Micro-expression Action Unit Recognition Based on Dynamic Image and Spatial Pyramid

Guanqun Zhou (954655899@qq.com)
Shandong University

Shusen Yuan (yuanshusen@mail.sdu.edu.cn)
Shandong University

Hongbo Xing (lingmeng@mail.sdu.edu.cn)
Shandong University

Youjun Jiang (874707833@qq.com)
Shandong University

Pinyong Geng (pygeng2020@163.com)
Shandong University

Yewen Cao (ycao@sdu.edu.cn)
Shandong University

Xianye Ben (benxianye@gmail.com)
Shandong University

Research Article

Keywords:

DOI: https://doi.org/

Additional Declarations: No competing interests reported.

Micro-expression Action Unit Recognition Based on Dynamic Image and Spatial Pyramid

Guanqun Zhou[1†], Shusen Yuan[1*†], Hongbo Xing[1], Youjun Jiang[1], Pinyong Geng[1], Yewen Cao[1*] and Xianye Ben[1]

[1*]Shandong University, Qingdao, 266200, Shandong, China.

*Corresponding author(s). E-mail(s): yuanshusen@mail.sdu.edu.cn; ycao@sdu.edu.cn;
†These authors contributed equally to this work.

Abstract

Most of the existing research focuses on the recognition of micro-expressions, and few studies how to recognize the action units of micro-expressions. This is due to the low intensity of the facial action unit, which is not easily to be recognized. To solve this problem, we proposed a micro-expression action unit recognition algorithm based on dynamic image and spatial pyramids. First, the video is passed through the dynamic image generation module to generate a dynamic image and extract the motion information contained in all frames. Then, given the subtle movement properties of micro-expressions, different levels of semantic features are obtained through spatial pyramids. It is also known that micro-expressions appear in the small range and are concentrated in local area of the face, so the regional feature network and attention mechanism are used for the image features of each layer. Finally, due to the weak correlation between each action unit, our models are trained separately. Experiments on CASME and CAS(ME)^2 datasets verify that the proposed algorithm has shown better action unit recognition performance compared with other advanced methods.

Keywords: Micro-expression, Action unit recognition, Dynamic image, Spatial pyramids

1 Introduction

The study found [1] that when liars lie, they will subconsciously try to suppress and disguise their true emotions. The basis of their suppression and disguise is the cognition and experience of various emotions. However, because the liar's cognition and experience of various emotions are biased, the "acted" expressions are unnatural, only superficial and fragmentary, and the real emotions will be revealed unconsciously through micro-expressions come out. Micro-expressions are short-lasting, difficult-to-detect, and uncontrollable facial movements that appear when a person tries to hide his or her true emotions, often reflecting the individual's true emotions. It is characterized by short duration, low intensity, and localized appearance on the face, usually about 1/25s to 1/5s of the face; in contrast, macro-expressions are controllable facial expressions, which usually appear on the face for about 1/5 second to 4 seconds, with obvious facial movement and covering a large facial area. The comparison between two expressions is shown in Fig. 1. Micro-expressions were first discovered by Haggard and Isaacs [2] in 1966 and called "micro-moment" expressions. In 1969, Ekman and Friesen [3] reported that they had discovered a special facial expression, which they named: micro-expressions. Analyzing micro-expressions is valuable for many potential applications, such as medical [4], law enforcement [5], political psychology [6], national security [7], etc.

(a) macro-expression

(b) micro-expression

Fig. 1: Comparison between macro-expression and micro-expression image [8].

Micro-expression recognition includes two important branches, namely, expression recognition and action unit recognition. Most of the existing research focuses on the expression recognition of micro-expressions [9–11], while few studies how to recognize the action units of micro-expressions. Expression recognition can only make general divisions of expressions, such as six basic human expressions such as happy, angry, disgusted, fearful, sad

Table 1: Commonly used AUs.

AU	Describe	Example
1	Inner Brow Raiser	
2	Outer Brow Raiser	
4	Brow Lowerer	
5	Upper Lid Raiser	
6	Cheek Raiser	
7	Lid Tightener	
9	Nose Wrinkler	
10	Upper Lip Raiser	
12	Lip Corner Puller	
14	Dimpler	
15	Lip Corner Depressor	
16	Lower Lip Depressor	
17	Chin Raiser	
45	Blink	

and surprised. Since human expressions are complex, in order to recognize the complete expression, it is necessary to use the face action unit (AU) to divide. AUs are the basic movements of a single muscle or muscle group, and different AU combinations can describe most expressions. Facial Action Coding System (FACS) [12] shows that successful facial action unit recognition can greatly facilitate the analysis of complex facial actions or expressions. Therefore, exploring AU is very important for in-depth interpretation of the facial behavior of micro-expressions. The commonly used AUs [13] are shown in Table 1.

Currently, there are many research methods on AU recognition of macro-expressions [16–19]. Compared with the AU recognition research on

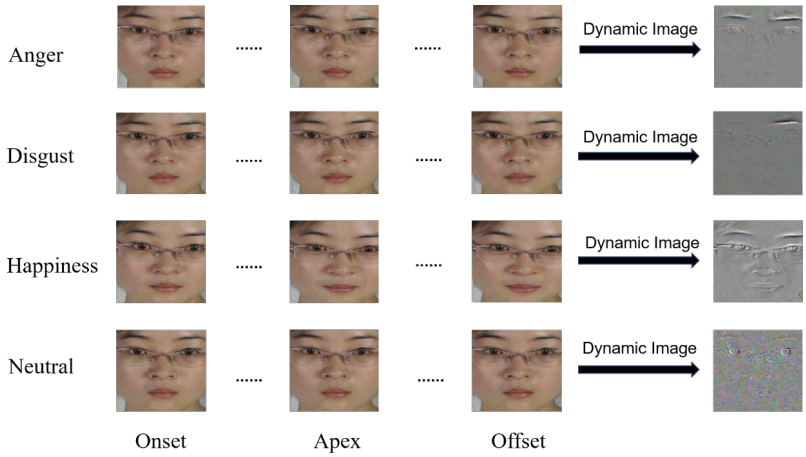

Fig. 2: Dynamic Image. It can be seen that in the angry expression video, the experimental subject frowned; in the disgusted expression video, the experimental subject's left eyebrow was raised; in the happy expression video, the experimental subject's mouth and eyebrows moved; In the neutral expression, the subject has no facial movements. The facial actions mentioned above can all be reflected in the dynamic image generated corresponding to the video [14, 15].

macro-expressions, there are relatively few AU recognition studies on micro-expressions [20–22] because of the following problems:

- The intensity of micro-expression AU recognition is much lower, and the duration of AU occurrence is much shorter, resulting in difficulty in recognition;
- Compared with the macro-expression AU dataset (such as the BP4D dataset [23]) (328 videos and a total of about 140,000 frames), the micro-expression AU dataset (such as CAS(ME)^2 dataset [8]) contains a very small number of samples;
- There are not several AUs coexisting in micro-expressions, that is, the correlation is weak, that is to say, the multi-label learning framework [24] commonly used in macro-expressions is not suitable for micro-expression AU recognition;
- The number of AU samples of micro-expressions is unbalanced. Some AU samples have many, such as AU4 (brow down), and some AU samples have only a few, such as AU10 (upper lip).

Unfortunately, micro-expression AU recognition also has some problems, that is, micro-expression has subtle and rapid facial muscle changes. Therefore, it is more difficult to recognize AUs on micro-expressions than that on macro-expressions.

To solve the above problems, we propose a micro-expression action unit recognition algorithm based on dynamic image and spatial pyramids.

The study found that the analysis of micro-expressions is mostly based on the analysis of video. Therefore, it is crucial to understand and represent video content accurately. A motion map, such as dynamic image [14, 15] is a single RGB image, equivalent to a still image, that captures the dynamics and appearance of an entire video sequence or subsequence, resulting in a long-term, stable representation of motion. Dynamic image can be applied to CNN network architectures commonly used in image tasks, such as VGG, ResNet and other networks, while the network can still infer long-term dynamics in videos and learn dynamic features. The dynamic image is shown in Fig. 2.

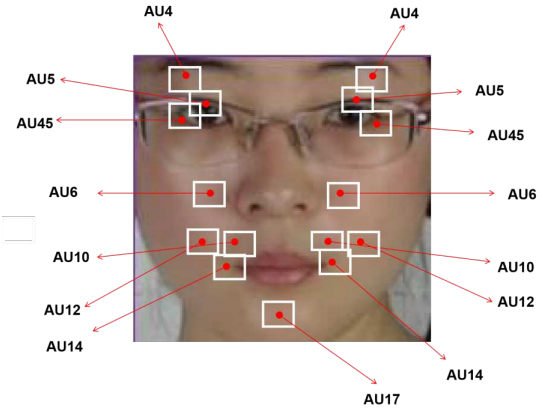

Fig. 3: AU distribution map in face.

Due to the subtle changes of micro-expressions, it is not easy to be captured and localized. We believe that this is similar to the feature localization problem considered in fine-grained image recognition. In the general research method, only the high-level features of the last network layer are used for final recognition, but because of the limitation of the perceptual field, only one range of local area feature information can be collected, and it is impossible to locate micro-expressions synthetically from the local area feature information of multiple range sizes. Furthermore, studies have shown that for convolutional neural networks, high-resolution low-level features help to capture detailed information of local regions, while low-resolution high-level features contain global semantic information that is crucial for classification. Therefore, we propose to use a spatial pyramid network to fuse multi-scale features from different layers to localize micro-expressions.

