applied sciences

MDPI

Article

Attention-Based Spatiotemporal-Aware Network for Fine-Grained Visual Recognition

Yili Ren [1,2,†], Ruidong Lu [3,†], Guan Yuan [3,*], Dashuai Hao [3] and Hongjue Li [4,5,*]

[1] Research Institute of Petroleum Exploration and Development, Beijing 100083, China; renyili@petrochina.com.cn
[2] State Key Laboratory of Continental Shale Oil, Daqing 163002, China
[3] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; luruidong@cumt.edu.cn (R.L.); haodashuai@cumt.edu.cn (D.H.)
[4] School of Astronautics, Beihang University, Beijing 100191, China
[5] Shenzhen Institute of Beihang University, Shenzhen 518063, China
* Correspondence: yuanguan@cumt.edu.cn (G.Y.); lihongjue@buaa.edu.cn (H.L.)
† These authors contributed equally to this work.

Abstract: On public benchmarks, current macro facial expression recognition technologies have achieved significant success. However, in real-life scenarios, individuals may attempt to conceal their true emotions. Conventional expression recognition often overlooks subtle facial changes, necessitating more fine-grained micro-expression recognition techniques. Different with prevalent facial expressions, weak intensity and short duration are the two main obstacles for perceiving and interpreting a micro-expression correctly. Meanwhile, correlations between pixels of visual data in spatial and channel dimensions are ignored in most existing methods. In this paper, we propose a novel network structure, the Attention-based Spatiotemporal-aware network (ASTNet), for micro-expression recognition. In ASTNet, we combine ResNet and ConvLSTM as a holistic framework (ResNet-ConvLSTM) to extract the spatial and temporal features simultaneously. Moreover, we innovatively integrate two level attention mechanisms, channel-level attention and spatial-level attention, into the ResNet-ConvLSTM. Channel-level attention is used to discriminate the importance of different channels because the contributions for the overall presentation of micro-expression vary between channels. Spatial-level attention is leveraged to dynamically estimate weights for different regions due to the diversity of regions' reflections to micro-expression. Extensive experiments conducted on two benchmark datasets demonstrate that ASTNet achieves performance improvements of 4.25–16.02% and 0.79–12.93% over several state-of-the-art methods.

Keywords: micro-expression recognition; deep learning; spatiotemporal feature extraction; attention mechanism

Citation: Ren, Y.; Lu, R.; Yuan, G.; Hao, D.; Li, H. Attention-Based Spatiotemporal-Aware Network for Fine-Grained Visual Recognition. *Appl. Sci.* **2024**, *14*, 7755. https://doi.org/10.3390/app14177755

Academic Editor: Andrea Prati

Received: 16 May 2024
Revised: 18 June 2024
Accepted: 15 July 2024
Published: 2 September 2024

1. Introduction

Facial expressions convey a wealth of communicative information, often surpassing that conveyed through language and bodily cues. Numerous studies have contributed to facial expression recognition; however, many of these primarily focus on extracting coarse-grained macroscopic features of facial expressions, potentially overlooking subtle, brief facial information. In this paper, we take micro-expressions as an example to delve into the exploration of visual recognition methods based on fine-grained features.

When human beings hide real emotions, subtle changes will occur in facial muscles. These changes result in transient and involuntary facial expressions. A micro-expression has the characteristics of real emotional reflection and is difficult to forge. Therefore, it plays an important role in the fields of medical diagnosis [1], lie detection [2], business negotiation [3] and so on. Due to micro-expression's small range and short duration, it is difficult to be captured and recognized by the naked eye. Even after professional

training, the accuracy of recognition by human eyes is only 47% [4]. Therefore, in the era of artificial intelligence, it is of great significance to study how to capture and recognize micro-expressions with the help of fast visual capture equipment and artificial intelligence technology. At present, the study of micro-expression is focused on video stream data. The existing methods mainly include: LBP (Local Binary Pattern) [5], optical flow and methods based on Deep Learning [6].

LBP can describe the local texture features of the image and extract them, which has the advantages of gray invariance and rotation invariance. LBP usually extracts the appearance and motion features from the symbol-based difference between two pixels, so it will lose some information about features. In order to extract more useful information, the improved methods based on LBP consider the information of amplitude and direction, but this also leads to higher feature dimension after extraction. Most of these improved methods will reduce the processing speed and find it difficult to achieve real-time detection.

Optical flow methods use the relevance between adjacent frames in an image sequence to calculate the motion information of objects. These methods can detect small motion changes, and are suitable for recognizing micro-expressions. Optical flow methods extract geometric features from images, so they have a relatively low dimension of features and a relatively fast processing speed. However, the effect of optical flow feature extraction will be greatly reduced in the scenes with large changes in brightness.

Methods based on LBP or optical flow usually extract the shallow features of micro-expression images, so they cannot fully describe the structural information of samples and can hardly distinguish the relevance between high-dimensional features. Therefore, it is still a challenge to extract useful information and obtain high-quality description from micro-expressions.

For micro-expression recognition, feature extraction is an important critical issue [7], especially for optimal flow-based micro-expression recognition. The methods based on deep learning can extract the deep features from micro-expression images and recognize them [8]. Due to the change in a micro-expression being relatively small, it is hard to distinguish the effectiveness of features by directly transforming its RGB image into matrix and inputting it into our network for convolution. Therefore, LBP or optical flow is usually used to extract the image's low-dimensional features before inputting them into the network to extract the deep-dimensional features.

Due to the weak intensity of micro-expressions, directly inputting their RGB images into the network for convolution makes it difficult to distinguish effective features. Therefore, we use micro-expression videos as the model input. We employ a deep learning-based video motion magnification network to amplify micro-expression features and use a convolutional neural network (CNN) to estimate optical flow, extracting the optical flow information between adjacent frames in the image sequence.

Existing networks may overlook the correlation of image information in channel and spatial dimensions. To address this issue, we combine ResNet and ConvLSTM to extract spatiotemporal information, adding attention mechanisms between them. Our model effectively focuses on the micro-expression information across multiple dimensions, achieving superior recognition performance.

Micro-expression datasets are typically small and imbalanced, which can hinder the training of deep networks and lead to overfitting. To address these challenges, we employ data augmentation techniques, such as random rotation, to expand the effective dataset. We exclude classes with very few samples to improve training efficiency. The ResNet parameters are initialized with ImageNet pre-trained weights, and the model is further trained on the macro-expression dataset CK+. Additionally, we incorporate dropout layers into our network to selectively ignore some neurons and reduce their interactions in each training batch, helping to prevent overfitting.

Figure 1 shows our research framework. The input to the model consists of video sequences containing micro-expressions. First, facial detection and alignment are performed on each frame to ensure consistent facial feature points across all frames. Next, an action

magnification network is used to amplify micro-expression features, making them easier to capture. Subsequently, a convolutional neural network (CNN) processes the optical flow information between adjacent frames to capture subtle facial muscle movements. Specifically, the optical flow describes the motion patterns between two consecutive frames. Following this, ResNet is utilized to extract spatial features, and ConvLSTM is employed to capture temporal variations. By combining these two methods, richer feature representations are obtained. Additionally, both channel attention mechanisms and spatial attention mechanisms are adopted to achieve multidimensional focus on micro-expressions, enhancing the model's ability to perceive important features. Finally, a fully connected layer maps the extracted features to the classification space, and a classifier is used for categorization.

