Graphical Abstract

Parallel Spatiotemporal Network to Recognize Micro-Expression

Jingting Li, Su-Jing Wang, Yong Wang, Haoliang Zhou, Xiaolan Fu

Highlights

Parallel Spatiotemporal Network to Recognize Micro-Expression

Jingting Li, Su-Jing Wang, Yong Wang, Haoliang Zhou, Xiaolan Fu

- A novel Temporal perception Unit (TPU) is proposed, with the temporal perception coefficient modulating the hidden state.
- Robust Principal Component Analysis (RPCA) is used to remove the identity information and only remain the subtle motion information of micro-expression.
- The element-wise addition with 1×1 convolutional kernel fusion model is proposed to better fuse the spatial and temporal features.

Parallel Spatiotemporal Network to Recognize Micro-Expression

Jingting Li[a,b], Su-Jing Wang[a,b,*], Yong Wang[c], Haoliang Zhou[d], Xiaolan Fu[e]

[a]*Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China*
[b]*Department of Psychology, University of the Chinese Academy of Sciences, Beijing, 100049, China*
[c]*College of Software, Xi'an Jiaotong University, Xi'anChina, 710000, China*
[d]*Tianjin University of Technology, Tianjin, 300382, China*
[e]*School of Psychology, Shanghai Jiao Tong University, Shanghai, 200240, China*

Abstract

Micro-expressions are fleeting spontaneous facial expressions that commonly occur in high-stakes scenarios and reflect humans' mental states. Thus, it is one of the crucial clues for lie detection. Furthermore, due to the brief duration of micro-expression, temporal information is important for micro-expression recognition. The paper proposes a Parallel Spatiotemporal Network (PSN) to recognize micro-expression. The proposed PSN includes a spatial sub-network and a temporal sub-network. The spatial sub-network is a shallow network with subtle motion information as the input. And the temporal sub-network is a network with a novel temporal feature extraction unit that extracts sparse temporal features of micro-expressions. Finally, we propose an element-wise addition with 1×1 convolutional kernel fusion model to fuse the spatial and temporal features. The proposed PSN gets better measurement metrics (such as recognition rate, F1 score, true positive rate, and true negative rate) than the other state-of-the-art methods on the consisted databases consisting of CASME, CASME II, CAS(ME)^2, and SAMM.

Keywords: Micro-expression Recognition, Deep Learning, Sparse Features, Affective Analysis, Lie Detection.

*Corresponding author: Su-Jing Wang (wangsujing@psych.ac.cn)

1. Introduction

Humans can convey emotions through various methods, such as body gestures, communication, and facial expression. As a kind of non-verbal behavior, facial expression can deliver much information. Mehrabian *et al.* found that the emotive information from the facial expressions accounts for 55% of emotive information from daily life [35]. Therefore, it is practical to know a person's mental activity state and genuine inner emotion from facial expressions. Usually, macro-expression, what we see in the daily interactions with people, lasts from 0.5 to 4 seconds [34]. However, one would disguise his or her macro-expression when trying to conceal his or her genuine emotion.

Nevertheless, genuine emotion cannot be fully controlled, and it always causes subtle and unconscious facial muscle movements that cannot be concealed or controlled. Haggard *et al.* [16] firstly found these facial muscle movements in 1966. Because of the short duration of these movements, they called them Rapid Expression. Ekman *et al.* [12] found this kind of facial muscle movement from an inpatient with psychotic, and they named it micro-expression. Micro-expressions are facial movements that last less than 0.5 seconds and occur when people hide their genuine emotions. It is widely used in security, judicial forensics, and clinical medicine as the critical clue of lie detection [10]. Micro-expression has three main characteristics: short duration, low intensity of movement, and fragmental action units [12].

A polygraph is a popular device for detecting lies. It can measure a person's blood pressure, pulse, respiration, and skin conductivity by answering a few questions during an interview [39]. During the lie detection, the polygraph connects multiple wires to the person, which will make him aware that he is being detected. Therefore, he may adopt some tactics to interfere with and disable the polygraph. Compared with lie detection based on polygraphs, lie detection based on micro-expressions is unobtrusive. Therefore, people will not realize they are monitored and will not develop countermeasures.

The research on micro-expression is the study of interdisciplinary psychology and computer science. The research from psychology focuses on the relationship between micro-expression and deception, human's ability to recognize micro-expressions, and so on. Ekman *et al.* [9] and Ten Brinke *et al.* [46] explored the relationship between micro-expression and deception, and found that micro-expression can be an effective clue to detect liars; Shen *et al.* [42] investigated how the micro-expression duration effects on recog-

nition accuracy of human beings, and found that recognition performance increases with increasing duration until reaching 200ms, then remains almost unchanged for 200ms; Zhang *et al.* [61] found that emotional context influences micro-expression recognition, and they further studied how emotional context modulates the processing of micro-expressions, understanding the processing mechanism [60][62].

To enhance the human's performance in recognizing micro-expression, Ekman [8] developed the Micro-Expression Training Tool (METT). However, not everyone can become a specialist in micro-expression recognition through training. Therefore, the development of an automatic recognition system for micro-expression is essential.

In computer science, researchers try to design some automatic methods to recognize micro-expression. These methods are divided into two main categories, one based on single-frame images[64, 55] and one based on temporal sequences [15, 31]. Although the single-frame-based micro-expression recognition methods can obtain a comparative accuracy, it is still not high. In particular, compared to macro-expressions, the static spatial information of single frames in micro-expression videos is not distinctly different from a neutral face due to the very low intensity of micro-expression. At the same time, the video contains more transition letter information of facial actions. This inter-frame dynamic information is vital for micro-expression recognition [31].

Therefore, many micro-expression studies have been conducted based on sequences. At the early stage of research, researchers classified micro-expressions by combining manual feature extraction with machine learning. Among these researches, many spatio-temporal features have been applied and improved, such as LBP-TOP [5], the combination of RPCA and Local Spatiotemporal Directional Features [51], LBP-TOP with integral projection [18], spatiotemporal completed local quantized patterns [21], local radon-based binary pattern [20], hierarchical Spatiotemporal Descriptor [65].

In recent years, micro-expression recognition methods combined with deep learning have also been gradually explored. On the one hand, to improve the network's efficiency in learning micro-expression features, researchers input the manually extracted spatio-temporal features of micro-expressions into the deep network to achieve the micro-expression classification task. For example, Liong et al. used multiple optical flow-derived components and an OFF-ApexNet structure to represent the facial subtle motion changes better [28]. Sun et al. proposed a multi-scale active-patches fusion-based

spatiotemporal LBP-TOP descriptor that considers the active contributions for different regional areas in faces [45]. Li et al. proposed a joint feature learning architecture that couples local and global information for recognizing micro-expressions [26]. Furthermore, they used self high-order statistics of spatio-wise and channel-wise features to detect Action Units in micro-expressions [27]. Mao et al. proposed a Region-inspired Relation Reasoning Network to model the relationships between various facial regions, addressing the issue of micro-expression recognition under partial occlusion [33].

On the other hand, the framework design considering both temporal and spatial dimensions becomes the key to the end-to-end network-based micro-expression recognition algorithm to extract micro-expression features effectively. For instance, Kim et al. proposed a two-scale spatio-temporal feature learning and utilized LSTM to recognize micro-expression [23]. Xia et al. proposed Spatiotemporal Recurrent Convolutional Networks, capturing the spatiotemporal deformations of micro-expression sequences [54]. Verma et al. proposed an AutoMER model, a spatiotemporal Neural Architecture Search for Micro-expression Recognition [48].

The micro-expression has two characteristics: short duration and low intensity of movement. They make data of micro-expression sparse in spatial and temporal domains. Thus, facial identity information in video clips becomes a massive noise for micro-expression recognition [51, 19, 18]. For recognizing micro-expression, the critical issue is how to extract the sparse spatial and temporal features effectively. In this paper, we propose a Parallel Spatiotemporal Network (PSN). It includes a spatial sub-network and a temporal sub-network to extract the sparse spatial and temporal features. In the temporal sub-network, we propose a novel Temporal perception Unit (TPU) based on temporal perception information to extract features in the temporal domain. Finally, we propose an element-wise addition fusion model to fuse the sparse spatial and temporal features.

The paper has three main contributions as follow:

- In the temporal domain, a novel Temporal perception Unit (TPU) is proposed. Compared with RNN, TPU introduces the temporal perception coefficient to modulate the hidden state. The temporal perception coefficient measures the amount of variation or dispersion of the feature difference between the current state and the previous state.

- In the spatial domain, Robust Principal Component Analysis (RPCA) is used to remove the identity information and only remain the subtle

motion information of micro-expression [51, 19, 18]. The subtle motion information better describes the spatial features of micro-expressions.

- In the fusion step, the element-wise addition with 1×1 convolutional kernel fusion model is proposed. The 1×1 convolutional kernel can better fuse the spatial and temporal features.

The rest of this paper is organized as follows: Section 2 presents our proposed Parallel Spatiotemporal Network (PSN) in detail; In Section 3, we expound the process of the experiments on the databases CASME [57], CASME II[56], CAS(ME)^2 [38] and SAMM [4]; In Section 4, conclusions are drawn and several issues for future work are discussed.

2. Parallel Spatiotemporal Network

This section proposes a novel Parallel Spatiotemporal Network (PSN) for micro-expression recognition. Fig. 1 shows the architecture of the network. The PSN includes a spatial sub-network and a temporal sub-network. For the temporal sub-network, we design a new recurrent unit based on Temporal Perception Information, called Temporal Perception Unit (TPU), to extract temporal features of micro-expressions. For the spatial sub-network, we use a shallow network to extract spatial features of micro-expressions. To efficiently extract spatial features, the input of the spatial sub-network is the information of micro-expressions without human identity information. To achieve this, we use RPCA to remove identity information from video clips of micro-expression [51]. Finally, element-wise addition is used to fuse spatial features and temporal features. To comprehensively fuse these two types of features, a convolutional layer with 1×1 size kernel is performed after the element-wise addition.