Unlike facial expression recognition, which only needs to analyze the entire face, the action unit (AU) of the face appears in a sparse facial area and needs to be analyzed in local areas. The different AU distributions are shown in Fig. 3. Most expression recognition methods use standard convolutional layers to

learn image features and assume that the weights of the convolutional kernels are shared across the image. But the human face is a structured image, such an assumption will not be able to capture the local subtle appearance changes, so different local feature extraction methods should be used for different facial regions. To this end, we propose a regional feature module to extract local features.

To highlight important features, we propose an attention module to emphasize decisive features and suppress invalid features, and then utilize residuals to improve robustness to partial face occlusion or camera viewpoint changes.

Finally, for each micro-expression AU model is a binary classification task. However, due to the unbalanced distribution of the number of samples in AUs, this makes it easier for the model to identify AUs with a large number of samples, and more difficult to identify AUs with a small number of samples. Therefore, we use the focal loss function to solve the problem of unbalanced distribution of AU samples.

Our main contributions are as follows:

- In order to solve the problems of low intensity and difficult recognition of the action units of micro-expressions, we propose a micro-expression action unit recognition algorithm based on dynamic image and spatial pyramids;
- We conduct in-depth experiments on two public micro-expression datasets with AU labels, and demonstrate the effectiveness of the algorithm for micro-expression action unit recognition;
- Since the micro-expression occurs in the local area of the face, we proposed a regional feature module to extract regional features, and reinforce the effectiveness of the micro-expression finding task in our action unit recognition algorithm;
- We proposed an attention module to emphasize the important features, weaken the impact of useless features, and enhance the robustness of the micro-expression action unit recognition algorithm.

2 Related Work

Currently, there are many research methods on AU recognition of macro-expressions [16, 17], mainly based on manual extraction of facial AU appearance features and geometric features. Appearance features represent local or global changes in the face. Commonly used appearance features are Haar feature [25], Histogram of Oriented Gradients (HOG) feature [26], LBP feature [27], Garbor wavelet feature [28], and Scale Invariant Feature Transform (SIFT) features [29]. Valstar et al. [30] extracted Gabor features at local regions of face landmarks and used support vector machines for classification tasks. [31] proposed joint patch and multi-label learning for AU recognition using SIFT descriptors near landmark points. Geometric features represent the changing direction or distance of face landmarks or skin. Lien et al. [32] developed a computer vision system that is very sensitive to subtle changes

in the face. The system includes three modules for extracting feature information: dense optical flow extraction based on wavelet motion model, face feature tracking, and edge and line extraction. The feature information thus extracted is input into a discriminative classifier or hidden Markov model, which is classified into different AUs. What??s more, Geometric changes can be measured by optical flow or displacement of facial landmarks [33, 34]. Fabian et al. [35] proposed a method combining geometric variation and local texture information. However, these hand-crafted features still do not represent facial variations well.

In recent years, deep learning methods have been widely studied in AU recognition of macro-expressions due to their strong nonlinear representation capabilities [36, 37]. Li et al. [38] proposed a local convolutional neural network (LCNN) to learn AUs on cropped regions centered on face landmarks, but this network suffers from severe instability in face landmark detection. Chu et al. [36] proposed a method based on Deep Region and Multi-label Learning (DRML), which utilizes region layers to obtain important face regions. It extracts facial structure information for good AU recognition results with subtle motion changes. In addition, Li et al. [39] proposed a local feature learning method to embed an attention map based on facial landmarks in cropped regions. These works strongly suggest that learned features can identify AUs well.

Compared with the AU recognition research on macro-expressions, there are relatively few AU recognition studies on micro-expressions. [20] proposed a deep Spatio-Temporal Adaptive Pooling (STAP) network with focal loss for micro-expression AU recognition. STAP is an end-to-end trainable network capable of identifying subtle and rapidly changing micro-expression AUs on specific regions with effective temporal information. Afterwards, Li et al. [21] proposed an end-to-end Spatial Channel Attention (SCA) network for micro-expression AU recognition, which consists of spatial and channel modules for spatial relations, respectively Modeling and local area representation. The SCA network efficiently identifies subtle AUs by using self-second-order statistics.

3 Micro-Expression Action Unit Recognition Algorithm Based on Dynamic Image And Spatial Pyramid

Aiming at the low strength of the action units of micro-expressions and the difficulty in being recognized, we propose a micro-expression action unit recognition algorithm based on dynamic image and spatial pyramids. The network is shown in Fig. 4. The network mainly includes 3 important modules: dynamic image generation module, feature extraction module and attention module. The dynamic image generation module generates dynamic image. The feature extraction module includes two parts, the spatial feature module and the regional feature module. The spatial feature module uses the structure

of the spatial pyramid to extract the subtle feature changes of the face; the regional feature module captures the local appearance changes of different facial regions. The attention module highlights important features at different layers of the spatial pyramid. In this section, we first discuss how dynamic image work, then introduce spatial pyramids, regional feature networks, and attention mechanisms, and finally, we introduce the Focal loss function, which addresses the imbalanced distribution of AU samples.

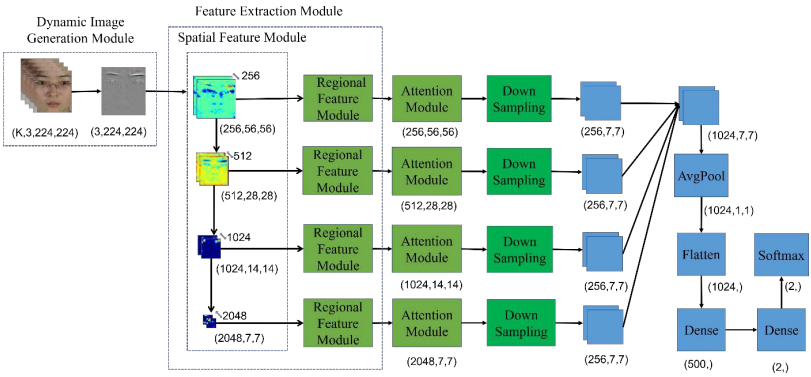

Fig. 4: Network framework of our proposed method, which consists of four sub-modules: Dynamic Image Generation Model, Spatial Feature Module, Regional Feature Module and Attention Module. Different scales of output of different Attention Module are all down sampled to $256 \times 7 \times 7$ and then concatenated to $1024 \times 7 \times 7$ for the classification.

3.1 Dynamic Image

The dynamic map is obtained by directly applying sorting pooling to the original image pixels of the video, and is the parameter of the sorting function finally obtained by solving the sorting support vector machine.

Suppose there is a micro-expression interval, which is represented as $\left[x_1, x_2 \ldots, x_{(K-1)}, x_K\right]$, where $x_t \in \mathbb{R}^{3\times 224\times 224}$ is the t-th frame, $t = 1, \ldots, K$. K is the number of micro-expression frames. Let $\psi(x_t)$ be the feature representation of the t-th frame x_t of the micro-expression, which can be directly represented by the original RGB image, that is, $\psi(x_t) = x_t$. After smoothing the original frame sequence, a new sequence $V = \left[v_1, v_2 \ldots, v_{(K-1)}, v_K\right]$ is obtained, and the smoothing process is shown in Equation (1).

$$v_t = \frac{1}{t}\sum_{i=1}^{t}\psi(x_i), t = 1, \ldots, K \tag{1}$$

Since action changes are time-related, and the frame order of micro-expression sequences is known. Then, the two frames of v_t and v_{t+1} in sequence are defined as $v_{t+1} \succ v_t$, that is, the v_{t+1} frame is after the v_t frame.

Let the sorting function be $f(x)$, according to the size of $f(x)$ to decide which frame is in the front and which frame is in the back. The properties of $f(x)$ are shown in Equation (2).

$$\begin{aligned} x_i \succ x_j &\Leftrightarrow f(x_i) > f(x_j) \\ x_i \prec x_j &\Leftrightarrow f(x_i) < f(x_j) \end{aligned} \tag{2}$$

In theory, $f(x)$ can be any function. For simplicity, it is assumed that it is a linear function $f(x) = \langle w, x \rangle$. From the properties of the linear function, for any two features x_i and x_j, the satisfying relationship is shown in Equation (3).

$$\begin{aligned} f(x_i) > f(x_j) &\Leftrightarrow \langle w, x_i - x_j \rangle > 0 \\ f(x_i) < f(x_i) &\Leftrightarrow \langle w, x_i - x_j \rangle < 0 \end{aligned} \tag{3}$$

For features v_i and v_j, the above relationship is also satisfied as shown in Equation (4).

$$\begin{aligned} f(v_i) > f(v_j) &\Leftrightarrow \langle w, v_i - v_j \rangle > 0 \\ f(v_i) < f(v_i) &\Leftrightarrow \langle w, v_i - v_j \rangle < 0 \end{aligned} \tag{4}$$

Define positive samples as $v_i - v_j$and negative samples as $v_j - v_i$, where time i is after j, and the corresponding ground-truth labels are set as shown in Equation (5).

$$y = \begin{cases} 1, \text{if } v_i - v_j \\ -1, \text{if } v_j - v_i \end{cases} \tag{5}$$

Equation (2) to Equation (5) establish a relationship, as shown in Equation (6).

$$\begin{aligned} \langle w, v_i - v_j \rangle > 0 \Leftrightarrow f(v_i) > f(v_j) \Leftrightarrow v_i \succ v_j \Leftrightarrow y = 1 \\ \langle w, v_i - v_j \rangle < 0 \Leftrightarrow f(v_i) < f(v_j) \Leftrightarrow v_i \prec v_j \Leftrightarrow y = -1 \end{aligned} \tag{6}$$

The above sorting problem is converted into a classification problem, which can be solved in the general way of Support Vector Machine (SVM). Learn the following convex optimization problem, as shown in Equation (7).

$$w^* = \arg\min_w E(w) \tag{7}$$

where,

$$E(w) = \frac{1}{2} \|w\|^2 + C \sum_{i>j} \max\{0, 1 - \langle w, v_i - v_j \rangle\} \tag{8}$$

From the equation, the parameter w can be learned, and then the sorting function $f(v) = \langle w, v \rangle$ can be obtained, and there is $\forall i, j, v_i \succ v_j \Leftrightarrow f(v_i) > f(v_j)$. Since the w^* vector contains enough information to rank all the frames in the video, it aggregates the information of all

the frames and can be used as a video descriptor. The above process of solving the video frame sequence w^* is called sorting pooling.