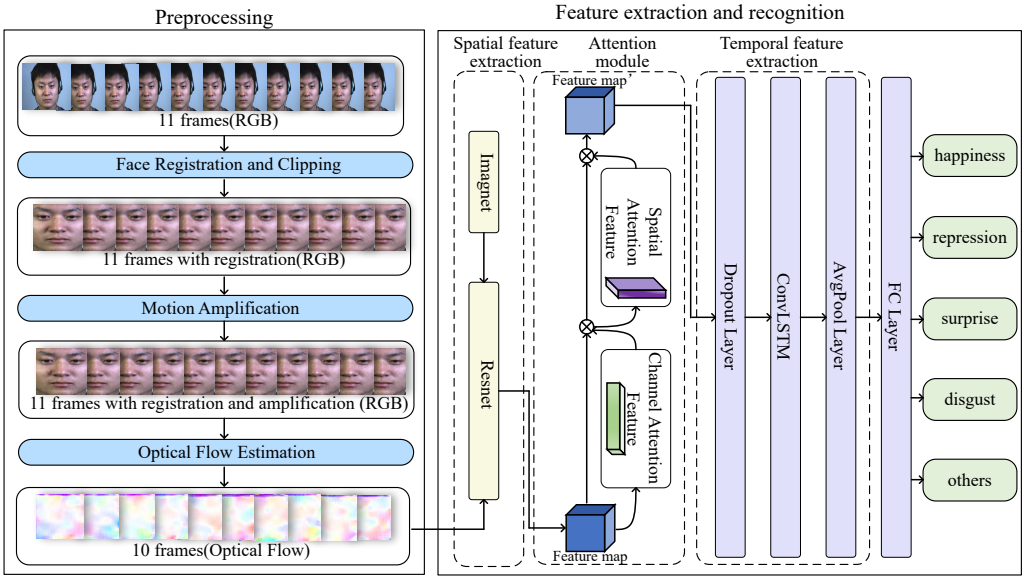


Figure 1. Research framework.

Overall, our contributions can be mainly summarized as follows:

- Instead of using traditional RGB images, we leverage video motion to recognize micro-expressions, which allows for better capture of expression dynamics.
- We combine spatiotemporal features with static features and introduce two types of attention mechanisms—channel attention and spatial attention—to model micro-expressions in a fine-grained manner from multiple dimensions.
- To address the challenge of limited micro-expression datasets, we employ effective data augmentation techniques such as rotation and flipping, improving the training process.
- We conduct extensive experimental validation, demonstrating that ASTNet significantly outperforms nine state-of-the-art open set recognition methods.

The structure of other parts of this paper is as follows: The second chapter briefly describes the existing feature extraction methods based on traditional or deep learning. The third chapter describes our processing method and proposes our network. The fourth chapter describes the experimental process, use parameters, results of experiments and the analysis of module effectiveness in detail. The last chapter summarizes our main work and contributions and looks forward to the possible direction of follow-up work.

2. Related Work

2.1. Micro-Expression Recognition Based on Optical Flow Features

Improved methods based on optical flow typically consider the displacement of feature regions rather than individual pixels, effectively recognizing subtle movements of objects and having relatively low feature dimensions, making them suitable for describing and processing small velocity changes in image motion [9]. Optical flow methods have

two assumptions: (1) The imaging surface is smooth and flat, and the brightness of the light illuminating the surface is uniformly distributed, meaning the image flow field is continuously differentiable in both spatial and temporal domains, with constant brightness in the time domain. (2) The variation in light reflectance is smooth and continuous, meaning the illumination on the object is continuous. Lucas et al. [10] used the least squares method to calculate the basic optical flow features of all pixels within a local region. Xu et al. [11] proposed the Facial Dynamics Map (FDM) based on optical flow, removing errors caused by facial translation and extracting the main directions of features in the spatiotemporal dimension, effectively reflecting the movement of micro-expressions. However, due to the long time required to compute dense optical flow fields, this method is unsuitable for real-time, large-scale micro-expression recognition.

Liu et al. [12] proposed the Main Directional Mean Optical Flow Feature (MDMO). In feature extraction, they divided the face into 36 non-overlapping regions and extracted optical flow features frame by frame, extracting the main direction in each region. Khor et al. [13] used the TV-L1 [14] method to extract optical flow and optical strain, which they combined with the original image and input into a CNN-LSTM network to highlight the dynamic changes of micro-expressions. Another approach involves changing the network structure, such as the Three-Dimensional Convolutional Neural Network Evolution (C3DEvol) proposed by Liang Zhengyou et al. [15], which simultaneously extracts spatiotemporal features of micro-expressions. This network uses global search and genetic algorithms to optimize the C3D network structure, improving micro-expression recognition performance. Compared to CNN-LSTM, this method can simultaneously extract spatiotemporal features of micro-expressions, but it still does not fully explore comprehensive micro-expression feature information, resulting in less-than-ideal recognition performance.

2.2. Micro-Expression Recognition Based on Action Units

The key to improving micro-expression recognition lies in capturing and extracting more comprehensive and discriminative micro-expression information. Besides the spatiotemporal dimensions of micro-expression information, the movement changes of facial muscles during micro-expressions are also related. Therefore, analyzing micro-expressions through the combination of action units (AUs) is an effective method.

For instance, Lo et al. [16] used a 3D convolutional neural network to extract AU-related features and then utilized a graph convolutional network to connect AU nodes with complex dependencies for emotion classification. This method focuses on local action changes but employs a model with a large number of parameters. The designed adjacency matrix is calculated based on the co-occurrence rate of AUs in the training set, making it susceptible to noisy actions and lacking robustness.

Similarly, Xie et al. [17] proposed AGACN, which combines AU information with micro-expression labels, modeling AUs in different regions based on facial muscle deformation and relational information. Ling et al. [18] represented micro-expressions based on partial facial key points and used depthwise convolution (DConv) to learn key point features, or node learning, and employed the Transformer encoder to learn relationships between nodes, or edge learning. The AU-GCN module in this method constructs and learns the AU feature matrix and adjacency matrix through word embedding and GCN. This method combines facial features with AU information, yielding good recognition results.

Puneet et al. [19] proposed the MERASTC method, which encodes the subtle changes in facial muscles in micro-expression image sequences by combining AU, key points, and appearance features during micro-expressions. They also proposed a face normalization method based on neutral features to accelerate the network's micro-expression recognition efficiency. Although MERASTC alleviates the network overfitting issue, it requires high sample data quality (the image sequence must contain neutral frames), thus having certain limitations.

Wang et al. [20] combined AU with facial key points to construct four regions of interest: eyes, nose, cheeks and mouth. They then weighted the corresponding regions

and extracted features using the proposed MER-AMRE network, improving the network's ability to extract local motion information.

To explore the relationship between facial expressions and AUs, Cen et al. [21] segmented micro-expression videos into multiple adjacent video clips, revealing the spatiotemporal feature changes in the three-dimensional neighborhood. They combined this with the proposed Multi-Task Facial Activity Pattern Learning Framework (MFAPLF) to promote the aggregation of similar micro-expressions.

3. Our Method

Firstly, we enlarge the features of the micro-expression and extract the optical flow features to improve the inter-class discrimination and enhance the micro-expression features. Then, we design a network structure, ASTNet, to further extract spatiotemporal features and identify emotions. In ASTNet, ResNet and ConvLSTM process the spatial and temporal features of the optical flow image, respectively. Finally, we add channel and spatial attention (CBAM) between them to generate the weight map of the feature map from the two dimensions of space and channel for feature adaptive learning. Figure 2 shows the structure of our network.

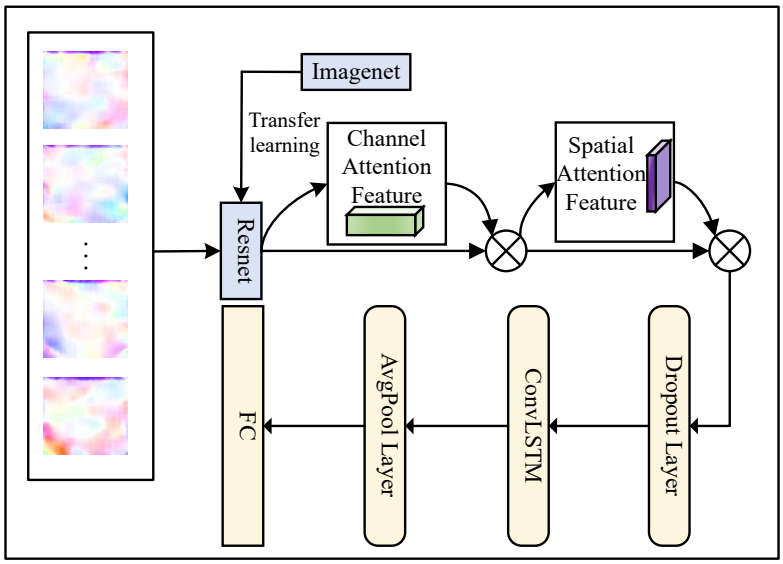


Figure 2. Framework of our network.