2.1. Spatial Sub-network

Whether for expression recognition or micro-expression recognition, spatial features are significant. They can be divided into appearance and geometrical. Appearance features use the image intensity information, while geometrical features measure distances, deformations, curvatures, and other geometric properties [2]. A micro-expression consists of one or more AUs. For some AUs, appearance features have better discrimination. For the others, geometrical features have better discrimination. For example, *Nose Wrinkler* (AU9) pulls the skin along the sides of the nose upwards towards the root of

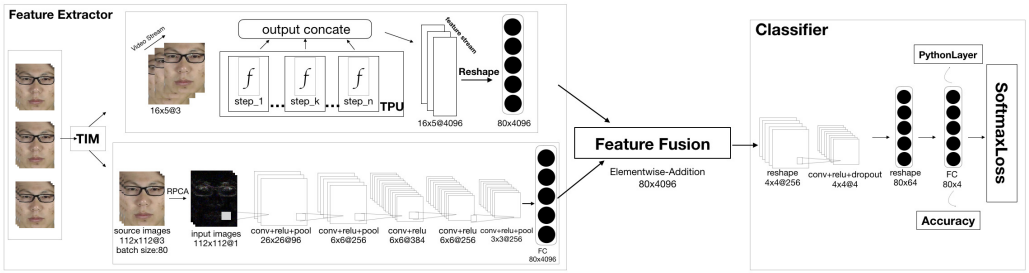
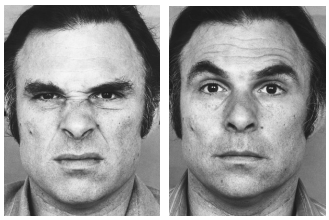


Figure 1: Architecture of the model we proposed in this paper.

the nose, causing wrinkles (see Fig. 2 (a)) to appear along the sides of the nose and across the root of the nose [11]. Appearance features of wrinkles include better discriminant information. However, *Outer Brow Raiser* (AU2) makes the eyebrows arched (see Fig. 2 (b)). The shape of eyebrows is easily described by geometrical features.



(a) (b)

Figure 2: Two examples of AUs. (a) and (b) describe AU9 and AU2 [11].

Consequently, the necessary spatial features include both appearance features and geometrical features. Firstly, Robust Principal Component Analysis (RPCA) is employed to extract preliminary appearance features. RPCA takes advantage of the fact that the data are characterized by low-rank subspaces [7]. It disassembles the data matrix $\mathbf{D}$ into two parts:

$$\mathbf{D} = \mathbf{A} + \mathbf{E} \tag{1}$$

where $\mathbf{A}$ is the deserved data in a low-rank subspace that contains principal information, and $\mathbf{E}$ is the error term, usually treated as noise. However, Wang *et.al* [51] treated $\mathbf{E}$ as deserved spatial features of micro-expressions. Meanwhile, since $\mathbf{A}$ includes the identity information, it is regarded as noise

for micro-expression recognition. In [18], the authors also used RPCA as the pre-processing step to improve micro-expression recognition performance.

In order to get $\mathbf{E}$, Eq.(1) can be rewritten as follow

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 \qquad \text{subject to} \qquad \mathbf{D} = \mathbf{A} + \mathbf{E} \tag{2}$$

where $\mathbf{D}$ is a video clip with micro-expressions, where each column represents each frame of the video. $\|\cdot\|_*$ represents the nuclear norm, i.e., the sum of its singular values, and $\|\cdot\|_1$ denotes ℓ_1-norm, which is the sum of the absolute values of matrix entries. $\lambda > 0$ is the regularizing parameter. For details, please refer to [51]. Fig. 3 shows the original frames (Figs. 3(a) - 3(i)) from SAMM and the corresponding $\mathbf{E}$ (Figs. 3(d) - 3(l)). From the figures at the bottom of Fig. 3, we can intuitively see that identity information has been removed in $\mathbf{E}$s.The highlight in Fig. 3(e) illustrates muscle movement in the eyebrows area. The corrugator supercilii muscle contraction may cause this movement, which usually expresses a negative emotion [22, 6]. Meantime, a movement in the left corner of the mouth can be represented by the highlight in Fig. 3(k). The movement maybe caused by the zygomaticus muscle contraction and usually express a positive emotion [22, 6].

In $\mathbf{E}$, the vast majority of elements are zero, and the tiny minority of elements are non-zero. In contrast to zero elements, these non-zero elements partly describe the geometrical features of micro-expressions. Among these non-zero elements, the different values of elements indicate the intensity information. In other words, these non-zero elements partly represent appearance features of micro-expressions. Hence, $\mathbf{E}$ contains part of spatial features of micro-expressions and is regarded as the input of the spatial sub-network.

Although deep network technology has dramatically improved performances of face recognition [41], expression recognition [30], etcetera, these exceptional performances depend on using massive labeled data to fit millions of weights and biases of the deep network. However, for micro-expression recognition, labeled data is less than one thousand. Such a small amount of labeled data is caused by the difficulty of collecting and labeling micro-expression data.

For the collecting data process, there are three kinds of difficulties. **(1) Videos instead of images.** Unlike the distinctive facial movement of single-image macro-expression, micro-expressions are almost indistinguishable by a single frame. Hence, micro-expression analysis needs to be performed on video samples. **(2) Strict recording environment.** Since

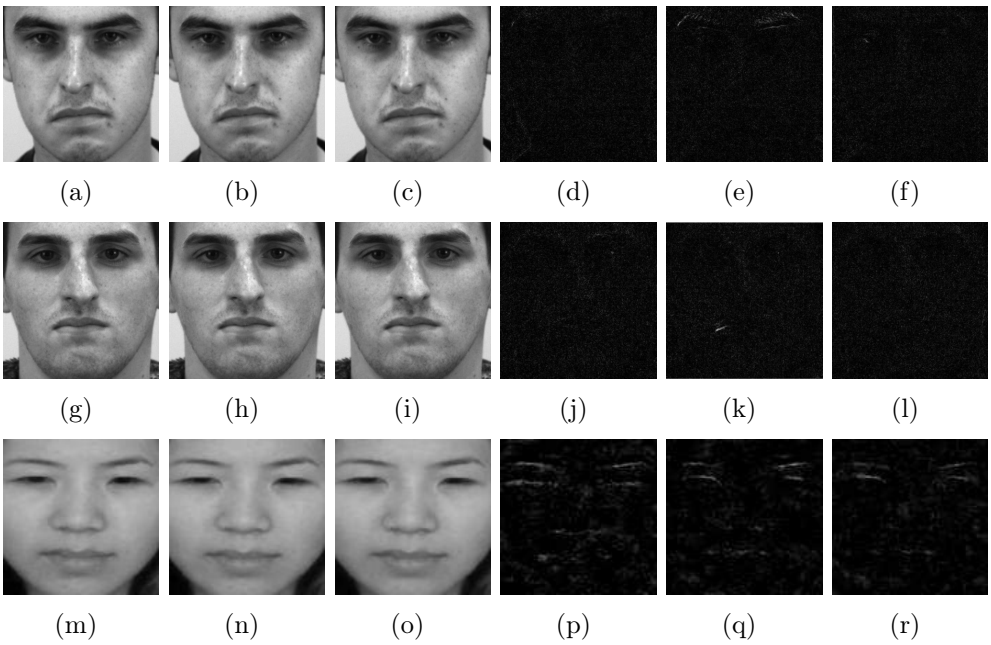


Figure 3: An example of sparse feature extraction. Fig. 3(a)-3(c) are the original micro-expression frames of angry videos of a Western male; Fig. 3(d)-3(f) are the corresponding extracted sparse features. Fig. 3(g)-3(i) are the original micro-expression frames of happy videos of another Western male; Fig. 3(j)-3(l) are the corresponding extracted sparse features. Fig. 3(m)-3(o) are the original micro-expression frames of happy videos of an Asian female; Fig. 3(p)-3(r)are the corresponding extracted sparse features.

micro-expression is a subtle facial movement, and current automatic analysis methods are not yet capable of performing experiments in nature, the sample collection requires a strictly controlled recording environment to avoid noises such as illumination variations. **(3) Difficulty inducing effective micro-expressions.** Since micro-expression appears when a person wants to hide his or her genuine emotion. Spontaneous samples are more appropriate than posed ones for real-life applications. Therefore, the eliciting process must be well designed, and the video used to induce micro-expression must contain strong emotional stimuli. Furthermore, since the participants are generally asked to be expressionless, this suppression makes micro-expressions extremely rare during the recording process.

For the labeling data process, there are three kinds of difficulties. **(1) The annotation is very laborious and time-consuming.** As micro-expression is a brief, subtle local facial movement, it is hard to be detected

in the video with naked human eyes. Moreover, the coder needs to be professionally trained, and the coding process requires at least two coders for reliable result control. **(2) Inconsistent sample classification criterion.** There are two kinds of micro-expression labels: Action Unit (AU) and emotion class. The relation between AU and emotion for micro-expression is still ambiguous. Besides, the emotion classification for micro-expression is unified, and samples may be categorized into three classes (positive, negative, and surprise), four classes (positive, negative, surprise, and others), six classes (six basic emotions) and so on. **(3) Masked expressions.** Due to the elicitation protocol, the participant may try to conceal the true emotion by covering it with blinking, smiling or other facial movements, i.e., micro-expression might be masked. It also increases the complexity of the annotation process.

Based on the above, due to the small amount of micro-expression samples, we use a shallow network to extract spatial features of micro-expression further. In some cases, the performance of shallow networks is not worse than the performance of deep networks. For example, Pan *et al.* [36] used a shallow network to predict salient areas in images and got similar results of using a deep network for the same prediction task. Wang *et al.* [53] used a shallow network parallel to a deep network for image super-resolution problems. Peng *et al.* [37] employed a shallow network with three convolutional layers for micro-expression recognition.

Here, we use AlexNet [24] with five convolutional layers as the backbone of the spatial sub-network. The kernel size of the first convolutional layer is reduced from 11×11 to 7×7 to avoid the over-fitting due to the micro-expression small samples size problem. Furthermore, the channel size of the fourth convolutional layer is reduced from 384 to 256. Zhang *et al.* [59] found that models without fully-connected layers have their activation map concentrated around the center object. Those with the fully-connected layers have activation maps that are more distributed. To get more distributed spatial features to better fuse with temporal features, we append a fully-connected layer with 4096 neurons to the spatial sub-network. Table 1 lists the kernel size, padding, and stride of each layer in the spatial sub-network. In Section 3, we also conducted some experiments to compare AlexNet backbone with VGG16 backbone and ResNet10 backbone. The experiments show that the AlexNet backbone gets better performance in most cases.

Table 1: Kernel size, padding and stride of each layers in the spatial sub-network

Layers	**Kernel size**	**Padding**	**Stride**
Conv1	$7 \times 7 \times 96$	1	2
Pool1	3×3	1	2
Conv2	$5 \times 5 \times 256$	2	2
Pool2	3×3	1	2
Conv3	$3 \times 3 \times 384$	1	1
Conv4	$3 \times 3 \times 256$	1	1
Conv5	$3 \times 3 \times 256$	1	1
Pool5	3×3	1	2
FC	4096	N/A	N/A

2.2. Temporal Sub-network

The intensity of micro-expressions is also very low in terms of facial muscles' movement [32]. Therefore, relying only on spatial features cannot provide enough discriminative information to recognize micro-expression effectively. Thus, we also designed a temporal sub-network to extract temporal features of micro-expressions. Recurrent Neural Network (RNN) [40] is a particular neural network to extract temporal features.