The above solution process is still complicated. In order to obtain dynamic image more easily, approximate sorting pooling is used. Let $w = 0$, then according to the properties of gradient descent, we can get $w^* = 0 - \eta \nabla E(w)|_{w=0} \propto -\nabla E(w)|_{w=0}$, and $\nabla E(w)|_{w=0}$ is shown in Equation (9).

$$\begin{aligned} \nabla E(w)|_{w=0} &\propto \sum_{i>j} \nabla \max\{0, 1 - \langle w, v_i - v_j \rangle\}|_{w=0} \\ &= \sum_{i>j} \nabla \langle w, v_j - v_i \rangle = \sum_{i>j} v_i - v_i \end{aligned} \tag{9}$$

The following Equation (10) is further obtained.

$$\begin{aligned} w^* \propto \sum_{i>j} v_i - v_j &= \sum_{i>j} \left[\frac{1}{i} \sum_{m=1}^{i} \psi(x_m) - \frac{1}{j} \sum_{m=1}^{j} \psi(x_m) \right] \\ &= \sum_{t=1}^{K} \alpha_t \psi(x_t) \end{aligned} \tag{10}$$

Among them, $\alpha_t = 2(K - t + 1) - (K + 1)(H_K - H_{t-1})$, $H_t = \sum_{i=1}^{t} 1/i$, so $w^* = \sum_{t=1}^{K} \alpha_t \psi(x_t)$.The calculated parameter value w^* is the required dynamic image.

3.2 Space Pyramid

Studies have shown that for convolutional neural networks, high-resolution low-level features help to capture detailed information in local regions, while low-resolution high-level: features contain global semantic information that is critical for classification. We use a ResNet50 network with 4 intermediate convolutional parts, and take the output features of the last residual block of each convolutional part as the features of one layer of the spatial pyramid. Due to the difference in the size of the receptive field of different layers, the scope of the local area context that can be observed contains important features is also different. Comprehensive consideration is helpful to locate the area of AU feature change. The spatial pyramid is shown in Fig. 5.

3.3 Regional Feature Network

Since human faces are structured image, in order to capture local subtle appearance changes, different local feature extraction methods should be used for different facial regions. To this end, we propose a Regional Feature Module (Reg) based on the literature [36] to extract local features, as shown in Fig. 6.

The regional feature module first divides the input image into 7×7 grids, each grid represents a local region, and then extracts features for each local region. Different from using only one convolutional layer in [36], in order to

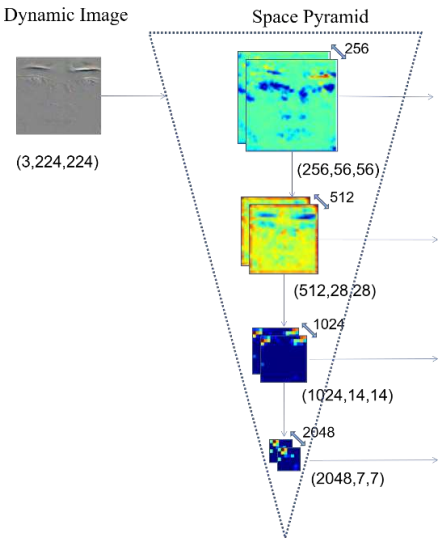

Fig. 5: Spatial pyramid model. The output of the last residual block in each of the four convolutions of ResNet50 is used as a feature of the spatial pyramid.

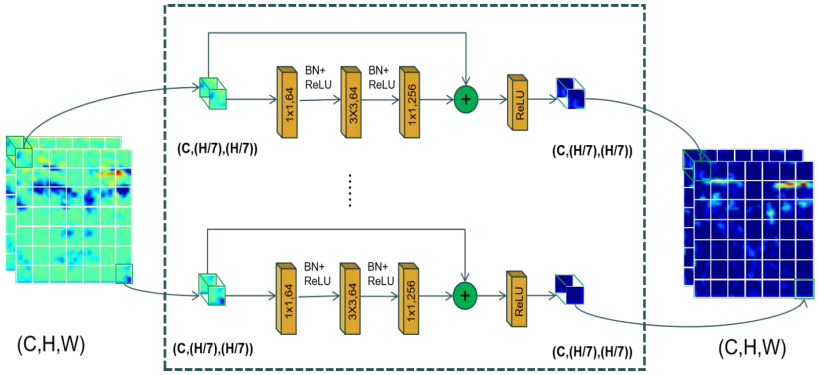

Fig. 6: Regional feature module. Different facial regions use different local feature extraction methods to capture subtle movements.

fully extract subtle features, this chapter uses two 1×1 convolutions and one 3×3 convolution for each local region. Two 1×1 convolutions, the former is used for dimensionality reduction and the latter is used for dimensionality enhancement, ensuring that the output and input size are the same. Batch normalization (BN) and ReLU activation functions are used after each convolution. Local convolutions are used to capture local appearance changes, and the learned weights for each local region are updated independently. Furthermore, if no useful information about AU is learned in the local region, the

original local region features are directly output using the residual. The output size of the local area after passing through the regional feature module is the same as the input size, and the position of the image is also the same. The generated feature map should be placed at the original local area position and combined with other local area output feature maps to form a new one. image. In this way, AUs are identified in sparse facial local regions.

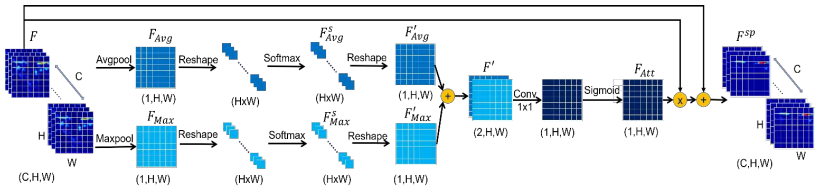

Fig. 7: Attention module.Generate attention maps that emphasize deterministic features and suppress invalid features, using the values of feature points to reflect the importance of location.

3.4 Attention Mechanism

To highlight important features, we propose the Attention Module (Att), as shown in Fig. 7.

The input feature map of the module is $F \in \mathbb{R}^{C\times H\times W}$. To calculate spatial attention, first apply max pooling and average pooling operations along the channel axis to obtain the feature maximum and average value of each channel at each position (i, j), which are used to represent the salient features of this position, and finally respectively Generate feature maps $F_{Avg} \in \mathbb{R}^{1\times H\times W}$ and $F_{Max} \in \mathbb{R}^{1\times H\times W}$. The max pooling and average pooling operations are shown in Equation (11) and Equation (12).

$$F_{Avg}(i,j) = \frac{\sum_{n=1}^{c} F^n(i,j)}{C} \tag{11}$$

$$F_{Max}(i,j) = \max\left(F^1(i,j), \ldots, F^C(i,j)\right) \tag{12}$$

where n is the channel index, referring to the n-th channel; C is the total number of channels. $F^n(i,j)$ represents the feature value at the position (i,j) of the n-th channel feature map of F. $F_{Avg}(i,j)$ is the feature average of all channel feature maps of F at position (i,j), $F_{Max}(i,j)$ is the feature maximum of all channel feature maps of F at position (i,j) value.

Then flatten it into a feature vector of length $H \times W$, and apply the softmax function to obtain the importance of the feature of each position point in the entire face space, and obtain the feature maps F^s_{Avg} and F^s_{Max}. The calculation method are shown in Equation (13) and Equation (14).

$$F_{Avg}^{s}(i,j) = \frac{e^{F_{Avg}(i,j)}}{\Sigma_{i,j} e^{F_{Avg}(i,j)}} \tag{13}$$

$$F_{Max}^{s}(i,j) = \frac{e^{F_{Max}(i,j)}}{\Sigma_{i,j} e^{F_{Max}(i,j)}} \tag{14}$$

Resize the generated feature maps F_{Avg}^{s} and F_{Max}^{s} to $F_{Avg}^{'}$ and $F_{Max}^{'}$. Then combined into feature map $F^{'} \in \mathbb{R}^{2\times H\times W}$. After 1×1 convolution kernel convolutional dimension reduction, the sigmoid function is used to limit all the values to the range of $0 \sim 1$ to obtain the final spatial attention feature map $F_{Att} \in \mathbb{R}^{1\times H\times W}$.