3.1. Processing

(1) Micro-expression motion amplification

Micro-expressions change subtly, which leads to low characteristic intensity and low category discrimination. Therefore, we use the motion amplification technology to properly amplify these subtle dynamic changes and improve the discrimination of emotions. Some methods of video action amplification can be used to complete this part of the work. Their working principles are as follows: First they use Euler Video Amplification [22] to decompose the input continuous segments into different spatial frequency bands. Then, they use the same time filtering for processing and give an amplification factor to amplify the signal of interest. Finally, they superimpose it on the original signal as output; the mathematical description is as follows: Let $I(x,t)$ be the image intensity of the image at time t and position x, and use the displacement function to express the observed intensity $\delta(t)$, that is, $I(x,t) = f(x+\delta(t)))$ and $I(x,0) = f(x)$. After amplifying the image signal α times, it is superimposed back to the original image to obtain the synthesized image signal $I'(x,t)$, as shown in Equation (1):

$$I'(x,t) = f(x + (I+\alpha)\delta(t)) \quad (1)$$

This paper focuses on the amplification of facial micro-expression movements, setting the amplification factor z to 2, 3 and 4 for comparison. Figure 3 shows the amplification effects of three types of emotions in the CASME II dataset, with each set of experiments using the same frame comparison.

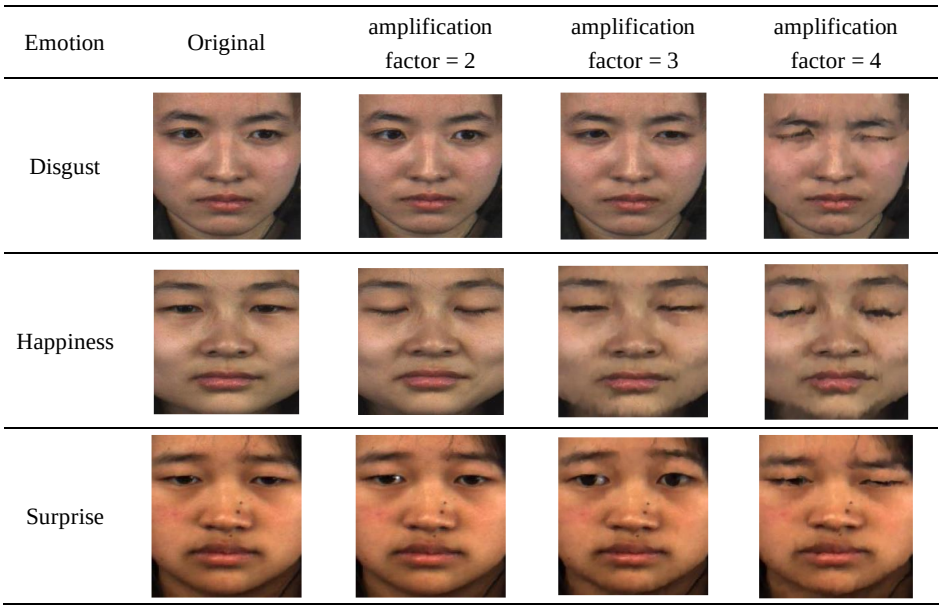


Figure 3. Comparison of before and after micro-expression magnification.

In the emotion category labeled as disgust, as the amplification factor increases, the muscle movements in the eyebrow area of the subjects become more noticeable, and the activation level of the frowning action increases. In the happiness category, the muscle movements in the mouth area of the subjects become more noticeable, and the activation level of the mouth corner lifting action increases. In the surprise category, the muscle movements in the eyebrow and mouth areas are generally amplified to varying degrees compared to other muscle regions.Additionally, it can be observed that in the happiness category, some subjects exhibit eye closure, while in the surprise category, the subjects' eyeballs are amplified. This is because the network enhances eye movements such as pupil changes and blinking through temporal information from consecutive frames. However, these eye movements do not conform to the definition of micro-expressions and do not contribute to their recognition.From the amplification results of the three emotions, although image noise increases when the amplification factor is set to 3, the amplification effect on motion is generally more noticeable than when the factor is set to 2. When the amplification factor is set to 4, facial muscles start to distort excessively, and image noise increases significantly. Considering these factors, this paper adopts an amplification factor of 3 for the network to amplify the actions of micro-expression samples.

We magnify images of micro-expression in SMIC and CASME II by 3 times through the Motion Amplification Network [23]. Figure 4 shows the comparison between the images before and after processing.

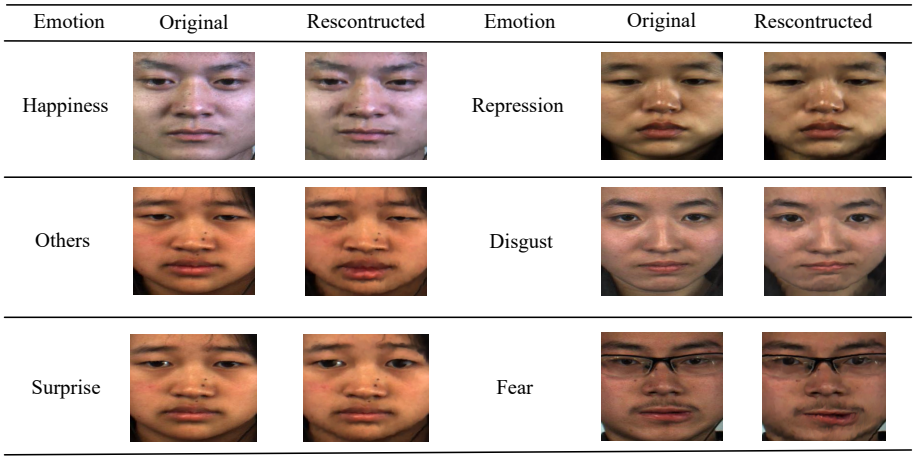


Figure 4. Comparison of before and after micro-expression magnification.

(2) Optical flow feature extraction

In micro-expression recognition, feature extraction is a critical issue. Micro-expressions are brief, involuntary facial expressions, and directly using RGB images of micro-expressions as network input to extract deep features cannot effectively recognize emotions [24]. Therefore, extracting effective features is essential [25].

In this paper, we employ an optical flow-based feature extraction method due to the superior performance of optical flow in handling dynamic information and subtle changes, which is crucial for micro-expression recognition. Micro-expressions are often difficult to detect, and the optical flow method can provide detailed motion vectors, aiding in the more accurate recognition of micro-expressions.

Therefore, we use features in the form of optical flow as the input to the network and model them using the Motion Constraint Equation [22] of optical flow.

Specifically, for the RGB keyframe sequence, $I(x, y, t)$ represents the light intensity of the pixel at position (x, y) and time t. The pixel moves a distance of (dx, dy) to the next frame over a time interval of dt. Since it is the same pixel, we assume that the light intensity of the pixel remains unchanged before and after the movement. It is shown in Equation (2):

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{2}$$

The flow field is continuously differentiable in both the space and time domain. According to Taylor series expansion, Equation (2) can be expanded, as shown in Equation (3):

$$I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \xi \tag{3}$$

ξ is the second-order or above estimator of time dt. When dt tends to infinity, the optical flow constraint equation can be obtained by combining Equations (2) and (3), as shown in Equation (4):

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t}\frac{dt}{dt} = 0 \tag{4}$$

Then, the corresponding optical flow vector calculation equation is shown in Equation (5):

$$\mathrm{V} = \left[u = \left(\frac{dx}{dt}\right), v = \frac{dy}{dr}\right]^{\mathrm{T}} \tag{5}$$

Figure 5 shows the use of LiteFlowNet [26] to extract the optical flow features between 11 frames of 'happiness' and visualize them. In the above optical flow images, different

colors represent different directions of motion, while different depths of color represent different intensities of motion.

