[CrossRef\]](#)
7. Cohen, I.; Sebe, N.; Garg, A.; Chen, L.S.; Huang, T.S. Facial expression recognition from video sequences: Temporal and static modeling. *Comput. Vis. Image Underst.* **2003**, *91*, 160–187. [\[CrossRef\]](#)
8. Waller, B.M.; Caeiro, C.; Peirce, K.; Burrows, A.M.; Kaminski, J. *DogFACS: The Dog Facial Action Coding System*; University of Portsmouth: Portsmouth, UK, 2013.
9. Chauhan, R.; Manchanda, H.; Tyagi, N. Classifying dog emotions using transfer learning and pre-trained models. In *Proceedings of 2024 14th International Conference on Cloud Computing, Data Science & Engineering*; IEEE: Noida, India, 2024; pp. 638–643.
10. Tribethran, S.; Hasan, N.J.P.; Rahman, A. Classification of Dog Emotion Using Transfer Learning on Convolutional Neural Network Algorithm. *Paradig.-J. Komput. Inform.* **2024**, *26*, 102–108. [\[CrossRef\]](#)
11. Kowalczyk, Z.; Czubenko, M.; Żmuda-Trzebiatowska, W. Categorization of emotions in dog behavior based on the deep neural network. *Comput. Intell.* **2022**, *38*, 2116–2133. [\[CrossRef\]](#)
12. Li, D.-L.; Lee, S.-K.; Tsai, Y.-C. Deep Learning-Based Dog Expression Recognition. *IEEE Access* **2026**, *14*, 5675–5683. [\[CrossRef\]](#)
13. Chen, H.-Y.; Lin, C.-H.; Lai, J.-W.; Chan, Y.-K. Convolutional neural network-based automated system for dog tracking and emotion recognition in video surveillance. *Appl. Sci.* **2023**, *13*, 4596. [\[CrossRef\]](#)
14. Franzoni, V.; Biondi, G.; Milani, A. Advanced techniques for automated emotion recognition in dogs from video data through deep learning. *Neural Comput. Appl.* **2024**, *36*, 17669–17688. [\[CrossRef\]](#)
15. Ferres, K.; Schloesser, T.; Gloor, P.A. Predicting dog emotions based on posture analysis using DeepLabCut. *Future Internet* **2022**, *14*, 97. [\[CrossRef\]](#)
16. Bhave, A.; Hafner, A.; Bhave, A.; Gloor, P.A. Unsupervised canine emotion recognition using momentum contrast. *Sensors* **2024**, *24*, 7324. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Boneh-Shitrit, T.; Amir, S.; Bremhorst, A.; Mills, D.S.; Riemer, S.; Fried, D.; Zamansky, A. Deep learning models for automated classification of dog emotional states from facial expressions. *arXiv* **2022**, arXiv:2206.05619. [\[CrossRef\]](#)
18. Caeiro, C.; Guo, K.; Mills, D. Dogs and humans respond to emotionally competent stimuli by producing different facial actions. *Sci. Rep.* **2017**, *7*, 15525. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Hobkirk, E.R.; Twiss, S.D. Domestication constrains the ability of dogs to convey emotions via facial expressions in comparison to their wolf ancestors. *Sci. Rep.* **2024**, *14*, 10491. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Hara, K.; Kataoka, H.; Satoh, Y. Learning spatio-temporal features with 3D residual networks for action recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 3154–3160. [\[CrossRef\]](#)
21. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [\[CrossRef\]](#)
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 3–19.
23. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
24. Carreira, J.; Zisserman, A. Quo Vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
25. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
26. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A spontaneous micro-expression database: Inducement, collection and baseline. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–6. [\[CrossRef\]](#)

27. Qu, F.; Wang, S.J.; Yan, W.J.; Fu, X. CAS(ME)²: A database of spontaneous macro-expressions and micro-expressions. In *Human-Computer Interaction. Novel User Experiences. HCI 2016*; Kurosu, M., Ed.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9733. [[CrossRef](#)]
28. Ben, X.; Ren, Y.; Zhang, J.; Wang, S.-J.; Kpalma, K.; Meng, W.; Liu, Y.-J. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 5826–5846. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.