Finally, the original feature map F and F_{Att} are multiplied, and the product result is added to the original feature map F to obtain the final output feature map F^{sp},forming a residual block to avoid the problem of vanishing gradients during training. The residual operation is shown in Equation (15).

$$F^{sp} = F \cdot F_{Att} + F \tag{15}$$

where $F^{sp} \in \mathbb{R}^{C\times H\times W}$ is the final output feature map, F is the original input feature map, and F_{Att} is the spatial attention feature map.

3.5 Focal Loss Function

To address the problem of unbalanced distribution of AU sample numbers, we use a focal loss function as shown in Equation (16).

$$\begin{aligned} Loss = &-\frac{1}{M}\sum_{i=1}^{M} \alpha y_i \left(1-\widehat{y}_i\right)^{\gamma} \log\left(\widehat{y}_i\right) \\ &+ \left(1-\alpha\right)\left(1-y_i\right)\left(\widehat{y}_i\right)^{\gamma} \log\left(1-\widehat{y}_i\right) \end{aligned} \tag{16}$$

Among them, M is the total number of samples, y_i is the true label of the i-th sample, if AU appears, it is 1, otherwise it is 0. $\widehat{y}_i$ is the predicted label of the i-th sample, representing the probability of AU occurrence. γ is usually taken as 2, and α is usually taken as 0.25. Among them, γ represents the weight of the hard samples, which is used to reduce the loss contribution of the easy samples, so that the network training process pays more attention to the hard samples. α is the class weight, which is used to weigh the imbalance of positive and negative samples.

4 Experiments

4.1 Settings

Datasets: We evaluate the proposed algorithm on two spontaneous micro-expression datasets, CASME [40] and CAS(ME)^2 [8]. Both micro-expression datasets contain AU labels, but the number of samples per AU label is quite different. In our experiments, we only explored 8 AUs related to lie recognition,

namely AU4, AU5, AU6, AU10, AU12, AU14, AU17 and AU45. The number of AU samples contained in different datasets is shown in Table 2.

Table 2: Number of AU samples for CASME and CAS(ME)^2 datasets.

Dataset	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45
CASME	69	0	1	3	9	23	13	0
CAS(ME)^2	120	11	63	6	120	46	9	14

CASME contains 195 micro-expression samples from 35 subjects at a frame rate of 60fps and an image resolution of 640×480. Since the average duration of video samples is only two or three seconds. Therefore, this dataset is only suitable for micro-expression recognition, not for micro-expression discovery. The dataset contains: AU4, AU6, AU10, AU12, AU14, AU17, etc. 6 AUs related to lies.

CAS(ME)^2 contains 357 expression samples from 22 subjects, of which the number of micro-expression samples is 57, the number of macro-expression samples is 300, the frame rate is 30fps, and the image resolution is 640×480. for micro-expression recognition and discovery tasks. The dataset contains: AU4, AU5, AU6, AU10, AU12, AU14, AU17, and AU45, 8 AUs related to lies.

Metrics: AU recognition is a binary classification problem. For binary classification tasks, especially in the case of unbalanced samples, the F1-score can better explain the performance of the algorithm. F1-score is the result of comprehensively considering the precision rate and recall rate of the model, and the size is $0 \sim 1$. F1-score is that bigger is better. In our evaluation, the F1-scores of 6 AUs in CASME and 8 AUs in CAS(ME)^2 were calculated according to the number and importance of AUs. The overall performance of the algorithm is evaluated by the average F1-score of all AUs.

Implementation: In our experiments, we used the cropped face images provided by the dataset. The input is a sequence of aligned RGB micro-expression images. Since the average number of micro-expression frames is 10 frames, it is necessary to expand each micro-expression image sequence to 10 frames through a temporal interpolation model. Finally, the preprocessed image sequence is input into our proposed algorithm for AU recognition. The training adopts the form of "one-vs-rest", that is, all samples that currently recognize AUs are marked as positive samples, and the samples of other AUs are marked as negative samples, and a binary classification model is trained for each AU. The ratio of training set and test set data size is 8:2. The optimizer uses the Adaptive Moment Estimation (Adam) method, the learning rate is set to 0.001, the training epoch is 100, and the batch size is 60.

4.2 Results

In this section, we conduct relevant comparative experiments on the CASME and CAS(ME)^2 datasets, respectively, and discuss them from four perspectives: ResNet network depth, ResNet network layer combination, image type, and method type.

ResNet network depth: It can be seen from Table 3 and Table 4 that in the CASME and CAS(ME)^2 datasets, the F1-score of ResNet50 is higher than that of ResNet101 and ResNet152, which are improved by 0.022 and 0.11, 0.041 and 0.135, respectively. This is because with the continuous increase of the network depth, the number of AU samples is too small, which leads to overfitting, and the performance of AU recognition begins to decline. Therefore, we choose ResNet50 as the spatial feature module.

Table 3: F1-score comparison of different ResNet network depth in the CASME dataset.

Depth	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
ResNet50	**0.701**	—	**0.181**	**0.242**	**0.451**	0.53	**0.625**	—	**0.455**
ResNet101	0.681	—	0.154	0.143	0.431	**0.643**	0.546	—	0.433
ResNet152	0.578	—	0.085	0.078	0.345	0.463	0.523	—	0.345

Table 4: F1-score comparison of different ResNet network depth in the CAS(ME)^2 dataset.

Depth	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
ResNet50	**0.901**	**0.667**	**0.91**	0.711	**0.954**	**0.863**	**0.635**	0.643	**0.786**
ResNet101	0.856	0.546	0.876	**0.713**	0.946	0.854	0.587	0.578	0.745
ResNet152	0.865	0.443	0.784	0.658	0.874	0.759	0.343	0.478	0.651

ResNet network layer combination: ResNet50 has a total of 4 intermediate convolution parts, namely conv2_x, conv3_x, conv4_x and conv5_x. When an image of size $3 \times 224 \times 224$ is input, the four parts output feature maps of size $256 \times 56 \times 56$, $512 \times 28 \times 28$, $1024 \times 14 \times 14$, and $2048 \times 7 \times 7$, respectively, corresponding to Features from low-level to high-level. The spatial pyramid is composed of these four feature maps. To facilitate analysis, the four intermediate convolutional parts are labeled as c1, c2, c3, and c4, corresponding to conv2_x, conv3_x, conv4_x, and conv5_x, respectively. As can be seen from Table 5 and Table 6, the performance of the network improves as more levels of features are used to combine. This shows that combining low-level features with high-level features can produce better results for AU recognition. This is because for low-level features, it has higher spatial resolution for AU localization, but the identified semantic information is lower;

for high-level features, it has higher semantic features, but due to pooling, convolutional and other operations, the spatial resolution is low, which is not conducive to the extraction of local area information. Therefore, using this spatial pyramid form, the spatial resolution and semantic information of the feature layer can be guaranteed.

Table 5: F1-score comparison of different combinations of ResNet network layers in the CASME dataset.

Index	Combination	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
1	c1	0.609	—	0.162	0.129	0.389	0.49	0.591	—	0.395
2	c2	0.602	—	0.165	0.129	0.388	0.486	0.595	—	0.394
3	c3	0.611	—	0.164	0.13	0.39	0.496	0.582	—	0.396
4	c4	0.615	—	0.163	0.131	0.378	0.488	0.588	—	0.394
5	c2+c1	0.62	—	0.176	0.232	0.41	0.496	0.601	—	0.423
6	c3+c1	0.623	—	0.177	0.234	0.423	0.502	0.601	—	0.427
7	c3+c2	0.598	—	0.175	0.232	0.415	0.503	0.605	—	0.421
8	c4+c1	0.624	—	0.175	0.234	0.42	0.512	0.61	—	0.429
9	c4+c2	0.634	—	0.176	0.23	0.425	0.509	0.612	—	0.431
10	c4+c3	0.627	—	0.179	0.235	0.43	0.515	0.609	—	0.433
11	c3+c2+c1	0.688	—	0.18	**0.242**	0.443	0.516	0.615	—	0.447
12	c4+c2+c1	0.686	—	0.181	0.241	0.441	0.523	0.614	—	0.448
13	c4+c3+c2	0.684	—	**0.182**	0.24	0.448	0.52	0.619	—	0.449
14	c4+c3+c2+c1	**0.701**	—	0.181	**0.242**	**0.451**	**0.53**	**0.625**	—	**0.455**

Table 6: F1-score comparison of different combinations of ResNet network layers in the CAS(ME)^2 dataset.