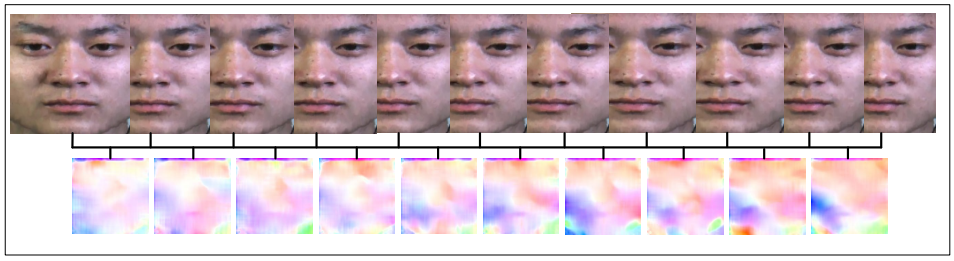


Figure 5. Optical flow visualization.

3.2. Network

ASTNet consists of three parts: ResNet for extracting spatial features of the micro-expression image sequence, an attention mechanism module for extracting attention maps and ConvLSTM for processing temporal features and recognizing emotions. ResNet can train very deep neural networks and extract deeper, fine-grained spatial features of micro-expression image sequences. The attention mechanism calculates weighted features through attention scores, allowing for a better focus on key features during the micro-expression change process. Since we use micro-expression video sequences that vary over time, temporal modeling is necessary. ConvLSTM excels in capturing spatiotemporal relationships and can simultaneously model spatial relationships to recognize facial emotions. In the following sections, we will detail each component of the model.

(1) ResNet

The first part of our network uses the transfer of learning; we initialize the parameters of ResNet by using ImageNet, and further train on the macro-expression dataset: CK+. Finally, we extract the spatial features of the optical flow image by using the processed ResNet.

ResNet was proposed by He Kaiming, Ren Shaoqing et al. [27]. They added a residual structure in their network to solve the degradation problem of the poor stacking effect when they deepened the depth of the network to a certain extent. ResNet also solved the problem of gradient disappearance or explosion of the deep network through the identity mapping method. Figure 6 shows the residual block.

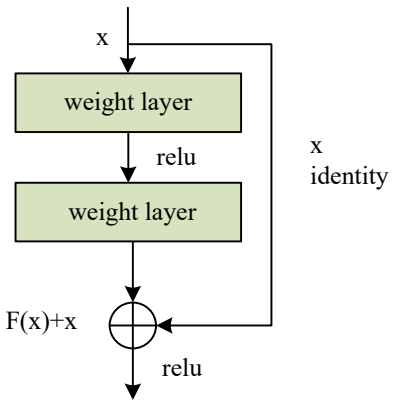


Figure 6. Residual block structure of ResNet.

We remove the full connection layer of ResNet and directly output the convoluted and pooled graph of features as the input of the next part of our network.

(2) Attention Mechanism

Each region or visual feature on micro-expression images has different importance for the recognition of the network. Therefore, in order to establish the correlation of the pixels

of a micro-expression image in the spatial and channel dimensions, we use CBAM [28] to construct the attention mechanism. CBAM is a feedforward convolutional neural network attention module that can be integrated into CNN. CBAM has two sequential submodules: channel and space. The feature map of the network convolution block can be adaptively refined through CBAM. Figure 7 shows CBAM's structure.

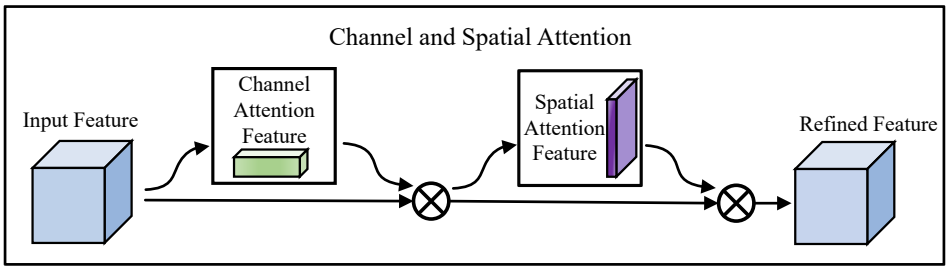


Figure 7. CBAM attention module.

Given a feature graph $F \in C \times H \times W$ as input, CBAM is successively transformed into a one-dimensional channel attention map $M_c \in R^{C\times1\times1}$ and a two-dimensional spatial attention map $M_s \in R^{1\times H\times W}$.

The channel attention module uses the channel relationship between features to generate the channel attention map. Firstly, the channel attention module uses the average pooling and maximum pooling to aggregate the spatial information of the feature graph, and generate average pooling F^c_{avg} and maximum pooling F^c_{max}. Then, it transmits F^c_{avg} and F^c_{max} to a shared network layer composed of multi-layer perceptrons to generate the channel attention graph $M_c \in R^{C\times1\times1}$. Finally, the feature vectors are merged. The channel attention model structure is shown in Figure 8.

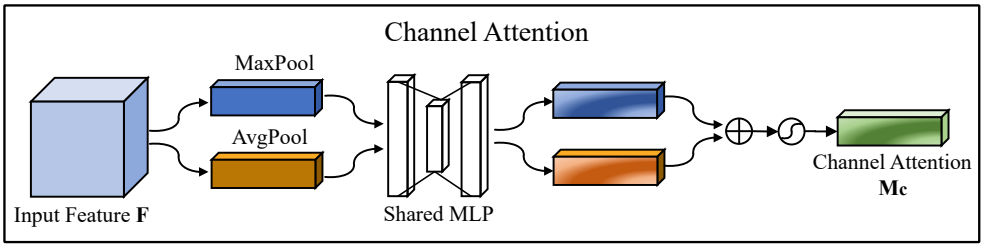


Figure 8. Channel attention module.

The spatial attention module uses the spatial relationship between features to generate a spatial attention map. Firstly, the average pooling and maximum pooling operations are used to aggregate the channel information of the characteristic graph along the channel direction to generate two characteristic graphs: $F^s_{avg} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$. Then, a standard convolution layer is used to connect and convolute them to generate a 2D spatial attention map. The structure of spatial attention model is shown in Figure 9.

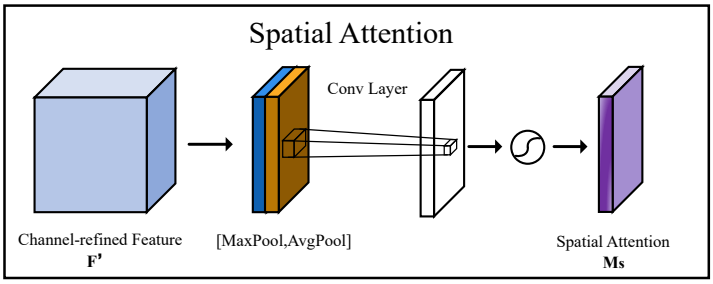


Figure 9. Spatial attention module.

The channel and spatial attention mechanisms are arranged in order. According to the experimental results, we prioritize the channel attention module, and then arrange the spatial attention module.

(3) ConvLSTM

The optical flow features of the micro-expression image sequence have strong temporal correlation. Thus, we use ConLSTM [29] to explore the temporal information from adjacent frames and retain the spatial information.

LSTM [30] has a good recognition effect for a video stream, but when processing spatiotemporal data, LSTM uses a complete connection to model the sequence information in the transformation of input–state and state–state, and then flattens the input image into a one-dimensional vector. Thus, LSTM does not encode the spatial information, resulting in the loss of spatial information. ConvLSTM uses convolution to replace matrix multiplication to act on the conversion between input and state and state and state. Therefore, ConvLSTM can retain the spatial information of the sequence. The structure of ConvLSTM is shown in Figure 10.