The standard RNN output $\mathbf{y}_t$ at a time step t is calculated using the following equations:

$$\mathbf{h}_t = tanh(\mathbf{W}_I \mathbf{x}_t + \mathbf{W}_H \mathbf{h}_{t-1}) \tag{3}$$

$$\mathbf{y}_t = tanh(\mathbf{W}_O \mathbf{h}_t) \tag{4}$$

where $\mathbf{W}_I$, $\mathbf{W}_H$, and $\mathbf{W}_O$ are the input, hidden, and output weight matrices, $\mathbf{x}_t$ tis the input vector at time step t, vectors $\mathbf{h}_t$ and $\mathbf{h}_{t-1}$ represent the hidden neuron activations at time steps t and $t-1$. For simplicity, neuron biases are omitted in the equations.

Next, we introduce Temporal Perception Information (TI) into the RNN unit and propose a novel Temporal Perception Unit (TPU) to construct the temporal sub-network.

TI is usually used to represent the temporal changes of videos, which can measure the movement intensity of every frame. TI is calculated based on the motion differential features, which refer to the difference in the exaction position of each pixel between two adjacent frames. The equation of motion differential features is as Eq.(5):

$$\mathbf{D}_t(i,j) = \mathbf{F}_t(i,j) - \mathbf{F}_{t-1}(i,j) \tag{5}$$

where $\mathbf{F}_t \in \mathbb{R}^{I \times J}$ is the t-th frame of a micro-expression video clip, (i,j) represents the coordinates of each pixels in the frame. Eq.(5) calculates the difference between adjacent frames. The adjacent frame difference shows the movement information between them and makes the movement between adjacent frames more obvious. Given a video clip with T frames, its TI is calculated by Eq.(6).

$$\mathrm{TI} = \max_{t=1}^{T-1} \left(\mathrm{std}_{i=1,j=1}^{I,J}(\mathbf{D}_t(i,j)) \right) \tag{6}$$

where $std(\mathbf{D})$ means the standard deviation of the matrix $\mathbf{D}$. The larger the value of TI is, the more the motion between adjacent frames is. In the concept of TI, the standard deviation of the adjacent frame difference plays an important role, and is used to modulate the hidden state of TPU.

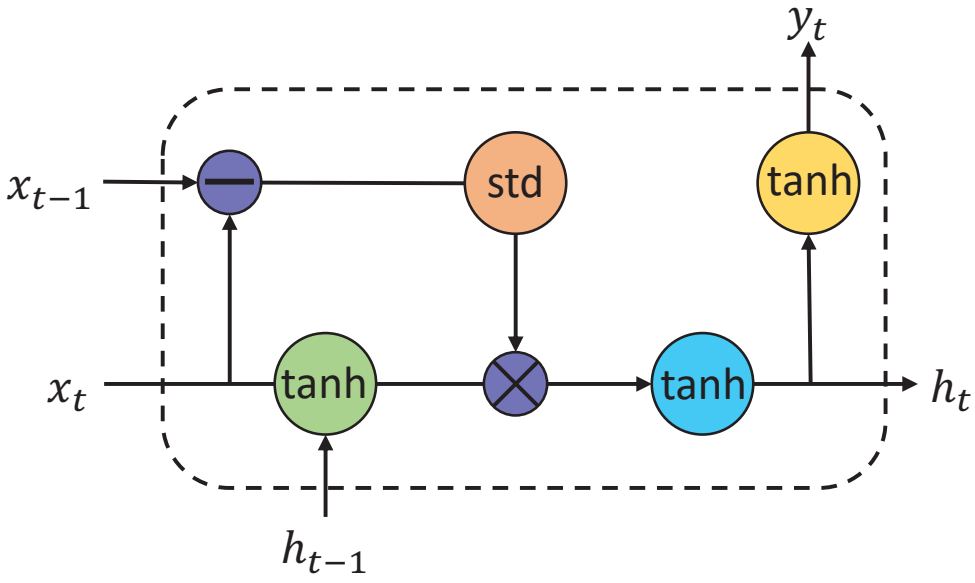


Figure 4: TPU's architecture based on TI.

Drawing on the idea of 2D image differencing, we perform a similar operation on 1D features. Fig.4 shows the structure of the proposed TPU. Compared with RNN, inputs of TPU include not only the current input $\mathbf{x}_t$

and the previous state $\mathbf{h}_t$ but also the previous input $\mathbf{x}_{t-1}$. The temporal perception coefficient tp is defined by Eq.(7) and is used to modulate the hidden state $\widetilde{\mathbf{h}}_t$.

$$tp = std(\mathbf{x}_t - \mathbf{x}_{t-1}) \tag{7}$$

To a certain degree, tp includes spatial motion information between the two adjacent frames. To better get the spatial information, the original frame image is directly vectorized as $\mathbf{x}_t$. The hidden state $\widetilde{\mathbf{h}}_t$ is calculated by the current input $\mathbf{x}_t$ and the previous state $\mathbf{h}_{t-1}$

$$\widetilde{\mathbf{h}}_t = tanh(\mathbf{W}_I \mathbf{x}_t + \mathbf{W}_H \mathbf{h}_{t-1}) \tag{8}$$

Then the cell state $\mathbf{h}_t$ is calculated by

$$\mathbf{h}_t = tanh(\mathbf{W}_{\widetilde{h}} \cdot (tp \times \widetilde{\mathbf{h}}_t)) \tag{9}$$

In Eq.(9), tp is used to modulate the hidden state $\widetilde{\mathbf{h}}_t$. This further embeds the spatial information into the temporal information. Similar to RNN, the output of TPU $\mathbf{y}_t$, is completed by Eq.(4). For better fusion in the following step, a full-connected layer is also appended at the end of the temporal sub-network. In addition, we also discuss the rationality behind the uses of standard deviation of difference between adjacent frame helps in learning the temporal information. Specifically, We use 'SAMM007_6_2' in the SAMM database as a toy example. We plotted the relationship curve between tp in Eq. 7 and frames, as shown in the following figure. The figure shows that the larger tp generally occurs in the middle of the clip. Features of the intermediate time interval are discriminative for micro-expression recognition. Meanwhile, the change in tp value can accurately reflect this variation.

In addition to the high probability that the most characteristic features are in the middle of micro-expression clips from the perspective of feature computation, we also demonstrate this from human visual perception. Specifically, we recruited subjects to watch the micro-expression clips and asked them to label one frame per video where the micro-expressions were most significant, i.e., the frame with the largest change in facial muscle movement. Twenty-four micro-expression clips from the SAMM database were selected, with three videos for each of the eight emotion types in the database. Furthermore, to remove the influence of facial familiarity on coders, the subject presenting the micro-expressions is different in each micro-expression clip we chose. Finally, among 24 micro-expression clips, the most significant micro-expression frames of 19 clips were annotated in the middle of the interval.

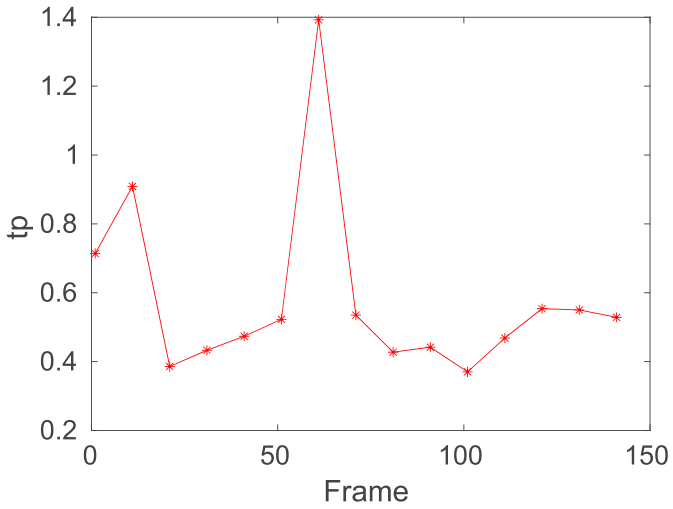


Figure 5: Toy example of the rationality behind the uses of standard deviation of difference between adjacent frame helps in learning the temporal information. (SAMM007_6_2)

In brief, the difference of the frames in the middle section matters in reflecting the movement variations of micro-expressions. Whereas t_p can capture this difference well, thus helping the network to learn the temporal dynamic change characteristics of micro-expressions.

2.3. Element-wise Addition with 1×1 *Convolutional Kernel Fusion Model*

The purpose of fusion is to make the spatial and temporal features complement each other. As mentioned above, micro-expression usually corresponds to multiple AUs simultaneously. These AUs maybe locate at different facial areas with different spatial distributions. Each of them has a specific temporal trajectory. Due to different spatial distributions, their temporal trajectories are separated. To effectively integrate temporal information and sparse spatial features, the model needs to use a feature combination method for feature fusion.

As shown in Fig. 1, before the feature fusion, the original video data is separately fed to the recurrent network and the convolutional network. After temporal and spatial features fusion, the fused features need to be fully converged and further refined into high-level features by convolution and fully-connect operations.

In addition, in the recurrent network part, each TPU unit has a single time step predicted output $\mathbf{y}_t$ and a hidden state h_t of the time step t. In order

to ensure that the data information can be fully utilized during the network training process, the model concatenates the output generated by each time step according to $\mathbf{F} = [\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_T]$, where T denotes the number of frames in each video. Thus, the output of the recurrent network can be represented as a 3rd order tensor $\mathcal{F}_{\text{temporal}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, where d_1, d_2 and d_3 are sample size, number of video streams, and feature dimension respectively. $\mathcal{F}_{\text{temporal}}$ will be reshaped into a temporal feature matrix $\mathbf{F}_{\text{temporal}} \in \mathbb{R}^{(d_1 \times d_2) \times d_3}$. Similarly, the output of the convolutional network can be represented as a 4th order tensor $\mathcal{F}_{\text{spatial}} \in \mathbb{R}^{e_1 \times e_2 \times e_3 \times e_4}$, which will be reshaped into a spatial feature matrix $\mathbf{F}_{\text{spatial}} \in \mathbb{R}^{e_1 \times (e_2 \times e_3 \times e_4)}$, where $e_1 = d_1 \times d_2$ and $d_3 = e_2 \times e_3 \times e_4$. And the features after fusion is:

$$\mathbf{F} = \mathbf{F}_{\text{temporal}} + \mathbf{F}_{\text{spatial}} \tag{10}$$

where $\mathbf{F}$ is a fusion feature matrix. After feature fusion, the fusion feature matrix $\mathbf{F}$ is reshaped into $e_1 \times e_2 \times e_3 \times e_4$ shape. As illustrated in Fig. 1, after temporal feature learning, we obtained a feature map of size 80×4096, and after spatial feature learning, we obtained a similar feature map of size 80×4096. We performed feature fusion using element-wise addition, where corresponding elements from both feature maps were added together, resulting in a combined feature map of size 80×4096. It is fed into a convolutional layer with 1×1 kernels to further fuse features. Finally, a fully-connected layer is used to classify.