Index	Combination	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
1	c1	0.873	0.641	0.89	0.685	0.933	0.852	0.599	0.623	0.762
2	c2	0.874	0.643	0.89	0.687	0.932	0.854	0.601	0.62	0.763
3	c3	0.879	0.64	0.889	0.677	0.936	0.855	0.603	0.625	0.763
4	c4	0.876	0.65	0.896	0.684	0.94	0.849	0.61	0.626	0.766
5	c2+c1	0.883	0.65	0.901	0.69	0.941	0.856	0.611	0.63	0.77
6	c3+c1	0.881	0.651	0.902	0.694	0.94	0.857	0.615	0.631	0.771
7	c3+c2	0.884	0.653	0.899	0.692	0.943	0.845	0.614	0.633	0.77
8	c4+c1	0.889	0.651	0.903	0.697	0.945	0.856	0.62	0.64	0.775
9	c4+c2	0.89	0.654	0.904	0.7	0.944	0.855	0.625	0.641	0.777
10	c4+c3	0.892	0.659	0.902	0.699	0.946	0.858	0.624	0.639	0.777
11	c3+c2+c1	0.893	0.667	0.909	0.702	0.953	0.86	**0.64**	**0.643**	0.783
12	c4+c2+c1	0.896	**0.67**	**0.91**	0.703	0.95	0.86	0.63	**0.643**	0.783
13	c4+c3+c2	0.899	0.668	0.909	0.705	0.952	**0.863**	0.633	0.642	0.784
14	c4+c3+c2+c1	**0.901**	0.667	**0.91**	**0.711**	**0.954**	**0.863**	0.635	**0.643**	**0.786**

Image type: To verify the reliability of dynamic image, we compare the dynamic image with average pooling image, max pooling image and micro-expression vertex frames on CASME and CAS(ME)^2 datasets. From Table 7

and Table 8, it can be seen that the experimental effect of dynamic image is the best, and the average F1-score values of 0.455 and 0.786 are obtained on the CASME and CAS(ME)^2 datasets, respectively, which shows that the dynamic image is better than other types of images. Contains more video action information. The experiment of max pooling image is the worst, getting an average F1-score of 0.389 and 0.722 on CASME and CAS(ME)^2 datasets, respectively, because max pooling loses a lot of important information. Average pooling image and vertex frames also produce good experimental results, because average pooling combines all feature information, and apex frames contain the feature information with the largest range of motion.

Table 7: F1-score comparison of different types of input images in the CASME datasets.

Image Type	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
Average pooling image	**0.702**	—	0.179	0.231	0.44	**0.531**	0.61	—	0.449
Max pooling image	0.612	—	0.075	0.198	0.354	0.499	0.6	—	0.389
Apex image	0.699	—	0.18	**0.242**	**0.452**	0.525	0.62	—	0.453
Dynamic image	0.701	—	**0.181**	**0.242**	0.451	0.53	**0.625**	—	**0.455**

Table 8: F1-score comparison of different types of input images in the CAS(ME)^2 datasets.

Image Type	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
Average pooling image	0.899	0.659	**0.915**	0.704	0.946	0.854	0.63	0.64	0.781
Max pooling image	0.753	0.579	0.876	0.7	0.875	0.798	0.597	0.597	0.722
Apex image	0.9	**0.668**	0.9	**0.723**	0.946	0.86	0.628	0.641	0.783
Dynamic image	**0.901**	0.667	0.91	0.711	**0.954**	**0.863**	**0.635**	**0.643**	**0.786**

Method type: From Table 9 to Table 10, it can be seen that with LBP-TOP [41] as the baseline, the F1-score of our algorithm in the CASME dataset and CAS(ME)^2 dataset is improved by 0.258 and 0.443, respectively, compared with the baseline. The F1-score value of handcrafted features is generally lower than that of learned features, because handcrafted features may ignore more detailed information of images, and learned features can capture more discriminative features for micro-expression AU recognition. Among the three handcrafted features LBP-TOP, LPQ-TOP [42] and LBP-SIP [43], LBP-TOP performs the best, obtaining F1-scores of 0.197 and 0.343 in CASME dataset and CAS(ME)^2 dataset, respectively. The learned feature I3D [44] is a network that augments 2DCNN to 3DCNN, which can achieve significant quality gain with less computation. SCA is an end-to-end spatial channel attention network for micro-expression AU recognition, which consists of spatial and channel modules for spatial relationship modeling and local region

representation, respectively. The SCA network can effectively identify subtle AUs through self-second-order statistics. The STAP network captures discriminative AU information based on region of interest and weighted temporal information fusion. SCA and STAP both use image sequences as input and perform poorly compared to dynamic image. Specifically, our proposed micro-expression action unit recognition algorithm outperforms SCA, STAP by 0.101 and 0.183, 0.025 and 0.050 on CASME and CAS(ME)^2 in terms of average F1-score, respectively. The results indicate that the dynamic image well preserves the motion information to a single image with a simple structure, and achieves better experimental results. Comparing with other advanced methods, in the CASME dataset, our algorithm has the highest recognition rate for AU4, AU10, AU12, and AU14, and the average F1-scores are higher than other methods; in the CAS(ME)^2 dataset, after recognition by our algorithm, six out of eight AUs have much higher F1-scores than other methods and are in the first place. The validity of our proposed network is again verified.

Table 9: F1-score comparison of different methods in the CASME datasets.

Methods	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
LBP-TOP [41]	0.472	—	0.14	0.152	0.206	0.214	0	—	0.197
LPQ-TOP [42]	0.568	—	0.012	0.015	0.189	0	0	—	0.131
LBP-SIP [43]	0.45	—	0.022	0.009	0.172	0.254	0	—	0.151
I3D [44]	0.588	—	0.138	0.235	0.389	0.513	0.61	—	0.412
SCA [21]	0.545	—	**0.184**	0.223	0.34	0.331	0.502	—	0.354
STAP [20]	0.624	—	0.143	0.23	0.435	0.511	**0.635**	—	0.43
Ours	**0.701**	—	0.181	**0.242**	**0.451**	**0.53**	0.625	—	**0.455**

Table 10: F1-score comparison of different methods in the CAS(ME)^2 datasets.

Methods	AU4	AU5	AU6	AU10	AU12	AU14	AU17	AU45	Avg
LBP-TOP [41]	0.767	0.378	0.181	0.01	0.452	0.391	0.223	0.389	0.343
LPQ-TOP [42]	0.701	0.317	0.024	0	0.33	0.343	0.228	0.301	0.281
LBP-SIP [43]	0.758	0.245	0.158	0	0.269	0.345	0.401	0.12	0.287
I3D [44]	0.842	0.665	0.864	**0.739**	0.845	0.651	0.605	0.603	0.727
SCA [21]	0.799	0.451	0.821	0.735	0.712	0.41	0.431	0.468	0.603
STAP [20]	0.866	0.653	0.864	0.709	0.805	0.742	**0.645**	0.6	0.736
Ours	**0.901**	**0.667**	**0.91**	0.711	**0.954**	**0.863**	0.635	**0.643**	**0.786**

In short, our proposed method introduces the dynamic image and spatial pyramids to recognize AUs with micro-expression information. Compared with LBP-TOP, LPQ-TOP, LBP-SIP and I3D, the results show the effectiveness of the micro-expression of improving the performance of the action unit recognition task. And the results also demonstrate the superiority of the dynamic

image for extracting micro-expression information compared with SCA and STAP.

5 Conclusion

Currently, micro-expression AU recognition is still an important and challenging task. In this paper, in view of the low intensity of the action units of micro-expressions and difficulty in being recognized, we propose a micro-expression action unit recognition algorithm based on dynamic image and spatial pyramids, which can recognize subtle and rapidly changing micro-expressions in local areas of the face. First, since there is little coexistence between micro-expression AUs, we train a model for each AU separately, and each model adopts the same network structure. Then, the motion and appearance characteristics of the entire video sequence or sub-sequence are captured using the dynamic image video representation. After that, the spatial pyramid network is used to extract subtle features at different layers, and the regional feature network is used to capture the local appearance changes of the face, and the attention mechanism is used to highlight important features. Finally, we conduct extensive experiments on the CASME dataset and the CAS(ME)^2 dataset to demonstrate the effectiveness of the algorithm. In the future, in view of the small sample size of the current micro-expression dataset, we need to design a new dataset to expand the current amount of data.

References

[1] Yang, P., Jin, H., Li, Z.: Combining attention mechanism and dual-stream 3d convolutional neural network for micro-expression recognition. In: 2022 7th International Conference on Image, Vision and Computing (ICIVC), pp. 51–59 (2022). https://doi.org/10.1109/ICIVC55077.2022.9886046

[2] Haggard, E.A., Isaacs, K.S.: Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy, pp. 154–165. Springer, Boston, MA (1966). https://doi.org/10.1007/978-1-4684-6045-2_14

[3] Ekman P, F.W.: Nonverbal leakage and clues to deception. Psychiatry **32**(1), 88–106 (1969). https://doi.org/10.1080/00332747.1969.11023575

[4] Yu, E.H., Choi, E.J., Lee, S.Y., Im, S.J., Yune, S.J., Baek, S.Y.: Effects of micro- and subtle-expression reading skill training in medical students: A randomized trial. Patient Education and Counseling **99**(10), 1670–1675 (2016). https://doi.org/10.1016/j.pec.2016.04.013

[5] Frank, M.G., Svetieva, E.: In: Mandal, M.K., Awasthi, A. (eds.) Microexpressions and Deception, pp. 227–242. Springer, New Delhi (2015). https://doi.org/10.1007/978-81-322-1934-7_11

[6] Döllinger, L., Laukka, P., Högman, L.B., Bänziger, T., Makower, I., Fischer, H., Hau, S.: Training emotion recognition accuracy: Results for multimodal expressions and facial micro expressions. Frontiers in Psychology **12** (2021). https://doi.org/10.3389/fpsyg.2021.708867

[7] Khan, W., Crockett, K., O'Shea, J., Hussain, A., Khan, B.M.: Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection. Expert Systems with Applications **169**, 114341 (2021). https://doi.org/10.1016/j.eswa.2020.114341

[8] Qu, F., Wang, S.-J., Yan, W.-J., Li, H., Wu, S., Fu, X.: Cas(me)^2 : A database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Transactions on Affective Computing **9**(4), 424–436 (2018). https://doi.org/10.1109/TAFFC.2017.2654440