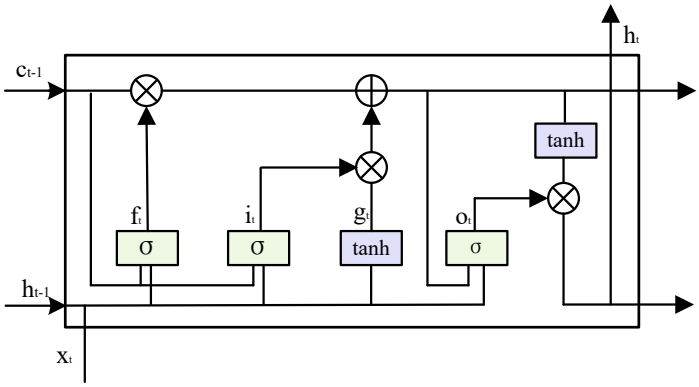


Figure 10. The structure of ConvLSTM.

ConvLSTM combines the advantages of CNNs and LSTMs, enabling the capture of spatial features while modeling the dynamic variations in time-series data. ConvLSTM controls the flow of information through three gate mechanisms (input gate, forget gate and output gate). The computations of these gates involve convolution operations instead of the fully connected operations in traditional LSTMs: Input Gate: Controls the impact of input data on the cell state. Forget Gate: Determines how much of the previous time step's cell state is forgotten. Output Gate: Regulates the influence of the current cell state on the hidden state. The calculation Equations for each gate are detailed in Equations (6)–(8):

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f) \tag{7}$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o) \tag{8}$$

where σ denotes the sigmoid function, $*$ represents the convolution operation, X_t is the input at the current time step, H_{t-1} is the hidden state from the previous time step, and W and b are the weights and biases, respectively. The process for updating the cell state is shown in Equations (9) and (10):

$$\widetilde{C}_t = tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{9}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C}_t \tag{10}$$

where $\circ$ denotes element-wise multiplication, *tanh* represents the hyperbolic tangent function, $\widetilde{C}_t$ is the new candidate cell state and C_{t-1} is the cell state from the previous time step. The update for the hidden state is calculated as Equation (11):

$$H_t = o_t \circ tanh(C_t) \tag{11}$$

Through the output gate control, the current cell state C_t is activated by the Tanh function to obtain the current hidden state H_t. ConvLSTM outputs the current cell state C_t and hidden state H_t at each time step. C_t serves as the processing result for the current time step, while H_t serves as the input for the next time step.

Therefore, combining the above three modules, our network model is shown in Table 1.

Table 1. Overall network structure.

Network Layer	Our Network
Layer 1	Conv0_1, BN1, RuLU, MaxPool
Layer 2	Conv1_1, BN1, ReLU, Conv1_2, BN2 Conv1_3, BN1, ReLU, Conv1_4, BN2
Layer 3	Conv2_1, BN1, ReLU, Conv2_2, BN2 Conv2_3, BN3 Conv2_4, BN1, ReLU, Conv2_5, BN2
Layer 4	Conv3_1, BN1, ReLU, Conv3_2, BN2 Conv3_3, BN3 Conv3_4, BN1, ReLU, Conv3_5, BN2
Layer 5	Conv4_1, BN1, ReLU, Conv4_2, BN2 Conv4_3, BN3 Conv4_4, BN1, ReLU, Conv4_5, BN2
Layer 6	CBAM, Dropout
Layer 7	ConvLSTM, AvgPool, FC

4. Experiment and Analysis

4.1. Datasets and Preprocessing

Considering the frame rate, emotion type and use frequency of each dataset, we use SMIC [31] and CASME II [32] for the experiments. SMIC contains 164 micro-expression samples of 16 subjects, which are divided into positive, negative and surprised. Among them, the positive emotion is happiness, and negative emotions include sadness, fear and disgust; CASME II contains 255 video samples from 26 participants, which are divided into seven emotions: happiness, surprise, repression, disgust, sadness, fear and others. Table 2 summarizes the characteristics of the two datasets.

Table 2. Characteristics of datasets.

Characteristics	SMIC	CASME II
Samples	164	255
Subjects	16	26
Ethnicities	3	1
FPS	100	200
Resolution	640 × 480	640 × 480
Facial Area	160 × 130	340 × 280
Emotion Classes	3	7
Objective Classes	-	7
FACS Coded	NO	YES

In CASME II, the number of samples of sadness and fear is less than 10. In order to avoid the imbalance of data distribution affecting the recognition effect, we ignore these two emotions in the experiment. Through image operations including mirror, random rotation and clipping, we expand the datasets 20 times. The distribution of original and expanded samples is shown in Table 3.

Table 3. Sample distribution of datasets.

Emotions	SMIC	Processed	Emotions	CASME II	Processed
			Happiness	32	640
Positive	51	1020	Surprise	25	500
			Disgust	63	1260
Negative	70	1400	Repression	27	540
			Sadness	7	-
Surprise	43	860	Fear	2	-
			Others	99	1980
Total Size	164	3280	Total Size	255	4920

The number of sample frames in the micro-expression datasets is not uniform. Therefore, we extract the continuous frames from the apex to the peak and fix the length as 11 frames. We preprocess each sample sequence by using face detection and alignment, facial motion feature amplification and optical flow feature extraction, as follows:

(1) Face detection and alignment

Firstly, we select the first frame image in each sample, and use ASM [33] to extract 68 facial key points; then, we use the Local Weighted Average [34] to transform these facial key points into matrices, and normalize and align all frames of the sample with these matrices. Finally, we locate the left eye coordinates, determine the distance between the two eyes and cut all frames of the samples.

(2) Facial motion feature amplification

We use the video action amplification network based on deep learning to amplify the features of micro-expression. According to our experiment, when the magnification is too small, the micro-expression features are not obvious enough; when the magnification is too large, it will cause too much noise interference and distort the micro-expression image. After many experiments, we determine that the amplification factor is 3.

(3) Optical flow extraction

We use LiteFlowNet to extract optical flow features from two adjacent micro-expression images in each sample. After a pair of images are processed through this network, an optical flow file (.flo) is output and visualized. Therefore, one sequence of 11 frames can be turned into an optical flow sequence of 10 frames.

4.2. Experimental Environment and Parameter Setting

The operating system of our experiments is ubuntu-16.04 and the development tool we used is Pycham. Our experiments are accelerated by a Tesla p100-pcie GPU and we selected the network parameters with the best effect after many experimental comparisons, as shown in Table 4.

Table 4. Hyperparameters of our network.

Parameter	Figure
epoch	50
batch_size	32
base_lr	1×10^{-5}
decay	1×10^{-7}
momentum	0.9
random deactivation rate	0.5

4.3. Detail of Experiment and Analysis

Our experiments use LOSO (Leave-One-Subject-Out) as the evaluation standard and use the accuracy and F1-score as the performance indexes of micro-expression recognition.

By reserving a specific participant sample as the test sample and using other participant samples as the training sample, we divide the micro-expression recognition accuracy of all participants equally. The final accuracy is $(\sum_{i=1}^{n} c_i)/n$, and the F1-score is $(\sum_{i=1}^{n} f_i)/n$.

The calculation equation of accuracy is shown in Equation (12):

$$\text{Accuracy } = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{12}$$

The F1-score is defined as Equation (13):

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \\ F1 = \frac{2*P*R}{P+R} \end{cases} \tag{13}$$

We compared four traditional methods and three methods based on deep learning with excellent performance in the past few years on the SMIC and CASME II datasets. Our method achieved the best results in accuracy and F1-score among these methods, which shows that it has significant advantages in micro-expression recognition. The specific experimental results are shown in Table 5.

Table 5. Comparison of experimental results.