In the proposed PSN, We use softmax as the loss function, which is widely used in the field of deep learning multi-classification. Eq.(11) shows the softmax loss function:

$$L = -\frac{1}{N} \sum\nolimits_{n=1}^{N} \log(\hat{p}_{n,j}) \tag{11}$$

where N is the number of a batch samples, $\hat{p}_{n,j}$ is the classification probability:

$$\hat{p}_j = \frac{e^{x_j}}{\sum_{i=1}^{K} e^{x_i}}, i, j \in K \tag{12}$$

where K is the number of categories.

3. Experiments

In this section, we conduct a series of experiments to evaluate the performance of PSN. After introducing the database, pre-process, and environment

configuration, we compare PSN with state-of-the-art methods to show the superiority of PSN on micro-expression recognition. Besides, the ablation study and training performance analysis are presented to verify the advantage of our proposed PSN architecture.

3.1. Databases

In the experiments, we will use four micro-expression databases: CASME, CASME II, CAS(ME)^2, and SAMM.

CASME: CASME database includes 189 spontaneous facial micro-expressions recorded by two 60 fps cameras. It is divided into two sets: Set A and Set B. The samples in Set A were recorded with a BenQ M31 consumer camera with 60fps, with the resolution set to 1280 $\times$720 pixels. The participants were recorded in natural light. The samples in Set B were recorded with a Point Grey GRAS-03K2C industrial camera with 60fps, with the resolution set to 640 $\times$ 480 pixels[50].

CASME II : Due to the low frame rate of CASME, short-duration facial movements may be missed. So CASME II [56] used a camera with 200 fps to record micro-expressions, and it contains 255 micro-expression samples from 26 subjects, including five categories (happiness, surprise, disgust, repression, and others). Among the 255 samples, there are seven samples, which RPCA can not extract. So, only 248 samples in CASME II are used in the experiments.

CAS(ME)^2: CAS(ME)^2 includes both spontaneous macro-expressions and micro-expressions in long videos (part A) and cropped expression samples with frames from onset to offset (part B) for automatic macro-expression and micro-expression spotting and recognition. To be consistent with CASME and CASME II, we split out the 57 micro-expression video clips from the long videos in CAS(ME)^2 for PSN performance validation. Among the 57 micro-expression samples, there are three samples that RPCA can not extract. So, only 54 samples in CAS(ME)^2 are used in the experiments.

SAMM: Davison *et al.* established the SAMM (Spontaneous Actions and Micro Movements) database in order to increase the ethnicity and age span of micro-expression samples. The SAMM database is a high-resolution database containing 159 spontaneous micro-movements induced with large variability in demographics. It includes seven basic emotions. The video samples are recorded by a camera with 200 fps.

The annotation criteria used in these databases are different, which results in some inconsistencies for labeling emotions. For instance, in CAS(ME)^2,

the appearance of AU1 and AU2 in the video clip disgust1_3 of the subject 2 means disgust, but in CASME II, the same AU movement was labeled as surprise in the video clip EP02_05 of the subject 12. Therefore, in our experiment, we re-labeled these facial expressions simply by the objective AU combinations according to Table 2 [38], inspired by [3]. Classification based on AU combinations eliminates human reporting bias and relies on ground-truth muscle movements for each micro-expression video. As a result, it can make the comparison of feature representation and recognition techniques fairer. There are 7 categories in Table 2. Categories I-VI refer to happiness, surprise, anger, disgust, sadness, and fear. Category VII relates to contempt and other AUs without any emotional link in EMFACS [13].

Furthermore, we classify the samples based on AU combinations into four categories, i.e., positive, negative, surprise, and other. Discovering negative emotions hidden under positive expressions or vice versa, such as covering the dagger with a smile, can help in lie recognition or emotional understanding of interpersonal interactions. Hence, there is more practical importance with this kind of classification. The criteria are listed in Table 3. Table 4 lists the numbers of each category in the consisted database.

Table 2: Subjective micro-expression classification and the corresponding AU combination

Category	Action Units	Related Emotion
I	AU6,AU12,AU6+AU12, AU6+AU7+AU12, AU7+AU12	Happiness
II	AU1+AU2,AU5,AU25, AU1+AU2+AU25,AU25+AU26, AU5+AU24	Surprise
III	AU23,AU4,AU4+AU7,AU4+AU5+AU7, AU17+AU24,AU4+AU6+AU7, AU4+AU38	Anger
IV	AU10,AU9,AU4+AU9,AU4+AU40, AU4+AU5+AU40,AU4+AU7+AU9, AU4+AU9+AU17,AU4+AU7+AU10, AU4+AU5+AU7+AU9,AU7+AU10	Disgust
V	AU1,AU15,AU1+AU4,AU6+AU15, AU15+AU17	Sadness
VI	AU1+AU2+AU4,AU20	Fear
VII	Others	Others

There are two reasons for evaluating the proposed methods on the con-

Table 3: 4 emotion categories of micro-expression and the corresponding criteria based on AU

	Category	Stands
0	Positive	AUs related to Happiness(at least AU6 or AU12)
1	Negative	AUs related to Anger, Disgust, Sadness and Fear
2	Surprise	At least AU1+AU2, AU25 and AU2
3	Others	Other Action Units on the face

Table 4: The numbers of each category in the consisted database.

	Positive	Negative	Surprise	Others	Total
CASME	8	100	14	67	189
CASME II	24	132	15	77	248
CAS(ME)^2	2	18	7	27	54
SAMM	21	30	14	94	159
Total	55	280	50	265	650

sisted database. First, the number of micro-expression samples in a single database is small, with the largest CASME II having only 256 samples. We want to increase the sample size by combining the databases and thus improve deep learning performance. Second, micro-expressions are weak and transient facial expressions, resulting in challenging feature learning. A micro-expression recognition model trained in a single database can hardly get similar recognition results in other databases. We trained the model by the cross-database approach to improve the algorithm's robustness. In addition, this combination helps alleviate the issue of sample imbalance to some extent. For example, before combining the datasets, the CAS(ME)^2 dataset contained only 2 samples for positive emotions. After merging with other databases, the number of positive emotion samples increased to 55, which significantly enhanced the model's ability to learn from a more balanced set of examples.

3.2. Preprocessing

Before being input to the network, the data should be preprocessed to remove irrelevant factors, with three steps as follows.

(1) Face segmentation and alignment processing Each frame of micro-expression videos is a bust image. If the bust images are fed directly into the network, the network will be affected by much irrelevant information

(such as parts of the body under the face and parts of the background). Thus, face detection and cropping are performed so that the network concentrates on extracting facial features.

(2) Normalization on spatial and temporal domains Due to different sizes of frames in the four databases mentioned above and the limit of GPU in our experimental environment, we resize each frame into 112×112 pixels. In this way, the batch size of the model could be enlarged. Furthermore, to fit the input size of TPU, the lengths of videos should be the same. In the four micro-expression databases, the longest length of videos is 142. Hence, we normalize all videos into 150 frames. The longer length means more columns of the matrix $\mathbf{E}$ in Eq. 2. The subtle motion information of micro-expressions extracted by RPCA may be better.

Finally, each video includes 16 frames, which are uniformly sampled from these 150 frames after normalization. In other words, the videos with 16 frames consist of the $1, 11, 21, \ldots, 141, 150$ frames of the videos with 150 frames. The reason for the down-sampling is the limited memory of GPUs. Moreover, the longer length of videos does not necessarily improve the model's performance [49]. Besides, regarding the choice of uniform sampling, it is because we then used RPCA to remove identity information and retain the action information. Specifically, the RPCA process requires a consistent time interval between frames to accurately extract facial movement from the neutral face to the occurrence of the expression and finally to the disappearance of the expression. For this reason, uniform sampling was the most suitable method for our approach. Moreover, regarding the concern about the possibility of missing the apex frame, since micro-expressions are characterized by very subtle movement in facial expressions, the difference between the apex frame and the surrounding frames is not very large. Sampled frames from the middle time period provide sufficient features to accurately extract micro-expression feature, even without including the apex frame.

(3) Training set and test set All data are randomly shuffled and divided into ten groups. In each group, the proportion of each class is consistent with that of the class in all samples. Each group includes 65 samples. There are ten folds in our experiments, and the leave-one-fold-out cross-validation is utilized for micro-expression recognition performance analysis.

3.3. Environment and Metrics

The experimental environment is based on CAFFE deep learning framework platform, running on a server with 8 GPUs (1080Ti, driver version

390.87) and operating system version 14.04.3 LTS Ubuntu. CUDA and cudnn versions are 8.0.61 and 6021, respectively. The preprocessing was performed by Python2.7 and MATLAB 2016b. During training, we perform data augmentation techinique, such as rotation and flipping. Dropout is implemented in the classifier. Four GPUs with the same specifications are used for simultaneous training to improve efficiency.

The recognition rate is one of the metrics for performance evaluation. Compared with the recognition rate, since the F1-score is the harmonic mean between precision and recall, it can better measure the model performance in the case of unbalanced sample distribution for micro-expressions. For the details, refer to [47].

3.4. Result and Analysis

3.4.1. PSN outperforms state-of-the-art methods

Table 5 shows the performance comparison between PSN and some state-of-the-art (SOTA) methods including LBP-SIP [52], Block Division Convolutional Network (BDCNN) [1], KFC [44], Recurrent Convolutional Network (RCN) [55], STSTNet [29], Feature Refinement (FR) [64], MERANet [14], FG-AU-Fusion [25] and MERSiamC3D [63]. LBP-SIP is a spatio-temporal descriptor, which removes the six repetitive coding intersections of LBP-TOP. In LBP-SIP experiment, the facial region is divided into 6×6, 7×7, and 8×8 blocks, respectively. Histograms are extracted from these blocks and then fed into an SVM with the RBF kernel.