[9] Duan, X., Dai, Q., Wang, X., Wang, Y., Hua, Z.: Recognizing spontaneous micro-expression from eye region. Neurocomputing **217**, 27–36 (2016). https://doi.org/10.1016/j.neucom.2016.03.090. SI: ALLSHC

[10] Wang, S.-J., Yan, W.-J., Sun, T., Zhao, G., Fu, X.: Sparse tensor canonical correlation analysis for micro-expression recognition. Neurocomputing **214**, 218–232 (2016). https://doi.org/10.1016/j.neucom.2016.05.083

[11] Sun, B., Cao, S., Li, D., He, J., Yu, L.: Dynamic micro-expression recognition using knowledge distillation. IEEE Transactions on Affective Computing **13**(2), 1037–1043 (2022). https://doi.org/10.1109/TAFFC.2020.2986962

[12] Wiggers, M., Vangelder, R., Heymans, P.: The evaluation of facial paralysis: a case study using the facial action coding system and electromyography. Journal of Clinical and Experimental Neuropsychology **9**, 278–279 (1987)

[13] Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE Transactions on Affective Computing **10**(3), 325–347 (2019). https://doi.org/10.1109/TAFFC.2017.2731763

[14] Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S.: Dynamic image networks for action recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3034–3042 (2016). https://doi.org/10.1109/CVPR.2016.331

[15] Bilen, H., Fernando, B., Gavves, E., Vedaldi, A.: Action recognition with dynamic image networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2799–2813 (2018). https://doi.org/10.1109/TPAMI.2017.2769085

[16] Zhao, K., Chu, W.-S., Martinez, A.M.: Learning facial action units from web images with scalable weakly supervised clustering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2090–2099 (2018). https://doi.org/10.1109/CVPR.2018.00223

[17] Han, S., Meng, Z., O'Reilly, J., Cai, J., Wang, X., Tong, Y.: Optimizing filter size in convolutional neural networks for facial action unit recognition. CoRR **abs/1707.08630** (2017). http://arxiv.org/abs/1707.08630

[18] Wang, S., Pan, B., Wu, S., Ji, Q.: Deep facial action unit recognition and intensity estimation from partially labelled data. IEEE Transactions on Affective Computing **12**(4), 1018–1030 (2021). https://doi.org/10.1109/TAFFC.2019.2914654

[19] Hoai, D.L., Lim, E., Choi, E., Kim, S., Pant, S., Lee, G.-S., Kim, S.-H., Yang, H.-J.: An attention-based method for multi-label facial action unit detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2453–2458 (2022). https://doi.org/10.1109/CVPRW56347.2022.00274

[20] Li, Y., Huang, X., Zhao, G.: Micro-expression action unit detection withspatio-temporal adaptive pooling. CoRR **abs/1907.05023** (2019). http://arxiv.org/abs/1907.05023

[21] Li, Y., Huang, X., Zhao, G.: Micro-expression action unit detection with spatial and channel attention. Neurocomputing **436**, 221–231 (2021). https://doi.org/10.1016/j.neucom.2021.01.032

[22] Li, Y., Peng, W., Zhao, G.: Micro-expression action unit detection with dual-view attentive similarity-preserving knowledge distillation. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 01–08 (2021). https://doi.org/10.1109/FG52635.2021.9666975

[23] Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014). https://doi.org/10.1016/j.imavis.2014.06.002

[24] Zhang, W., Wang, L., Yan, J., Wang, X., Zha, H.: Deep extreme multi-label learning. CoRR **abs/1704.03718** (2017) https://arxiv.org/abs/1704.03718. http://arxiv.org/abs/1704.03718

[25] Whitehill, J., Omlin, C.W.: Haar features for facs au recognition. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06), pp. 5–101 (2006). https://doi.org/10.1109/FGR.2006.61

[26] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–8931 (2005). https://doi.org/10.1109/CVPR.2005.177

[27] Jiang, B., Valstar, M.F., Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 314–321 (2011). https://doi.org/10.1109/FG.2011.5771416

[28] Bazzo, J.J., Lamar, M.V.: Recognizing facial actions using gabor wavelets with neutral face average difference. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings., pp. 505–510 (2004). https://doi.org/10.1109/AFGR.2004.1301583

[29] Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–11572 (1999). https://doi.org/10.1109/ICCV.1999.790410

[30] Valstar, M., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), pp. 149–149 (2006). https://doi.org/10.1109/CVPRW.2006.85

[31] Zhao, K., Chu, W.-S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit and holistic expression recognition. IEEE Transactions on Image Processing **25**(8), 3931–3946 (2016). https://doi.org/10.1109/TIP.2016.2570550

[32] Lien, J.J.-J., Kanade, T., Cohn, J.F., Li, C.-C.: Detection, tracking, and classification of action units in facial expression. Robotics and Autonomous Systems **31**(3), 131–146 (2000). https://doi.org/10.1016/S0921-8890(99)00103-7

[33] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101 (2010). https://doi.org/10.1109/CVPRW.2010.5543262

[34] Valstar, M.F., Pantic, M.: Fully automatic recognition of the temporal phases of facial actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **42**(1), 28–43 (2012). https://doi.org/10.1109/TSMCB.2011.2163710

[35] Benitez-Quiroz, C.F., Srinivasan, R., Martinez, A.M.: Emotionet: An

accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5562–5570 (2016). https://doi.org/10.1109/CVPR.2016.600

[36] Zhao, K., Chu, W.-S., Zhang, H.: Deep region and multi-label learning for facial action unit detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3391–3399 (2016). https://doi.org/10.1109/CVPR.2016.369

[37] Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing **13**(3), 1195–1215 (2022). https://doi.org/10.1109/TAFFC.2020.2981446

[38] Li, W., Abtahi, F., Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6766–6775 (2017). https://doi.org/10.1109/CVPR.2017.716

[39] Li, W., Abtahi, F., Zhu, Z., Yin, L.: Eac-net: Deep nets with enhancing and cropping for facial action unit detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(11), 2583–2596 (2018). https://doi.org/10.1109/TPAMI.2018.2791608

[40] Yan, W.-J., Wu, Q., Liu, Y.-J., Wang, S.-J., Fu, X.: Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–7 (2013). https://doi.org/10.1109/FG.2013.6553799

[41] Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(6), 915–928 (2007). https://doi.org/10.1109/TPAMI.2007.1110

[42] Päivärinta, J., Rahtu, E., Heikkilä, J.: Volume local phase quantization for blur-insensitive dynamic texture classification. In: Heyden, A., Kahl, F. (eds.) Image Analysis, pp. 360–369. Springer, Berlin, Heidelberg (2011)

[43] Wang, Y., See, J., Phan, R.C.-W., Oh, Y.-H.: Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) Computer Vision – ACCV 2014, pp. 525–537. Springer, Cham (2015)

[44] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision

and Pattern Recognition (CVPR), pp. 4724–4733 (2017). https://doi.org/10.1109/CVPR.2017.502

Figures

Figure 1

Figure 2

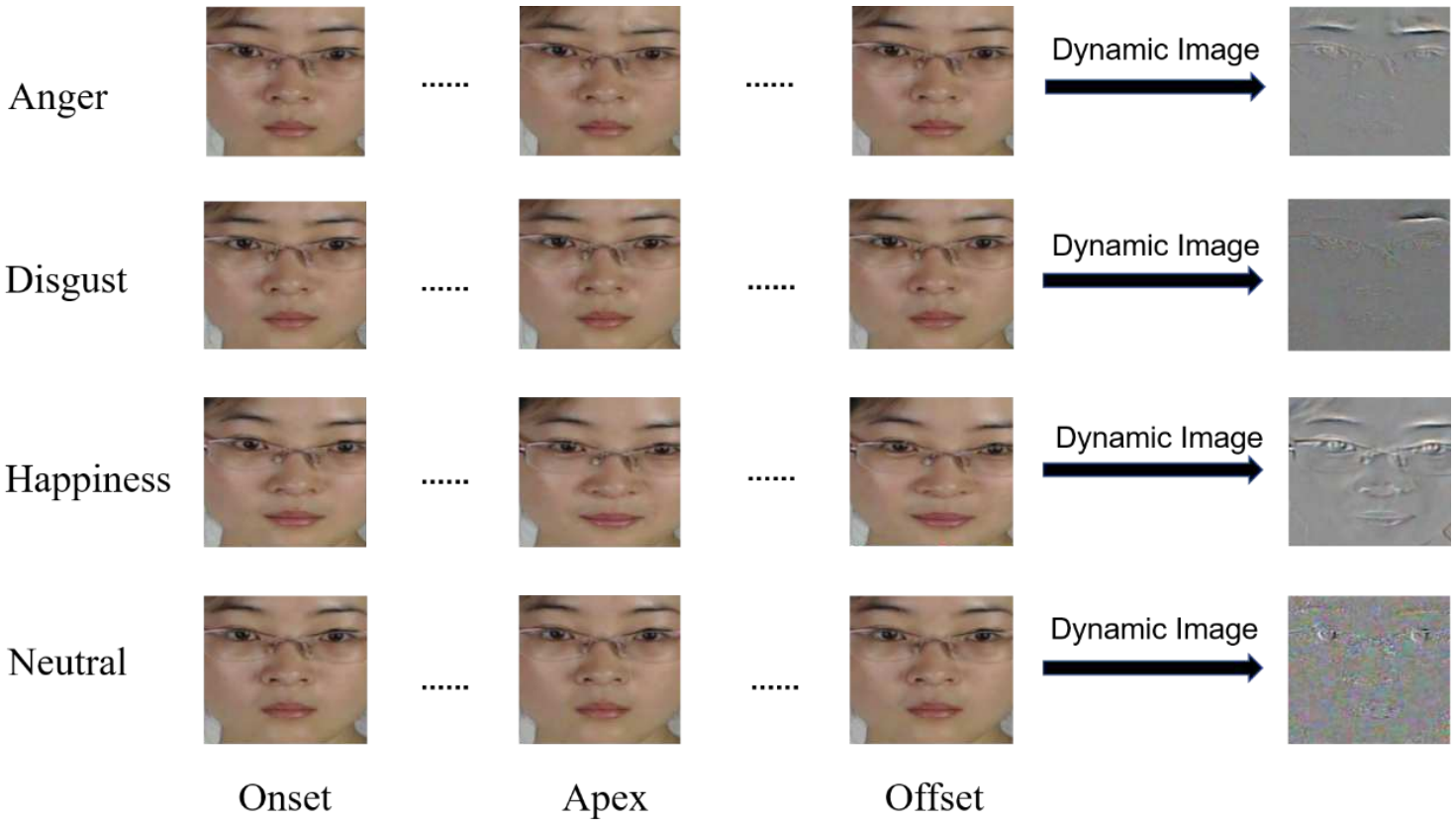
Anger
Disgust
Happiness
Neutral
Dynamic Image
Dynamic Image
Dynamic Image
Dynamic Image
Onset
Apex
Offset