Methods	SMIC		CASME II	
	ACC (%)	**F1-Score**	**ACC (%)**	**F1-Score**
LBP-TOP [35]	43.38	0.34	46.46	0.42
Bi-WOOF+Phase [36]	68.29	0.67	62.55	0.65
FDM [11]	54.88	0.54	45.93	0.40
Sparse MDMO [37]	70.51	0.70	66.95	0.69
DSSN [38]	63.41	0.65	70.78	0.73
GEME [39]	65.24	0.61	75.02	0.73
OFF-Apex Net [40]	67.68	0.67	68.94	0.70
ASTNet	**73.57**	**0.72**	**75.61**	**0.74**

Bold text indicates the best results.

The proposed method outperforms other methods in recognition performance on both the CASME II and SMIC datasets. This is because our proposed method does not use traditional RGB images but instead utilizes video motion to recognize micro-expressions, thereby better capturing the dynamics of expressions. Additionally, we combine spatiotemporal characteristics with static features, modeling micro-expressions from multiple perspectives in a fine-grained manner.

DSSN captures micro-expressions from different angles but only focuses on micro-expression changes at the image level without considering optical flow features and dynamic changes. MDMO divides the face into 36 non-overlapping regions and extracts optical flow features frame by frame. However, it only applies optical flow features, neglecting the continuous spatiotemporal changes in micro-expressions, resulting in less effective performance compared to our model.

4.4. Hyperparameter Optimization

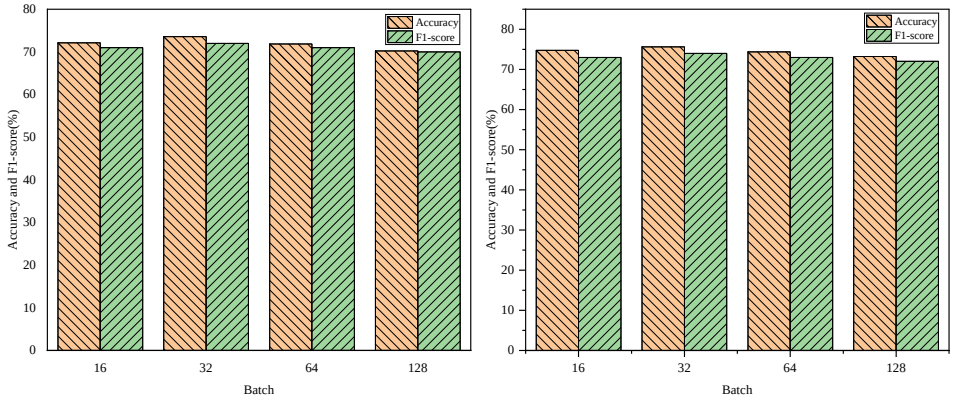
First, we compare and experiment with different learning rate settings. Due to the weak intensity of micro-expression features, the initial learning rate of the network should be relatively low. In this experiment, we set the learning rates to 1×10^{-3}, 1×10^{-4} and 1×10^{-5} and tested them on the CASME II dataset. When the learning rate was 1×10^{-3}, the network training experienced oscillations, and the accuracy could not improve steadily. This was because the excessively high learning rate prevented the network from reaching local minima. When the learning rate was 1×10^{-5}, the network's fitting speed was too

slow, resulting in a waste of time. Therefore, considering these factors, the learning rate was set to 1×10^{-4} for the experiment.

When training the network, the initial weights need to be initialized according to a certain distribution to ensure faster convergence of the loss function during training, thereby achieving better optimization results. However, when initializing network weights randomly according to a certain distribution, inappropriate initial weights might cause the loss function to get stuck in local minima during training, preventing it from reaching a global optimum. Momentum can partially solve this problem. The larger the momentum, the more energy is converted into potential energy, making it more likely to escape the confines of local concave regions and enter global concave regions. The most common setting for momentum is 0.9.

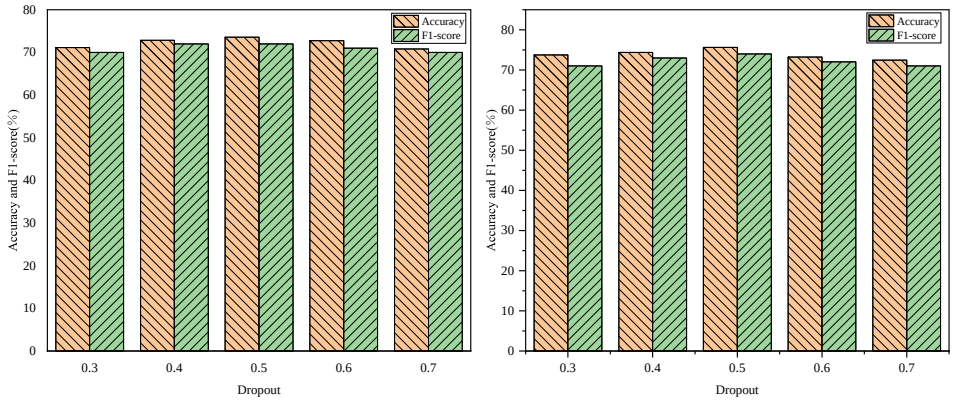
Other parameter settings are as follows: the optimizer used is Stochastic Gradient Descent (SGD), and lr_scheduler is used to automatically update the learning rate. We compared batch sizes of 16, 32, 64 and 128, and dropout rates of 0.3, 0.4, 0.5, 0.6 and 0.7.

Figures 11 and 12 show the impact of Batch Size and Dropout on network recognition, respectively. When the batch is selected as 32 and the dropout parameter is set to 0.5, the recognition of our network is the best. Dropout prevents overfitting by discarding some features and improves the generalization ability of the network. However, when the parameter is set too large, too much effective information is lost, which affects the learning ability of our network; while when its parameter is set too small, there are too many redundant neurons to extract enough effective features.



(**a**) SMIC (**b**) CASME II

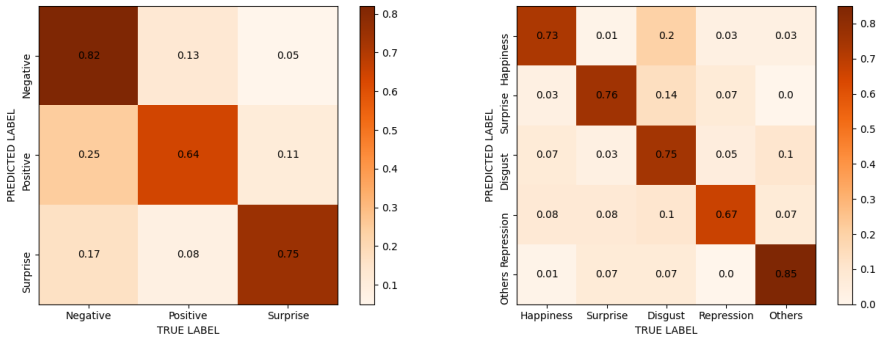
Figure 11. Influence of different batch sizes on our network.



(**a**) SMIC (**b**) CASME II

Figure 12. Influence of different dropouts on our network.

The confusion matrices on SMIC and CASME II are shown in Figure 13. It can be seen from Figure 13 that among the three emotions identified in the SMIC, 'negative' has the highest accuracy, followed by 'surprise', and 'positive' has the lowest accuracy. The possible reason is that the number of 'negative' samples is the largest and it has high discrimination from other emotions. Among the five emotions identified in the CASME II, the accuracy of 'others' is the highest, while the accuracy of 'repression' is the lowest. The reason is that the number of 'others' is the largest, 'repression' has a small number of samples and the class differentiation between other emotions is small. While the number of samples is also small, the accuracies of 'disgust' and 'surprise' are better than that of 'repression', which is because the former two are highly distinguished from other emotions. It can be seen that the number of samples and the intensity of features can have a great impact on the recognition.



(**a**) SMIC (**b**) CASME II

Figure 13. Confusion matrix.