On the other hand, optical flow calculates the motion between two image frames. The spatial and temporal coordinates are both involved in the optical flow feature. Some deep learning methods utilized this spatio-temporal feature to recognize micro-expressions, such as STSTNet, RCN, and FR. FR emphasizes the salient and discriminative expression-specific feature learning using two channels of optical flow. Meanwhile, STSTnet could extract discriminative high-level features and details of micro-expression based on the combination of optical flow and optical strain. Furthermore, through the same spatio-temporal feature, the recurrent RCN explores the shallower-architecture and lower-resolution input data, shrinking model, and input complexities simultaneously. RCN has three extensions, i.e., a-extension, s-extension, and w-extension. They denote attention unit extension, shortcut connection extension, and wide expansion extension. In addition to these three methods, BDCNN, KFC, MERANet, FG-AU-Fusion, and MER-

SiamC3D are all recently published SOTA methods that focus on spatiotemporal motion changes in micro-expression recognition.

All the methods are reproduced based on the same experimental configuration for a fair comparison. Recognition rate and F1 scores of 10-fold-cross validation is listed in Table 5. PSN has the highest recognition rate 64.92% and the best F1 score 0.5211.

Table 5: Recognition rate and F1 score of PSN and other state-of-the-art methods. HF-ML and DL represent the Handcrafted feature based machine leaning and Deep leaning, respectively.

Category	Methods		Recognition Rate (%)	F1 score
HF-ML	LBP-SIP [52]	6×6	53.08	0.2891
		7×7	55.08	0.4016
		8×8	54.62	0.3910
	KFC [44]		26.31	0.2284
	BDCNN [1]		49.84	0.2713
DL	RCN [55]	a-extension	50.62	0.2876
		s-extension	49.69	0.2909
		w-extension	53.54	0.3563
	STSTNet [29]		53.23	0.2895
	FR [64]		56.00	0.3727
	MERANet [14]		54.00	0.4183
	FG-AU-Fusion [25]		57.23	0.4474
	MERSiamC3D [63]		59.38	0.4670
	PSN		64.92	0.5211

As listed in Table 6, we present the results of these two metrics (UAR and UF1) for the CASME, CASME II, CAS(ME)^2, and SAMM datasets, along with comparisons to SOTA methods. Our approach achieves the best performance on most individual databases and database combinations, demonstrating that our method improves the accuracy and robustness of micro-expression recognition by effectively learning both the dynamic and spatial features of micro-expressions.

Moreover, we have also conducted a deeper exploration of our method's performance across different datasets. The SAMM dataset, which includes individuals from multiple ethnic groups, has a relatively higher proportion of Caucasian subjects. However, SAMM constitutes only about 24% of the total sample size in our study. It is worth noting that, apart from the SAMM

Table 6: Performance comparison with SOTA methods among single databases.

Method	Combined Database		CASME		CASME II		CAS(ME)^2		SAMM	
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
LBP-SIP 7×7 [52]	0.3452	0.4016	0.2882	0.2704	0.3041	0.2595	0.2731	0.2432	0.2861	0.2575
KFC [44]	0.2603	0.2284	0.2924	0.2164	0.2308	0.2026	**0.4509**	0.3666	0.2657	0.2316
BDCNN [1]	0.2974	0.2713	0.2917	0.2686	0.3047	0.2745	0.2593	0.2143	0.3066	0.2553
RCN-w [55]	0.3237	0.3563	0.3104	0.2856	0.3019	0.2732	0.3657	0.3086	0.2244	0.2065
STSTNet [29]	0.3167	0.2895	0.2887	0.2724	0.2614	0.2633	0.2639	0.2175	0.2445	0.2158
FR [64]	0.3563	0.3727	0.3031	0.2869	0.4044	0.4066	0.3102	0.2638	0.2841	0.2635
MERANet [14]	0.4139	0.4183	**0.4272**	**0.4241**	0.3492	0.3425	**0.4944**	**0.4546**	**0.4760**	0.4521
FG-AU-Fusion [25]	0.4387	0.4474	0.3998	0.3868	0.4175	0.4210	0.4054	0.4111	**0.5055**	**0.4843**
MERSiamC3D [63]	**0.4473**	**0.4670**	0.3829	0.4015	**0.4618**	**0.4552**	0.3306	0.3282	0.4297	**0.4548**
PSN	**0.4871**	**0.5211**	**0.4349**	**0.4591**	**0.4347**	**0.4685**	0.4481	**0.4295**	0.4378	0.4545

dataset, all other datasets include subjects from China. We found that, compared to other methods (which involve training and testing on the SAMM database with model fine-tuning), our method did not achieve the best performance, but only relatively better results. However, on the other databases (with Chinese subjects), our method achieved the best performance. This suggests that ethnicity does have an impact on micro-expression recognition performance, particularly in a generalized model.

Table 7 lists seven confusion matrices of LBP-SIP, KFC, BDCNN, RCN, STSNet, FR, MERANet, FG-AU-Fusion, MERSiamC3D and the proposed PSN. For LBP-SIP and RCN, the confusion matrices come from at the highest recognition rate cases. From the table, we can see that PSN can get the highest recognition rates of *Negative* categories, while the recognition rates of *Positive*, *Surprise*, and *Other* three categories are also relatively advanced. Meantime, the unbalanced distribution of samples not only affects our method but also significantly impacts the performance of other SOTA methods in recognizing these four categories, with significant differences.

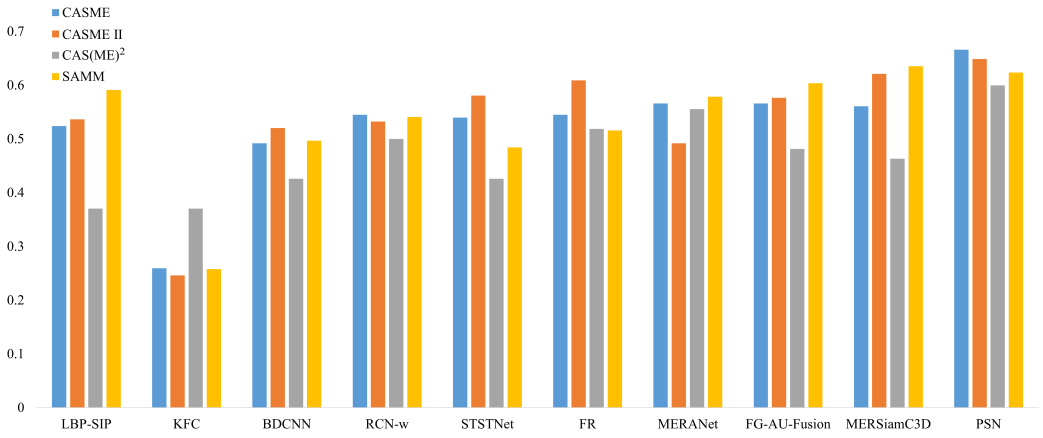
Table 8 lists these confusion matrices for each of the four different micro-expression databases. Similar conclusions can be drawn from SAMM database. For SAMM, the miss rates classified into *Other* are high. In these high miss rates, the miss rates of PSN are the lowest. Fig. 6 shows recognition rate, true positive rate, and true negative rate of PSN and other state-of-the-art methods on four micro-expression databases. From Fig. 6, we can see that the performances of PSN are better than those of other state-of-the-art methods in most cases.

Table 7: Confusion matrices for PSN and other state-of-the-art methods. P, N, S, and R denote *Positive*, *Negative*, *Surprise*, and *Others*, respectively. The number in parentheses indicates the number of samples.

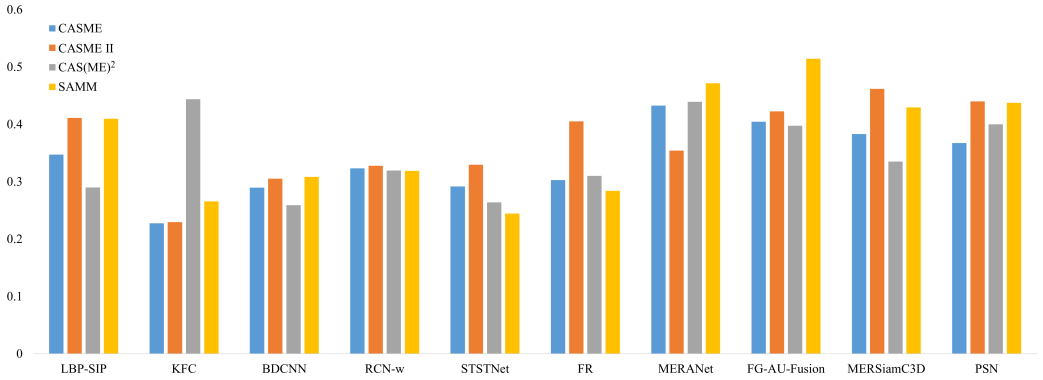
	LBP-SIP 7 × 7				KFC				BDCNN				RCN-w				STSTNet			
	P	N	S	O	P	N	S	O	P	N	S	O	P	N	S	O	P	N	S	O
P (55)	**3**	25	1	26	**13**	17	17	8	**0**	11	0	44	**1**	14	0	40	**0**	15	0	40
N (280)	1	**210**	4	65	84	**91**	72	33	0	**163**	0	117	0	**162**	0	118	0	**192**	0	88
S (50)	1	21	**3**	25	15	13	**14**	8	0	26	**0**	24	0	21	**0**	29	0	25	**0**	25
O (265)	1	119	3	**142**	83	61	68	**53**	0	104	0	**161**	1	79	0	**185**	3	108	0	**154**
	FR				MERANet				FG-AU-Fusion				MERSiamC3D				PSN			
	P	N	S	O	P	N	S	O	P	N	S	O	P	N	S	O	P	N	S	O
P (55)	**5**	12	1	37	**12**	13	3	27	**20**	14	1	20	**10**	15	0	30	**10**	20	0	25
N (280)	5	**181**	6	88	10	**181**	14	75	6	**171**	9	94	8	**191**	6	75	2	**226**	3	49
S (50)	0	26	**1**	23	3	14	**12**	21	4	18	**6**	22	1	19	**14**	16	1	21	**6**	22
O (265)	6	80	2	**177**	19	83	17	**146**	14	63	13	**175**	15	73	6	**171**	6	79	0	**180**

Table 8: Confusion matrices for PSN and other state-of-the-art methods. P, N, S, and R denote *Positive*, *Negative*, *Surprise*, and *Others*, respectively. The number in parentheses indicates the number of samples.