Figure 3

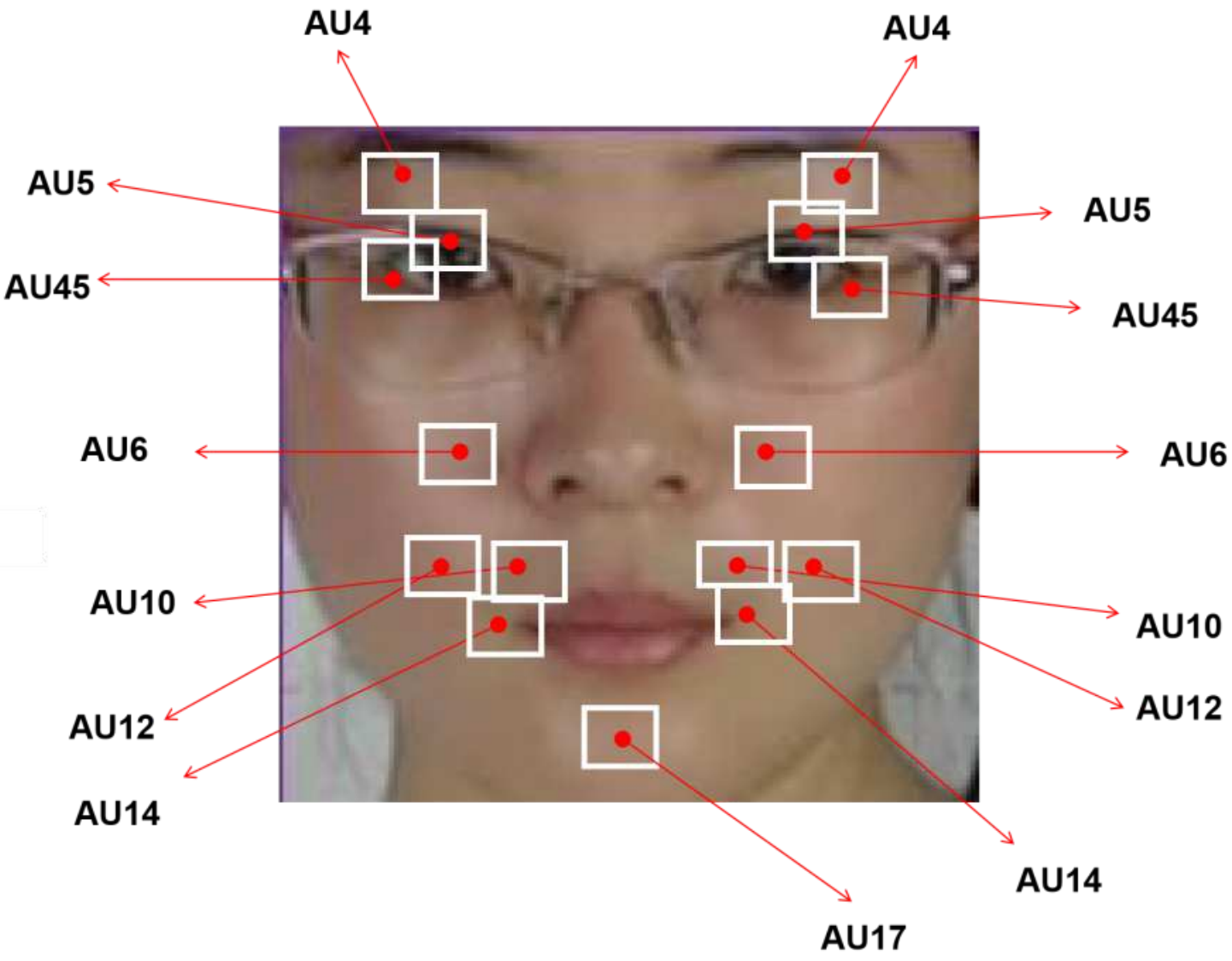
AU4
AU4
AU5
AU5
AU45
AU45
AU6
AU6
AU10
AU10
AU12
AU12
AU14
AU14
AU17

Figure 4

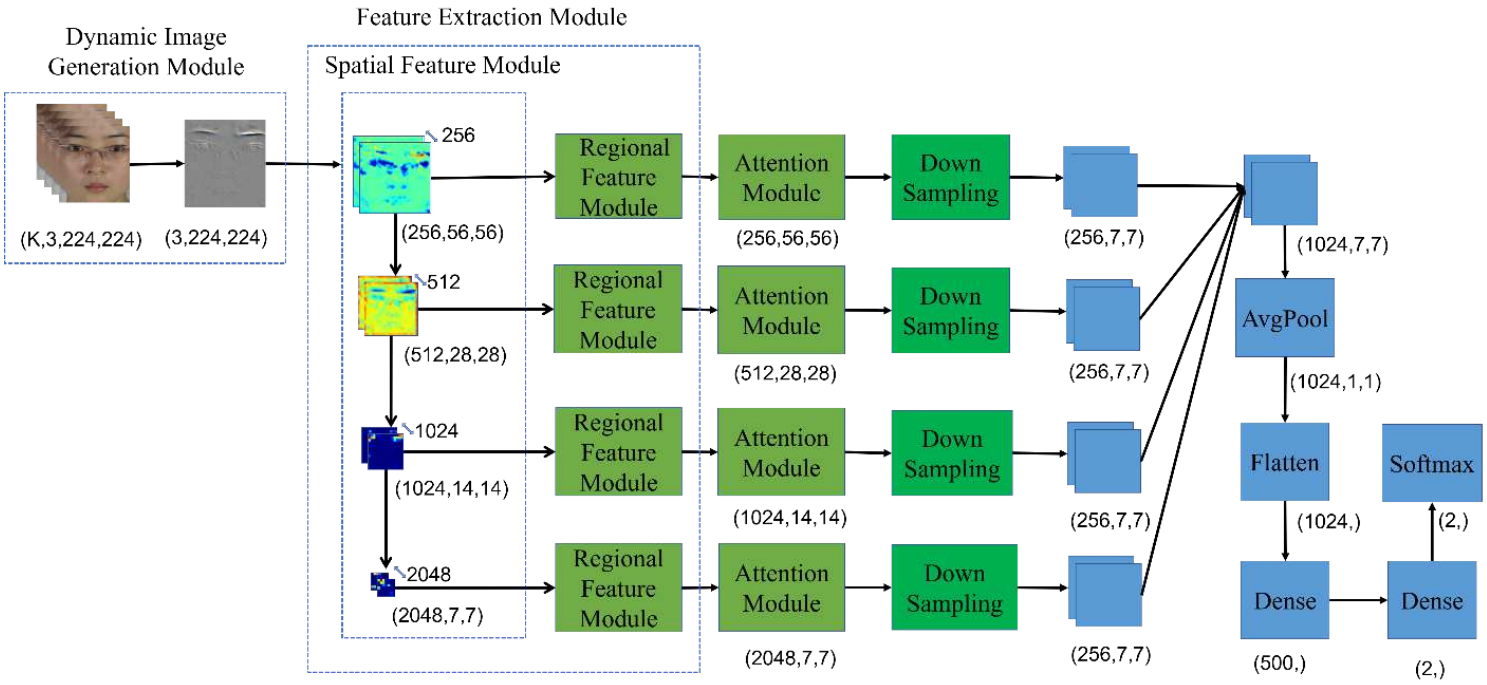
Dynamic Image Generation Module
Feature Extraction Module
Spatial Feature Module
(K,3,224,224)
(3,224,224)
256
(256,56,56)
512
(512,28,28)
1024
(1024,14,14)
2048
(2048,7,7)
Regional Feature Module
Regional Feature Module
Regional Feature Module
Regional Feature Module
Attention Module
Attention Module
Attention Module
Attention Module
(256,56,56)
(512,28,28)
(1024,14,14)
(2048,7,7)
Down Sampling
Down Sampling
Down Sampling
Down Sampling
(256,7,7)
(256,7,7)
(256,7,7)
(256,7,7)
(1024,7,7)
AvgPool
(1024,1,1)
Flatten
(1024,)
Dense
(500,)
Dense
(2,)
Softmax
(2,)