4.5. Ablation and Analysis

To validate the effectiveness of each module, we conducted ablation experiments, testing the effectiveness of the action amplification module, the spatial attention mechanism and the channel attention mechanism. We generated four variants: the ResNet-ConvLSTM network without the action amplification module and both attention mechanisms. ResNet-ConvLSTM with the action amplification module but without both attention mechanisms. The ResNet-SpatialAttention-ConvLSTM network without the channel attention module. The ResNet-ChannelAttention-ConvLSTM network without the spatial attention module. We compared the results of these variants with our ASTNet.

Table 6 illustrates the effectiveness of micro-expression amplification and compares the improvement of different attention modules on the network. It can be seen from Table 6 that although action amplification increases the noise of the micro-expression image and causes a certain degree of image distortion, it has a significant amplification effect on micro-expression features. On the whole, its recognition effect on the micro-expression has been improved. ResNet+ConvLSTM has more degradation without the attention mechanism, which indicates that the importance of pixels in different regions and different channels of the micro-expression image is different and verifies the effectiveness of our attention mechanism. Compared with spatial attention and channel attention, CBAM improves the network recognition effect more significantly. At the same time, it can be seen that the spatial attention mechanism improves the network recognition effect less than the channel attention mechanism. This is because the convolution of ConvLSTM is used for conversion between input and state and state and state, which can better capture the spatial information of images than LSTM. Therefore, it can be considered that there is an overlap between the spatial attention mechanism and the convolution of ConvLSTM.

We also try to add CBAM to different locations within ResNet to explore the impact of location and structure on the CBAM module computing attention map and improve the effect of the network model. As shown in Table 7, in our proposed network model, ResNet is composed of five middle layers. We integrate CBAM modules between each two middle

layers to form four different networks. In Table 8, we show the effect of these networks on recognizing micro-expressions and analyze the possible reasons. The results show that they improve the accuracy of recognition less than adding CBAM between ResNet and ConvLSTM. The results show that although integrating the attention mechanism into different locations within the ResNet will improve the recognition rate of our network, their improvement is less than that of integrating CBAM between ResNet and ConvLSTM. It is difficult to find the law of the effect of the attention mechanism integrated into the network with the change in location, but they will affect the effect of transfer learning to a certain extent, so the overall improvement effect is not as good as the effect of integrating attention outside the network.

Table 6. Effectiveness analysis of amplification and attention models.

Methods	SMIC	CASME II
	ACC (%)	ACC (%)
[1] ResNet-ConvLSTM	64.21	66.83
[2] ResNet-ConvLSTM	69.59	71.54
[2] ResNet-Spatial Attention-ConvLSTM	71.83	73.96
[2] ResNet-Channel Attention-ConvLSTM	72.27	74.35
[2] ASTNet	**73.57**	**75.61**

[1] indicates no action amplification. [2] indicates action amplification. Bold text indicates the best experimental results.

Table 7. Network structure of CBAM integrated into different layers of ResNet.

	ResNet in Our Model	Fusion Position of Attention
Layer1	Conv0_1, BN1, RuLU, MaxPool	
		CBAM(1)
Layer2	Conv1_1, BN1, ReLU, Conv1_2, BN2 Conv1_3, BN1, ReLU, Conv1_4, BN2	
		CBAM(2)
Layer3	Conv2_1, BN1, ReLU, Conv2_2, BN2 Conv2_3, BN3 Conv2_4, BN1, ReLU, Conv2_5, BN2	
		CBAM(3)
Layer4	Conv3_1, BN1, ReLU, Conv3_2, BN2 Conv3_3, BN3 Conv3_4, BN1, ReLU, Conv3_5, BN2	
		CBAM(4)
Layer5	Conv4_1, BN1, ReLU, Conv4_2, BN2 Conv4_3, BN3 Conv4_4, BN1, ReLU, Conv4_5, BN2	

CBAM indicates that the attention mechanism is added here.

Table 8. Comparison of adding CBAM in different locations.

	ResNet+ConvLSTM	Network(1)	Network(2)
SMIC	-	↑3.59%	↑3.52%
CASME II	-	↑3.64%	↑3.79%
	Network(3)	**Network(4)**	**Proposed**
SMIC	↑3.64%	↑3.59%	↑**3.98%**
CASME II	3.62%	↑3.61%	↑**4.07%**

↑ indicates the improvement compared with '-'. Bold text indicates the best experimental results.

5. Application

The ASTNet model has been integrated into a multimodal teaching behavior intelligent analysis system for smart classrooms (This system is designed based on the open source smart classroom accessed on 1 October 2023 (https://github.com/hongyao-hongyao/smart_classroom).), aiming to enhance teaching behavior through advanced technology. This innovative system utilizes video data to perform detailed analysis of students' micro-expressions in the classroom and accurately perceive collective emotions. This emotional insight assists teachers, enabling them to dynamically adjust teaching methods to improve student engagement and learning efficiency. The system performs real-time detection and analysis of classroom emotions through three aspects: attention, emotional state and facial expressions.

The specific application scenario is a classroom with a frame width of 640 pixels, height of 480 pixels and frame rate of 30 frames per second, ensuring high-quality data acquisition and analysis. Figure 14 shows a screenshot of the prototype system, which we now introduce: Firstly, the system uses three specific metrics to evaluate micro-expressions: focused emotion, group emotional state and expressed emotion. The three graphs on the right represent the score curves of these metrics, with the vertical axis showing the metric scores. Higher scores indicate higher levels of focused emotion, group emotional state and expressed emotion. The horizontal axis represents time, showing the emotional values at each time point.

The classroom group emotion perception curve below represents the overall emotional value of the classroom. The radar chart in the lower left corner shows the ratio of occurrences of each behavior to the total number, used as weights. By weighting and summing each behavior, the classroom group emotion value is obtained.

The system interface also includes real-time video data captured during the class, with the video capture time displayed in the upper left corner. The bounding boxes in the video indicate identified students, and the number above each box represents the individual student's emotion value. These emotion values are crucial for teachers, as higher emotion values indicate better student engagement, allowing teachers to adjust their teaching methods accordingly.

The contribution of this application is significant. By providing real-time, detailed insights into students' emotions, the system enables teachers to create more responsive and effective learning environments. This proactive educational approach not only improves individual student performance but also enhances the overall classroom atmosphere, fostering a more supportive and engaging learning experience.

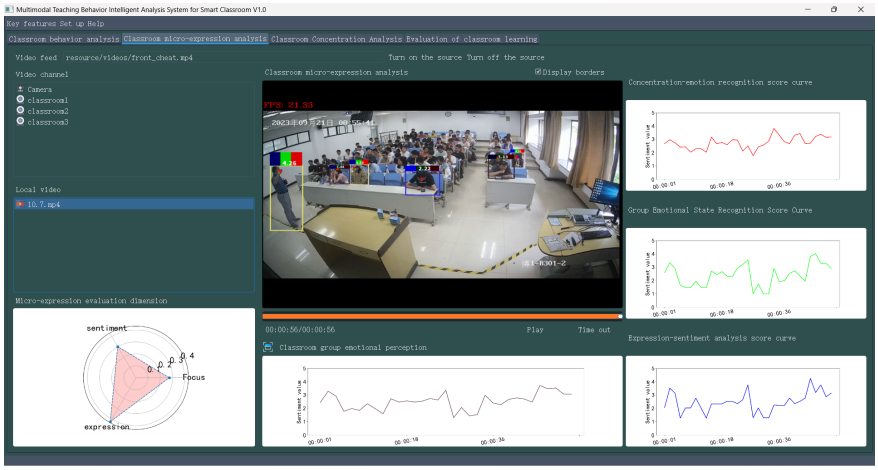


Figure 14. The multimodal teaching behavior intelligent analysis system for smart classrooms.

6. Conclusions

We propose a network combining ResNet with ConvLSTM and integrate an attention mechanism, which achieve a good result in the task of micro-expression recognition. In our model, ResNet extracts the features in the spatial domain, ConvLSTM processes the features in the temporal domain and the CBAM model integrated between ResNet and ConvLSTM make our network pay more attention to the effective information of images and improve our network's representation ability of the feature channel and space. We also verify the effectiveness of the attention mechanism in the proposed network and compare the ability of the spatial attention mechanism, channel attention mechanism and CBAM to improve the recognition of our network. Finally, we try to explore the impact of the position of CBAM in the network on the improvement of the recognition effect. In the follow-up work, we can try to further optimize the network structure and explore the different combinations of attention mechanism and ConvLSTM, so as to mine the micro-expression features more effectively and further improve the recognition effect.