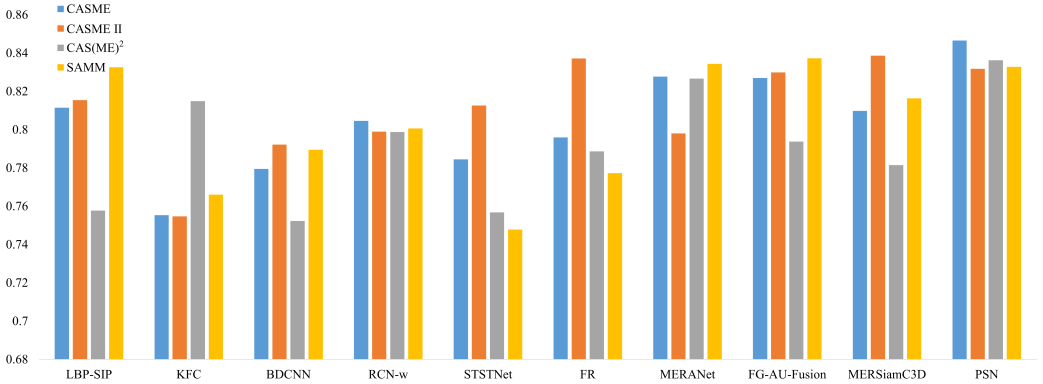
		LBP-SIP 7 × 7				KFC				BDCNN				RCN-w				STSTNet			
		P	N	S	O	P	N	S	O	P	N	S	O	P	N	S	O	P	N	S	O
CASME (189)	P (8)	**0.00**	0.50	0.00	0.50	**0.13**	0.30	0.29	0.28	**0.00**	0.23	0.00	0.77	**0.00**	0.38	0.00	0.63	**0.00**	0.38	0.00	0.63
	N (100)	0.00	**0.72**	0.00	0.28	0.50	**0.32**	0.21	0.30	0.00	**0.54**	0.00	0.46	0.00	**0.57**	0.00	0.43	0.00	**0.72**	0.00	0.28
	S (14)	0.00	0.57	**0.00**	0.43	0.25	0.26	**0.29**	0.24	0.00	0.67	**0.00**	0.33	0.00	0.64	**0.00**	0.36	0.00	0.64	**0.00**	0.36
	O (67)	0.00	0.57	0.00	**0.43**	0.13	0.12	0.21	**0.18**	0.00	0.38	0.00	**0.62**	0.00	0.33	0.00	**0.67**	0.03	0.55	0.00	**0.45**
CASME II (248)	P (24)	**0.00**	0.83	0.04	0.13	**0.33**	0.21	0.29	0.17	**0.00**	0.24	0.00	0.76	**0.00**	0.25	0.00	0.75	**0.00**	0.29	0.00	0.71
	N (132)	0.00	**0.98**	0.02	0.00	0.33	**0.30**	0.24	0.13	0.00	**0.62**	0.00	0.38	0.00	**0.64**	0.00	0.36	0.00	**0.77**	0.00	0.23
	S (15)	0.00	0.80	**0.20**	0.00	0.40	0.33	**0.13**	0.13	0.00	0.44	**0.00**	0.56	0.00	0.33	**0.00**	0.67	0.00	0.60	**0.00**	0.40
	O (77)	0.00	0.95	0.01	**0.04**	0.32	0.26	0.26	**0.16**	0.00	0.40	0.00	**0.60**	0.00	0.43	0.00	**0.57**	0.01	0.45	0.00	**0.55**
CAS(ME)^2 (57)	P (2)	**0.00**	0.50	0.00	0.50	**0.50**	0.50	0.00	0.00	**0.00**	0.00	0.00	1.00	**0.00**	0.00	0.00	1.00	**0.00**	0.00	0.00	1.00
	N (18)	0.00	**0.39**	0.11	0.50	0.33	**0.44**	0.22	0.00	0.00	**0.56**	0.00	0.44	0.00	**0.72**	0.00	0.28	0.00	**0.61**	0.00	0.39
	S (15)	0.00	0.00	**0.00**	1.00	0.14	0.29	**0.57**	0.00	0.00	0.43	**0.00**	0.57	0.00	0.29	**0.00**	0.71	0.00	0.43	**0.00**	0.57
	O (27)	0.00	0.22	0.07	**0.70**	0.26	0.15	0.33	**0.26**	0.00	0.52	0.00	**0.48**	0.00	0.26	0.00	**0.74**	0.04	0.52	0.00	**0.44**
SAMM (159)	P (21)	**0.14**	0.00	0.00	0.86	**0.14**	0.33	0.38	0.14	**0.00**	0.16	0.00	0.84	**0.05**	0.24	0.00	0.71	**0.00**	0.24	0.00	0.76
	N (30)	0.03	**0.03**	0.00	0.93	0.13	**0.40**	0.33	0.13	0.00	**0.58**	0.00	0.42	0.00	**0.03**	0.00	0.77	0.00	**0.23**	0.00	0.77
	S (14)	0.07	0.07	**0.00**	0.86	0.29	0.21	**0.29**	0.21	0.00	0.50	**0.00**	0.50	0.00	0.36	**0.00**	0.64	0.00	0.29	**0.00**	0.71
	O (94)	0.01	0.02	0.00	**0.97**	0.34	0.18	0.24	**0.23**	0.00	0.34	0.00	**0.66**	0.01	0.18	0.00	**0.81**	0.02	0.23	0.00	**0.74**



(a) recognition rate



(b) true positive rate



(c) true negative rate

Figure 6: Recognition rate, true positive rate, and true negative rate of PSN and other state-of-the-art methods on four micro-expression databases.

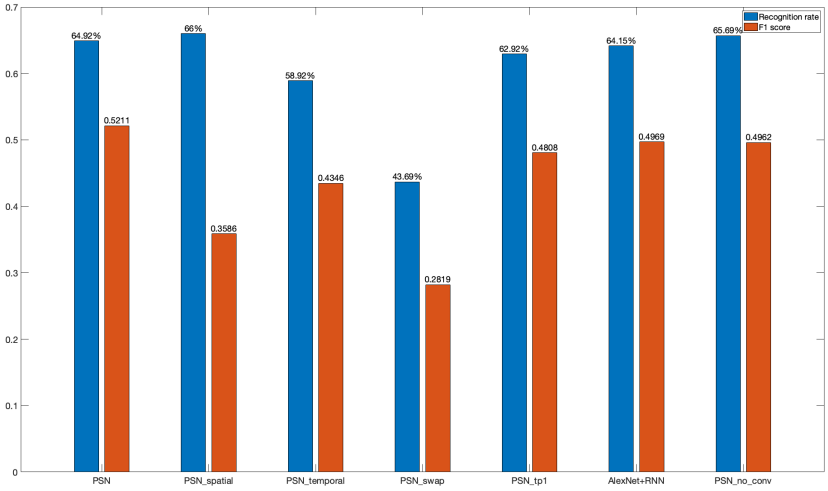


Figure 7: Ablation study of PSN. Recognition rate and F1 score are evaluated. The conditions are PSN, PSN only with spatial sub-network (PSN_spatial), PSN only with temporal sub-network (PSN_temporal), PSN with two sub-network inputs interchanged (PSN_swap), PSN with $t_p = 1$ in TPU (PSN_tp1), AlextNet + RNN, and PSN without convolution module in final fusion (PSN_no_conv), respectively.

3.4.2. Ablation Study

In order to study the necessity of each sub-module in the PSN, we conducted an ablation experiment. Fig. 7 shows the recognition rate and F1 score comparisons under different conditions which will be presented in detail in the following paragraphs.

Necessity of parallel spatio-temporal network architecture: First of all, with the aim of studying the indispensability of the parallel network structure, the control experiment includes the PSN only with the spatial sub-network (PSN_spatial) and only with the temporal sub-network (PSN_temporal) respectively. As illustrated in Fig 7, PSN_spatial gets the highest recognition rate 66%. However, its F1 score only is 0.3586. From the confusion matrix of PSN_spatial, we also can see that predicted results of PSN_spatial are *Negative* and *Others* classes, which have the largest sample size. And for *Positive* or *Surprise*, the recalls (hit rates) of PSN_spatial are zeros. Besides, the performance of PSN_spatial is the weakest because PSN_spatial does not extract the temporal characteristic of the micro-

expression, and it is difficult to recognize the micro-expression based on the subtle spatial features effectively. In addition, to prove that the inputs of the two sub-networks are suitable for each, we also crossed the inputs of the two sub-networks and conducted experiments (PSN_swap). The performances of PSN are severely affected. The main reason is that the input of the spatial sub-network does not go through RPCA, and a large amount of irrelevant facial information interferes with the classification ability of the network.

Backbone comparison for temporal sub-network: This part analyzes the advantage of TPU design on temporal feature extraction. The backbone of our proposed network is Alexnet and the proposed Temporal Perception Unit (TPU) is an RNN-based approach. The idea behind TPU is to adjust the weights of the RNN based on the feature difference between two frames. As shown in Fig. 7, by comparing the Alexnet+RNN and PSN, it could be seen that the temporal inception information in TPU improves the micro-expression recognition performance. The reason for the improvements is that the frame differences are utilized to compute the temporal inception information in TPU, and then it is used as the weight of the hidden states (see Eq. 7-9). This progress is similar to the RPCA function, and it can extract the sparse features in the temporal domain, which enriches the feature space so that the model can offset the negative impact of small samples. In addition, although AlexNet+RNN has slightly lower accuracy than our PSN method, our PSN method has significantly higher F1 scores. Regarding the uneven distribution of samples in micro-expression recognition, the F1 score can better reflect the superiority of the model. And the improvement of the Uf1 score indicates that our network's recognition ability has improved across different classifications, which further demonstrates the effectiveness and robustness of our PSN method.

Moreover, the importance of temporal perception coefficient t_p is explored by the comparison between PSN with $t_p = std(x_t - x_{t-1})$ (Eq.7) and $t_p = 1$ in TPU (PSN_tp1). As illustrated in Fig. 7, the recognition rate and F1 score of PSN outperform those of PSN_tp1. t_p reflects the motion difference on each pixel between two adjacent frames. This information provides the network with more time-varying characteristics of micro-expression, thereby enhancing the distinguishing ability of the classifier.