Figure 5

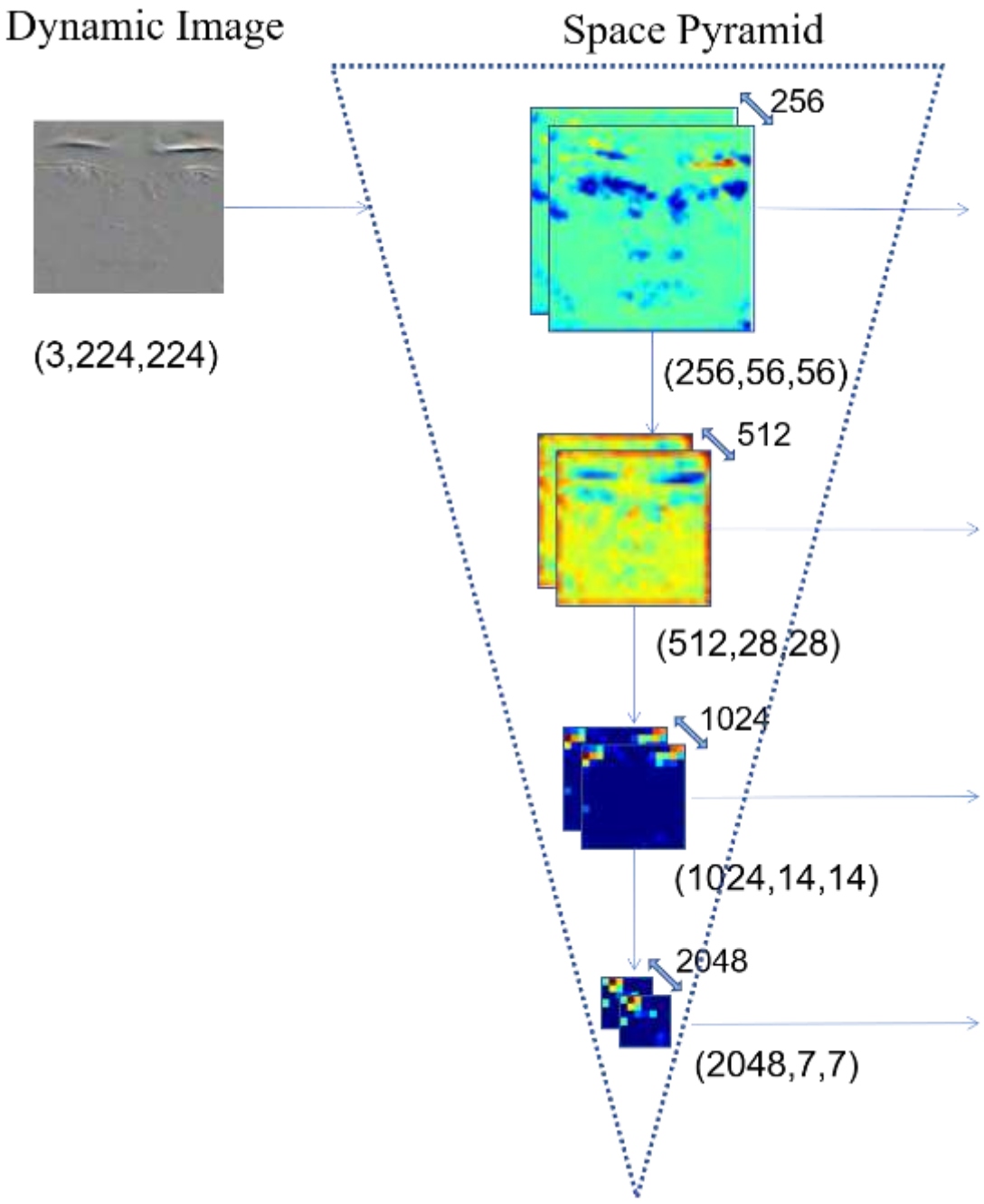
Dynamic Image
Space Pyramid
256
(3,224,224)
(256,56,56)
512
(512,28,28)
1024
(1024,14,14)
2048
(2048,7,7)

Figure 6

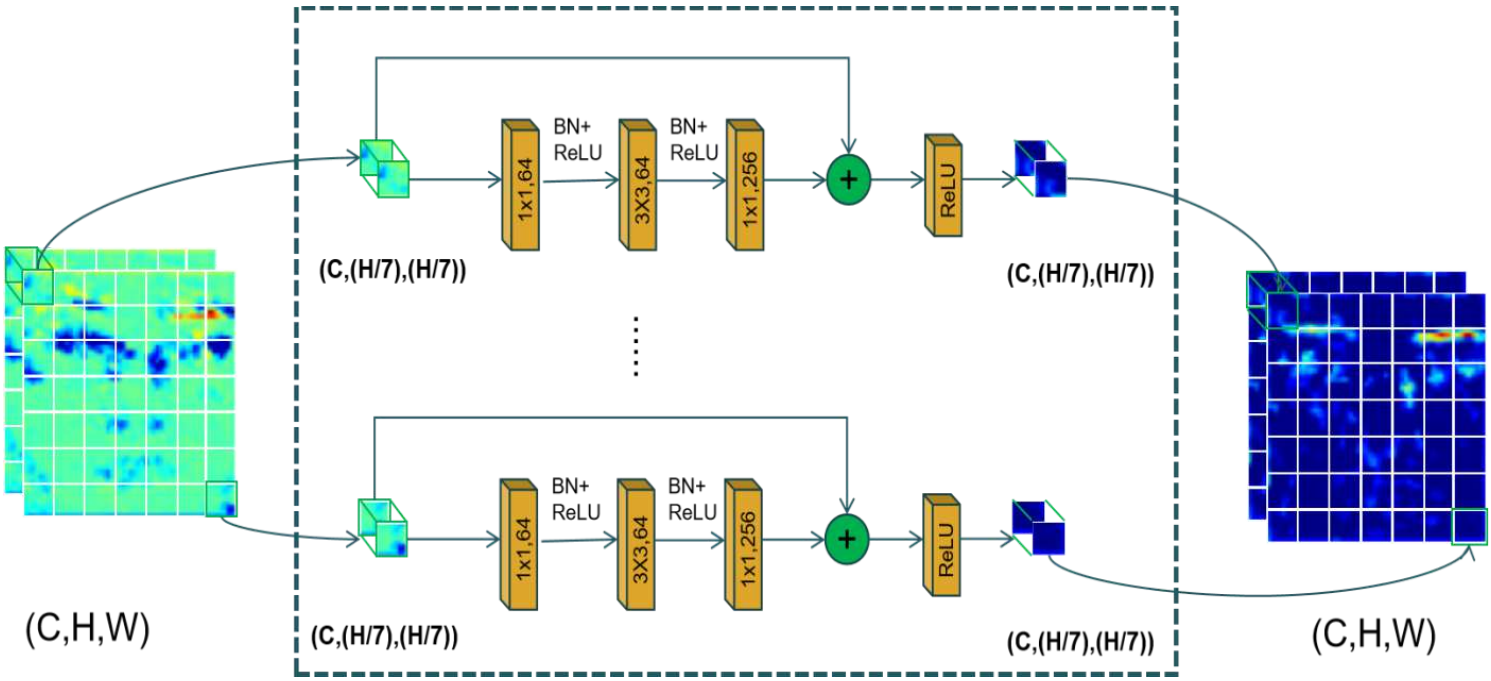

Figure 7

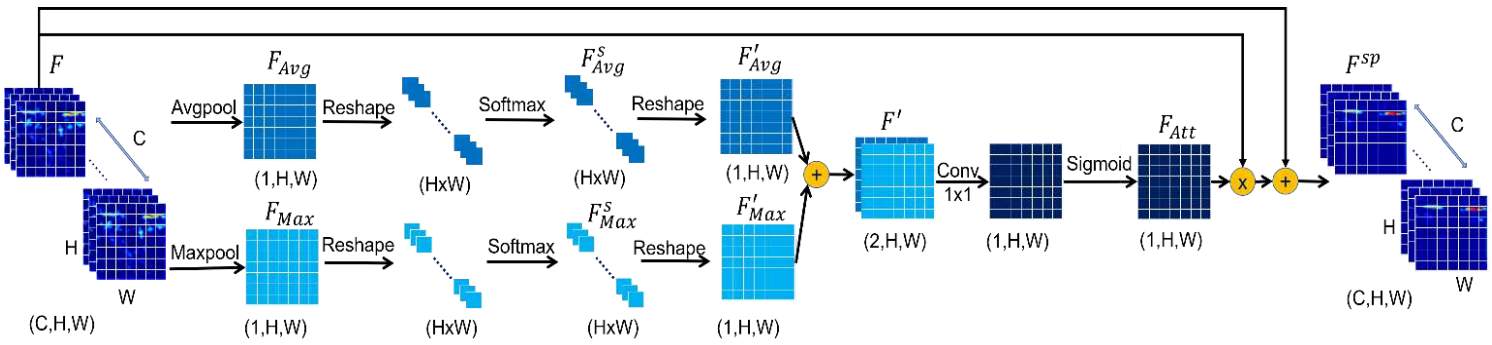

Figure 8

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- snjnl.cls
- snbibliography.bib