Author Contributions: Data collection, D.H. and R.L.; conceptualization, Y.R. and H.L.; methodology and formal analysis, Y.R. and R.L.; resources, H.L. and G.Y.; data curation, D.H. and Y.R.; experiments, Y.R., R.L. and D.H.; writing—original draft preparation, Y.R.; writing—review and editing, R.L., H.L. and G.Y.; visualization, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Xuzhou K&D Program, grant number KC23296. This work was also supported in part by the Guangdong Basic and Applied Basic Research Foundation, grant number 2022A1515110837. This work was also supported in part by the general program of the NSFC intelligent thin section identification method for an oil and gas reservoir based on knowledge and data fusion, grant number 42372175. This work was also supported in part by the CNPC Innovation Foundation, grant number 2021DQ02-0904.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from the Center for Machine Vision and Signal Analysis (dataset SMIC) and State Key Laboratory of Brain and Cognitive Science (dataset CASME II) and are available from Xiaobai Li (Xiaobai.Li@oulu.fi dataset SMIC) and accessed on 23 December 2021 http://casme.psych.ac.cn/casme/c2 (dataset CASME II) with the permission of the Center for Machine Vision and Signal Analysis and State Key Laboratory of Brain and Cognitive Science.

Acknowledgments: Thanks to Guangdong Basic and Applied Basic Research Foundation, the general program of NSFC intelligent thin section identification method for an oil and gas reservoir based on knowledge and data fusion, and the CNPC Innovation Foundation for funding this study.

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

References

1. Zhu, C.; Chen, X.; Zhang, J.; Liu, Z.; Tang, Z.; Xu, Y.; Zhang, D.; Liu, D. Comparison of ecological micro-expression recognition in patients with depression and healthy individuals. *Front. Behav. Neurosci.* **2017**, *11*, 199. [CrossRef]
2. Gan, Y.; Liong, S. Bi-directional vectors from apex in cnn for micro-expression recognition. In Proceedings of the 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, China, 27–29 June 2018; pp. 168–172.
3. Collins, R.; Visual micro-sociology and the sociology of flesh and blood: Comment on Wacquant. *Qual. Sociol.* **2015**, *38*, 13–17. [CrossRef]
4. Warren, G.; Schertler, E.; Bull, P. Detecting deception from emotional and unemotional cues. *J. Nonverbal Behav.* **2009**, *33*, 59–69. [CrossRef]
5. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, 9–13 October 1994; pp. 582–585.
6. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]

7. Takalkar, M.; Xu, M.; Wu, Q.; Chaczko, Z. A survey: Facial micro-expression recognition. *Multimed. Tools Appl.* **2018**, *77*, 1–25. [CrossRef]
8. Zhao, G.; Li, X.; Li, Y.; Pietikäinen, M. Facial micro-expressions: An overview. *Proc. IEEE* **2023**, *111*, 1215–1235. [CrossRef]
9. Li, Y.; Wei, J.; Liu, Y.; Kauttonen, J.; Zhao, G. Deep learning for micro-expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2028–2046. [CrossRef]
10. Lucas, B.D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, BC, Canada, 24–28 August 1981 ; Volume 2, pp. 674–679.
11. Xu, F.; Zhang, J.; Wang, J.Z. Micro-expression identification and categorization using a facial dynamics map. *IEEE Trans. Affect. Comput.* **2017**, *8*, 254–267. [CrossRef]
12. Liu, Y.J.; Zhang, J.K.; Yan, W.J.; Wang, S.J.; Zhao, G.; Fu, X. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affect. Comput.* **2015**, *7*, 299–310. [CrossRef]
13. Khor, H.Q.; See, J.; Phan, R.C.; Lin, W. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 667–674.
14. Pérez, J.S.; Meinhardt, L.E.; Facciolo, G. TV-L1 optical flow estimation. *Image Process. Line* **2013**, *1*, 137–150. [CrossRef]
15. Liang, Z.; He, J.; Sun, Y. A three-dimensional convolution neural network evolution method for automatic recognition of micro expression. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval, Shenzhen, China, 6–8 August 2020; pp. 79–84.
16. Lo, L.; Xie, H.X.; Shuai, H.H.; Cheng, W.H. MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks. *Comput. Sci.* **2018**, *47*, 227–232.
17. Xie, H.X.; Lo, L.; Shuai, H.H.; Cheng, W.H. AU-assisted graph attention convolutional network for micro-expression recognition. In Proceedings of the MM '20: Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 24–29.
18. Lei, L.; Chen, T.; Li, S.; Li, J. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Nashville, TN, USA, 20–25 June 2021; pp. 76–79.
19. Gupta, P. MERASTC: Micro-expression recognition using effective feature encodings and 2D convolutional neural network. *IEEE Trans. Affect. Comput.* **2021**, *99*, 11–13. [CrossRef]
20. Wang, Y.; Zheng, S.; Sun, X.; Guo, D.; Lang, J. Micro-expression recognition with attention mechanism and region enhancement. *Multimedia Syst.* **2022**, *27*, 1–9. [CrossRef]
21. Cen, S.; Yu, Y.; Yan, G.; Yu, M.; Guo, Y. Multi-task facial activity patterns learning for micro-expression recognition using joint temporal local cube binary pattern. *Signal Process.-Image* **2022**, *103*, 1–9. [CrossRef]
22. Wu, H.-Y.; Rubinstein, M.; Shih, E.; Guttag, J.; Durand, F.; Freeman, W. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.* **2012**, *31*, 1–8. [CrossRef]
23. Oh, T.H.; Jaroensri, R.; Kim, C.; Elgharib, M.; Durand, F.E.; Freeman, W.T.; Matusik, W. Learning-based video motion magnification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 633–648.
24. 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**, *41*, 5826–5846. [CrossRef]
25. Zhou, L.; Shao, X.; Mao, Q. A survey of micro-expression recognition. *Image Vis. Comput.* **2021**, *105*, 104043. [CrossRef]
26. Hui, T.W.; Tang, X.; Loy, C.C. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8981–8989.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
29. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.
30. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
31. 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.
32. Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.; Liu, Y.J.; Chen, Y.H.; Fu, X. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **2014**, *9*, e86041. [CrossRef]
33. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models-their training and application. *Comput. Vis. Image Und.* **1995**, *61*, 38–59. [CrossRef]
34. Goshtasby, A. Image registration by local approximation methods. *Image Vis. Comput.* **1988**, *6*, 255–261. [CrossRef]
35. Pfister, T.; Li, X.; Zhao, G.; Pietikäinen, M. Recognising spontaneous facial micro-expressions. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1449–1456.

36. Liong, S.T.; Wong, K. Micro-expression recognition using apex frame with phase information. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 534–537.
37. Liu, Y.J.; Li, B.J.; Lai, Y.K. Sparse mdmo: Learning a discriminative feature for micro-expression recognition. *IEEE Trans. Affect. Comput.* **2018**, *12*, 254–261. [CrossRef]
38. Khor, H.Q.; See,J.; Liong, S.T.; Phan, R.C.; Lin, W. Dual-stream shallow networks for facial micro-expression recognition. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 36–40.
39. Nie, X.; Takalkar, M.A.; Duan, M.; Zhang, H.; Xu, M. GEME: Dual-stream multi-task GEnder-based micro-expression recognition. *Neurocomputing* **2021**, *427*, 13–28. [CrossRef]
40. Gan, Y.S.; Liong, S.T.; Yau, W.C.; Huang, Y.C.; Tan, L.K. OFF-ApexNet on micro-expression recognition system. *Signal Process. Image Commun.* **2019**, *74*, 129–139. [CrossRef]