Backbone comparison for spatial sub-network and fusion methods comparison: As introduced in the previous section, the backbone of the spatial sub-network is AlexNet. We conduct an experiment to compare AlexNet with VGG16 [43] and ResNet10 [17]. In the meantime, in this ex-

periment, different fusion models are also compared to show the superiority of our method. They are element-wise addition, element-wise multiplication, and feature concatenation, respectively. Table 9 lists the mean of recognition rate, the standard deviation of recognition rate, and F1 score. The combination of AlexNet and feature concatenation achieves the highest mean of recognition rate 65.08%. The combination of AlexNet and element-wise addition achieves the second highest mean of recognition rate 64.92%. The difference between the top two means of recognition rate is only 0.16%. But the F1 score of the combination of AlexNet and element-wise addition is the highest. Moreover, the standard deviation of the combination of AlexNet and element-wise addition is less than the standard deviation of the combination of AlexNet and feature concatenation. Overall, we choose the combination of AlexNet and element-wise addition.

This combination is reasonable. AlexNet, with its shallower architecture and local feature extraction ability, is particularly well-suited for micro-expression recognition with small samples and weak. Furthermore, element-wise addition could avoid dimensionality expansion while retaining all the feature information effectively.

Table 9: F1 scores and recognition rates of different backbone with different fusion models. SUM, PROD, CONCAT mean element-wise addition, element-wise multiplication, and feature concatenation, respectively. RR_{mean} and RR_{std} denote the mean and the standard deviation of recognition rate, respectively.

Backbones	Fusion models	RR_{mean}	RR_{std}	F1 score
AlexNet	SUM	**64.92**	0.084	**0.5211**
	PROD	62.77	0.057	0.4702
	CONCAT	65.08	0.086	0.5073
VGG16	SUM	61.08	0.072	0.4799
	PROD	61.23	0.061	0.4659
	CONCAT	61.23	0.071	0.4823
ResNet10	SUM	62.15	0.056	0.4902
	PROD	63.38	0.059	0.4671
	CONCAT	61.54	**0.042**	0.4510

Weight reduction through convolution module in final fusion: The final analysis of the architecture of PSN is to prove the necessity of the convolution module in the fusion process. The comparison experiment is conducted between PSN with and without convolution module (PSN_no_conv).

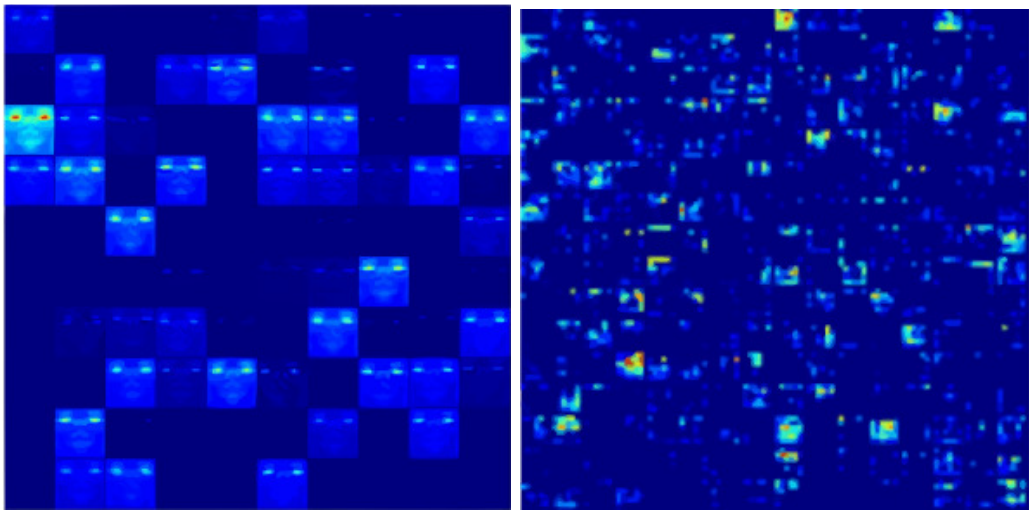
As shown in Fig. 7, PSN with convolution modules has better recognition performance. The convolution module is placed between the feature concatenation and the fully connected layer (See Fig. 1). Through a one-dimensional convolution feature conversion, the number of weights in the network is greatly reduced. Compared with directly inputting the fused features into the fully connected layer, the number of parameters is reduced from 20,976,640 $(64 \times (4096 + 1) \times 80)$ to 82,240 $(4 \times (256 + 1) \times 80)$. Thus, such a setting reduces the difficulty of training, avoids over-fitting, and improves recognition performance.

3.5. Discussion

Feature map analysis for spatial sub-network Furthermore, the feature extraction capacity of the spatial sub-network is analyzed. Fig. 8 shows the spatial features of the first and fifth convolutional layers. Generally, the first layers of the network extract some basic features, such as shape, etc. The later layers extract more advanced features relevant to the task [58]. Since our training set is face images, facial morphological shapes are basic features in the proposed network. Therefore, the first layer of the proposed network extracts facial morphological features. The network's comprehensive feedback on spatiotemporal features in the fifth layer makes the spatial features progressively sparse, and only valuable information is retained for the final classification. With the advanced feature extraction process, spatial feature extraction gradually obtains sparse features from the inputs.

Recognition rate analysis during PSN training stage: For further analysis of network training performance, Fig. 9(a) shows the trend of recognition rate during the training phase. The recognition rate has an uptrend at the beginning of the testing phase, and it gets stable at 30 iterations, which means the model converges ultimately during the training phase.

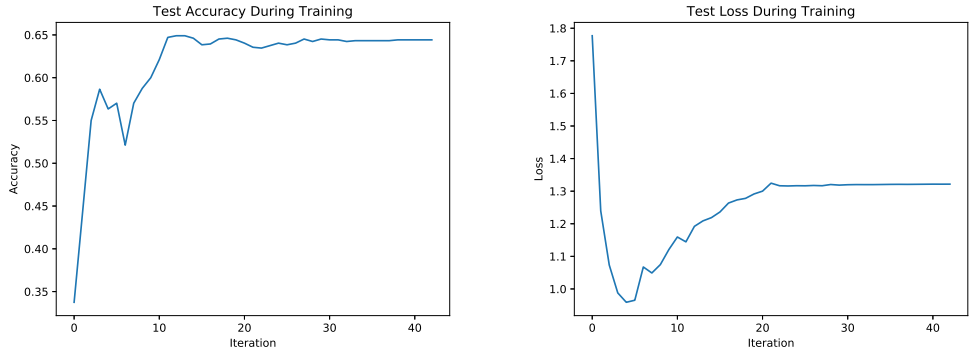
Impact of imbalanced sample distribution on recognition performance: As listed in Table 7, all methods have a relatively weak ability to recognize *Positive* and *Surprise* samples. Besides, the testing loss trend of PSN is shown in fig.9(b). The loss value should have been on a downward trend, but when it reaches the minimum, it continues to rise and then stabilize. Through the analysis of the sample size, the reason for the poor recognition performance of *Positive* and *Surprise* is that the data size of the above two categories is too small compared with the other two categories. Furthermore, this also makes the loss eventually rise slightly. According to the recognition results of *Negative* and *Other*, if the number of samples of



(a) the first layer (b) the fifth layer

Figure 8: Spatial feature map of the first and fifth convolutional layers

Positive and *Surprise* can be expanded, PSN could achieve better performance in these two emotion classifications.

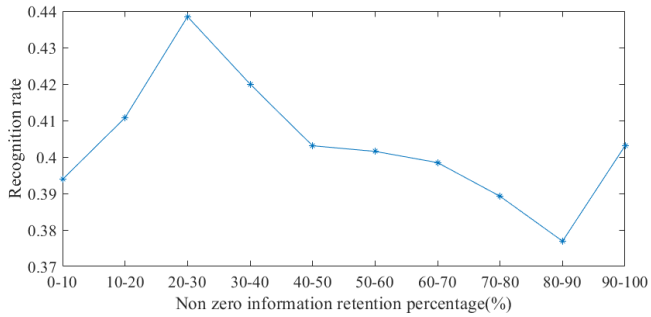


(a) Test recognition rate (Accuracy) during training of PSN.

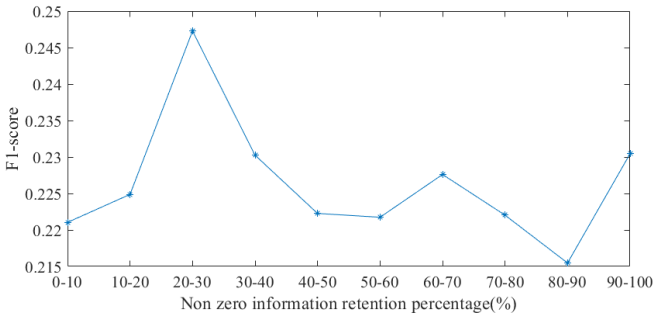
(b) Loss trend of PSN

Figure 9: Curves of recognition rate and loss of PSN.

The difference between the TPU and the optical flow based methods: The optical flow based methods extract different local features between adjacent frames. While the standard deviation of the adjacent frame



(a) recognition rate



(b) F1-score

Figure 10: Comparison of micro-expression recognition based on different RPCA non-zero information retention ratios. (a) Recognition rate; (b) F1-score.

difference is a different global feature between adjacent frames. It is further used to modulate the hidden state of TPU.

Analysis on the necessity of removing identity information through RPCA: First, we performed a parameter analysis on the RPCA process. Specifically, we set the RPCA parameters to retain different percentages of non-zero information on the face image (0-10%, 10-20%, etc.). We then use apex frames at different retention levels for micro-expression recognition. As illustrated in Fig. 10, the recognition rate of micro-expressions is highest when 20-30% of the information is retained. This is because facial identity information accounts for a large portion of the facial image, which can interfere with the model's ability to extract micro-expression features that only represent a small percentage of the image. Thus, retaining too much irrelevant information can lead to lower micro-expression recognition rates.

Second, we visually inspected the processed images and found that the

facial action information was well preserved while harmful inherent features were eliminated. We have included a GIF video of the processed images in the supplementary material. These experimental results could provide strong evidence of the effectiveness of our approach in separating identity information while retaining dynamic information for accurate micro-expression recognition.

4. Conclusion

For micro-expression recognition, this paper proposes the Parallel Spatiotemporal Network (PSN), which includes two sub-networks to extract spatial and temporal features, respectively. One is a shallow spatial network with subtle motion information as the input. Another is a temporal network with a novel temporal feature extraction unit TPU extracts sparse temporal features of micro-expressions.

In the paper, we can again see the importance of temporal information for micro-expression recognition. Moreover, the introduction of temporal perception information further improves the performance of micro-expression recognition. In the spatiotemporal fusion, 1×1 convolution also contributes to the improvement.

However, the subtle motion information was extracted by RPCA with some empirical parameters. In future research, we will design a network to extract subtle motion information and implement an end-to-end spatiotemporal network to recognize micro-expression.

Besides, research related to data imbalance, sample diversity, and generalizability needs further exploration. On one hand, this could enhance the performance of micro-expression recognition, and on the other hand, it would help improve the generalization ability of micro-expression recognition.

Acknowledgment

This work is supported, in part, by grants from the National Natural Science Foundation of China (62476269, 62276252, 62106256), and in part, by a grant from the Youth Innovation Promotion Association CAS.

References

[1] Chen, B., Liu, K.H., Xu, Y., Wu, Q.Q., Yao, J.F., 2022. Block division convolutional network with implicit deep features augmentation for micro-expression recognition. IEEE Transactions on Multimedia .

[2] Corneanu, C.A., Simón, M.O., Cohn, J.F., Guerrero, S.E., 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE transactions on pattern analysis and machine intelligence 38, 1548–1568.

[3] Davison, A., Merghani, W., Yap, M., 2018. Objective classes for micro-facial expression recognition. Journal of Imaging 4, 119.

[4] Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H., 2016. SAMM: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing 9, 116–129.

[5] Davison, A.K., Yap, M.H., Costen, N., Tan, K., Lansley, C., Leightley, D., 2014. Micro-facial movements: an investigation on spatio-temporal descriptors, in: European conference on computer vision, Springer. p. 111–123.

[6] Dong, Z., Wang, G., Lu, S., Li, J., Yan, W., Wang, S.J., 2021. Spontaneous facial expressions and micro-expressions coding: From brain to face. Frontiers in Psychology 12.

[7] Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218.

[8] Ekman, P., 2002. Microexpression training tool (METT). San Francisco: University of California .

[9] Ekman, P., 2003. Darwin, deception, and facial expression. Annals of the New York Academy of Sciences 1000, 205–221.

[10] Ekman, P., 2009. Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition). WW Norton & Company.

[11] Ekman, P., Friesen, W., Hager, J., 2002. Facs investigator¡¯s guide. A Human Face .

[12] Ekman, P., Friesen, W.V., 1969. Nonverbal leakage and clues to deception. Psychiatry 32, 88–106.

[13] Ekman, P., Friesen, W.V., 1978. Facial action coding system: Investigator's guide. Consulting Psychologists Press.

[14] Gajjala, V.R., Reddy, S.P.T., Mukherjee, S., Dubey, S.R., 2021. Meranet: facial micro-expression recognition using 3d residual attention network, in: Proceedings of the twelfth Indian conference on computer vision, graphics and image processing, pp. 1–10.

[15] Gupta, P., 2021. MERASTC: Micro-expression recognition using effective feature encodings and 2d convolutional neural network. IEEE Transactions on Affective Computing .

[16] Haggard, E.A., Isaacs, K.S., 1966. Methods of Research in Psychotherapy. New York: Appleton-Century-Crofts. chapter Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. pp. 154–165.

[17] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

[18] Huang, X., Wang, S.J., Liu, X., Zhao, G., Feng, X., Pietikäinen, M., 2019. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Transactions on Affective Computing 10, 32–47.

[19] Huang, X., Wang, S.J., Zhao, G., Piteikainen, M., 2015. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection, in: Proceedings of the IEEE international conference on computer vision workshops, pp. 1–9.

[20] Huang, X., Zhao, G., 2017. Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern, in: the Frontiers and Advances in Data Science (FADS), 2017 International Conference on, IEEE. p. 159–164.

[21] Huang, X., Zhao, G., Hong, X., Zheng, W., Pietikäinen, M., 2016. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing 175, 564–578.

[22] Hubert, W., de Jong-Meyer, R., 1990. Psychophysiological response patterns to positive and negative film stimuli. Biological psychology 31, 73–93.

[23] Kim, D.H., Baddar, W.J., Ro, Y.M., 2016. Micro-expression recognition with expression-state constrained spatio-temporal feature representations, in: Proceedings of the 2016 ACM on Multimedia Conference, ACM. p. 382–386.

[24] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, pp. 1097–1105.

[25] Lei, L., Chen, T., Li, S., Li, J., 2021. 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, pp. 1571–1580.

[26] Li, Y., Huang, X., Zhao, G., 2020. Joint local and global information learning with single apex frame detection for micro-expression recognition. IEEE Transactions on Image Processing 30, 249–263.

[27] Li, Y., Huang, X., Zhao, G., 2021. Micro-expression action unit detection with spatial and channel attention. Neurocomputing 436, 221–231.

[28] Liong, S.T., Gan, Y., Zheng, D., Li, S.M., Xu, H.X., Zhang, H.Z., Lyu, R.K., Liu, K.H., 2020. Evaluation of the spatio-temporal features and gan for micro-expression recognition system. Journal of Signal Processing Systems , 1–21.

[29] Liong, S.T., Gan, Y.S., See, J., Khor, H.Q., Huang, Y.C., 2019. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE. pp. 1–5.

[30] Liu, X., Kumar, B., You, J., Jia, P., 2017. Adaptive deep metric learning for identity-aware facial expression recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE. pp. 20–29.

[31] Liu, Y.J., Li, B.J., Lai, Y.K., 2018. Sparse MDMO: Learning a discriminative feature for micro-expression recognition. IEEE Transactions on Affective Computing 12, 254–261.

[32] Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X., 2016. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing 7, 299–310.

[33] Mao, Q., Zhou, L., Zheng, W., Shao, X., Huang, X., 2022. Objective class-based micro-expression recognition under partial occlusion via region-inspired relation reasoning network. IEEE transactions on affective computing 13, 1998–2016.

[34] Matsumoto, D., Hwang, H., 2011. Evidence for training the ability to read microexpressions of emotion. Motivation and Emotion 35, 181–191.

[35] Mehrabian, A., 1968. Communication without words. Psychology Today 2.

[36] Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E., 2016. Shallow and deep convolutional networks for saliency prediction, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 598–606.

[37] Peng, M., Wang, C., Chen, T., Liu, G., Fu, X., 2017. Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in Psychology 8, 1745.

[38] Qu, F., Wang, S.J., Yan, W.J., Li, H., Wu, S., Fu, X., 2017. CAS(ME)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Transactions on Affective Computing 9, 424–436.

[39] Rosenfeld, J.P., 1995. Alternative views of bashore and rapp's (1993) alternatives to traditional polygraphy: A critique. Psychological Bulletin 117, 159–166.

[40] Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. nature 323, 533–536.

[41] Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 815–823.

[42] Shen, X.b., Wu, Q., Fu, X.l., 2012. Effects of the duration of expressions on the recognition of microexpressions. Journal of Zhejiang University Science B 13, 221–230.

[43] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations.

[44] Su, Y., Zhang, J., Liu, J., Zhai, G., 2021. Key facial components guided micro-expression recognition based on first & second-order motion, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE. pp. 1–6.

[45] Sun, Z., Hu, Z.p., Zhao, M., Li, S., 2020. Multi-scale active patches fusion based on spatiotemporal lbp-top for micro-expression recognition. Journal of Visual Communication and Image Representation 71, 102862. doi:10.1016/j.jvcir.2020.102862.

[46] Ten Brinke, L., Porter, S., Baker, A., 2012. Darwin the detective: Observable facial muscle contractions reveal emotional high-stakes lies. Evolution and Human Behavior 33, 411–416.

[47] Tharwat, A., 2020. Classification assessment methods. Applied Computing and Informatics .

[48] Verma, M., Reddy, M.S.K., Meedimale, Y.R., Mandal, M., Vipparthi, S.K., 2021. Automer: Spatiotemporal neural architecture search for microexpression recognition. IEEE Transactions on Neural Networks and Learning Systems , 1–13doi:10.1109/TNNLS.2021.3072290.

[49] Wang, S.J., Li, B.J., Liu, Y.J., Yan, W.J., Ou, X., Huang, X., Xu, F., Fu, X., 2018. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. Neurocomputing 312, 251–262.

[50] Wang, S.J., Yan, W.J., Li, X., Zhao, G., Zhou, C.G., Fu, X., Yang, M., Tao, J., 2015. Micro-expression recognition using color spaces. IEEE Transactions on Image Processing 24, 6034–6047. doi:10.1109/TIP.2015.2496314.

[51] Wang, S.J., Yan, W.J., Zhao, G., Fu, X., Zhou, C.G., 2014a. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features, in: European Conference on Computer Vision, Springer. pp. 325–338.

[52] Wang, Y., See, J., Phan, R.C.W., Oh, Y.H., 2014b. LBP with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in: Asian Conference on Computer Vision, Springer. pp. 525–537.

[53] Wang, Y., Wang, L., Wang, H., Li, P., 2019. End-to-end image super-resolution via deep and shallow convolutional networks. IEEE Access 7, 31959–31970.

[54] Xia, Z., Hong, X., Gao, X., Feng, X., Zhao, G., 2020a. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. IEEE Transactions on Multimedia 22, 626–640. doi:10.1109/TMM.2019.2931351.

[55] Xia, Z., Peng, W., Khor, H.Q., Feng, X., Zhao, G., 2020b. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. IEEE Transactions on Image Processing 29, 8590–8605.

[56] Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X., 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. Plos One 9, e86041.

[57] Yan, W.J., Wu, Q., Liu, Y.J., Wang, S.J., Fu, X., 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces, in: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, IEEE. pp. 1–7.

[58] Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer. pp. 818–833.

[59] Zhang, C.L., Luo, J.H., Wei, X.S., Wu, J., 2017. In defense of fully connected layers in visual representation transfer, in: Pacific Rim Conference on Multimedia, Springer. pp. 807–817.

[60] Zhang, M., Chen, Y.H., Fu, X., 2016. Suppression of alpha oscillation during micro-expression recognition, in: Asian Conference on Computer Vision, pp. 544–551.

[61] Zhang, M., Fu, Q., Chen, Y.H., 2014. Emotional context influences micro-expression recognition. Plos One 9, e95018.

[62] Zhang, M., Fu, Q., Chen, Y.H., Fu, X., 2018. Emotional context modulates micro-expression processing as reflected in event-related potentials. PsyCh Journal 7, 13–24.

[63] Zhao, S., Tao, H., Zhang, Y., Xu, T., Zhang, K., Hao, Z., Chen, E., 2021. A two-stage 3d cnn based learning method for spontaneous micro-expression recognition. Neurocomputing 448, 276–289.

[64] Zhou, L., Mao, Q., Huang, X., Zhang, F., Zhang, Z., 2022. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. Pattern Recognition 122, 108275.

[65] Zong, Y., Huang, X., Zheng, W., Cui, Z., Zhao, G., 2018. Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. IEEE Transactions on Multimedia 20, 3160–3172.