Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Micro-expression recognition using a multi-scale feature extraction network with attention mechanisms

Yan Wang
Jiangsu University of Science and Technology
Qingyun Zhang
Jiangsu University of Science and Technology
Xin Shu
shuxin@just.edu.cn
Jiangsu University of Science and Technology

Research Article

Keywords: Micro-expression recognition, Attention mechanism, Multi-scale features, Optical flow, Deep learning

Posted Date: September 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3089932/v1

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Signal, Image and Video Processing on May 8th, 2024. See the published version at https://doi.org/10.1007/s11760-024-03221-1.

Abstract

Micro-expressions are instantaneous flashes of facial expressions that reveal a person's true feelings and emotions. Micro-expression recognition (MER) is challenging due to its low motion intensity, short duration, and the limited number of publicly available samples. Although the present MER methods have achieved great progress, they face the problems of a large number of training parameters and insufficient feature extraction ability. In this paper, we propose a lightweight network MFE-Net with Res-blocks to extract multi-scale features for MER. To extract more valuable features, we incorporate Squeeze-and-Excitation (SE) attention and multi-headed self-attention (MHSA) mechanisms in our MFE-Net. The proposed network is used for learning features from three optical flow features (i.e. optical strain, horizontal and vertical optical flow images) which are calculated from the onset and apex frames. We employ the LOSO cross-validation strategy to conduct experiments on CASME II and the composite dataset selected by MEGC2019, respectively. The extensive experimental results demonstrate the viability and effectiveness of our method.

1 Introduction

Facial expression is a form of non-verbal communication, which expresses people's mental state and emotions and plays a crucial role in our daily communication. Facial expressions can usually be divided into six categories: happiness, sadness, fear, anger, disgust, and surprise. Researchers have achieved excellent recognition performance on macro expressions. Numerous expression recognition [1] systems are developed, which can reach more than 95% classification accuracy [2, 3]. Compared to expression studies, micro-expression has a shorter history. It was first proposed by Haggard et al. in 1966 [4], who argued that micro-expression is related to ego defense mechanisms and expresses repressed emotions. Ekman and Friesen also discovered micro-expression in 1969 [5].

Micro-expression is a rapid, unconscious, spontaneous facial movement that occurs when a person is experiencing strong emotions. Micro-expression, neither faked nor suppressed, is produced when people try to hide their inner emotions [4]. Micro-expression is characterized by a short duration, typically lasting 1/25 ~ 1/3s [6]. Another characteristic is low-intensity movement so that it does not occur simultaneously in the upper and lower part of the face.

Micro-expression is commonly applied in clinical diagnosis, emotional intelligence, judicial investigation, etc. Although the Micro-Expression Training Tool (METT) [7] has been developed to train professionals, the results of human recognition are still not ideal, with only 47% reported in the literature [8]. Hence MER needs to be realized automatically by a computer, which can handle large-scale MER tasks at an inexpensive cost whenever an efficient and stable model is trained [8].

MER mainly involves the establishment of micro-expression datasets, preprocessing techniques, and micro-expression recognition algorithms. Until now, only a few micro-expression datasets are available and according to the method of elicitation, they are classified into two categories: posed and spontaneous. This paper is conducted on the spontaneous micro-expression datasets entirely. The main publicly available spontaneous micro-expression datasets are SMIC [9], CASME [10], CASME II [11], CAS(ME)^2 [12] and SAMM [13].

Significant progress has been made in MER based on the release of these datasets mentioned above. However, the current works still suffer from the excessive amount of model parameters and insufficient extraction of micro-expression features. To address the above deficiencies, we propose a novel network named MFE-Net, which reduces the number of model parameters significantly, obtains more critical and essential features, and suppresses useless information as well. The results on public benchmarks demonstrate that MFE-Net is viable for MER. The contributions of this paper are summarized as follows:

(1) We propose a novel MER network with three branches that have different convolution kernels to extract multi-scale micro-expression features.

(2) The channel attention SE and MHSA are embedded in Res-blocks to focus on the most informative channels and extract valuable features.

(3) Extensive experiments are conducted on multiple micro-expression datasets, and the results show that the proposed method outperforms or is comparable to the state-of-the-art methods on public and composite datasets.

The remainder of this paper is organized as follows. Section 2 introduces the related works. Section 3 details the proposed and theoretical derivation of our approach. A detailed description of the experiments is given in Section 4. Finally, Section 5 draws a brief conclusion of our approach.

2 Related Work

2.1 Handcrafted methods

Early research on MER mostly focused on extracting features manually, which can be mainly divided into appearance-based methods and optical flow-based methods. In terms of the former, Zhao et al. [14] used the LBP-TOP operator to extract features from the video XY plane, XT plane, and YT plane, respectively. Inspired by [14], many variations of LBP-TOP have emerged [15]. Wang et al. [16] proposed the LBP-SIP descriptor, which considers a special case of LBP-TOP to calculate the relationship between each four pixels in three planes and the central pixel. LBP-SIP reduces the dimensions of the features, decreases the redundant information of LBP-TOP, and improves the efficiency of feature extraction. Huang et al. [17] proposed STLBP-IP, which uses the integral projection technique to extract facial shape information based on LBP. STLBP-IP extracts relatively important micro-expression features and discards unimportant features. Other improvement methods for LBP-TOP include Hierarchical STLBP-IP [18], DiSTLBP-RIP [19], etc.

In addition to the appearance-based feature extraction methods, there are many approaches based on optical flow information. Xu et al. [20] proposed facial dynamics map (FDM) to suppress the anomalous optical loss caused by noise or illumination changes, which has high time complexity. FDM divides the optical flow sequence into spatiotemporal segmentation blocks and then calculates the main optical flow direction for each spatiotemporal segmentation block. Liu et al. [21] proposed the main directional mean optical flow feature (MDMO) by extracting the main direction in the video sequence and calculating the mean optical flow feature in the block of the face part. The computational efficiency is ensured by its fewer dimensions. The Sparse MDMO method was further proposed by Liu et al. [22] to preserve the popular structure information in the feature space using the distance measure.

Handcrafted feature extraction, which chiefly relies on manually designed rules, requires specialized knowledge and a complex parameter adjustment process. The obtained features cannot explain the physical meaning of each specific dimension. Meanwhile, the generalization capability and robustness of the engineered methods are limited.

2.2 Deep learning methods

Recently, deep learning methods receive unprecedented attention and have been considered an efficient way to learn feature representation. Patel et al. [23] first adopted convolution neural networks for MER. Since the small amount of micro-expression sample data brings a challenge to training the network model adequately, Patel et al. used transfer learning to migrate features from macro-expression to the micro-expression task. The results obtained from their work are not better than traditional methods due to the possibility of model overfitting.

Then, many researchers used the temporal and spatial information of micro-expression to learn feature representation. Kim et al. [24] proposed combining CNN with Long Short-Term Memory (LSTM) network to extract spatiotemporal information of micro-expression. Xia et al. [25] introduced the spatiotemporal recurrent convolutional networks (STRCN) model, which uses recurrent convolution networks to encode micro-expression video sequences. To handle the temporal data, temporal deformations are modeled in facial appearance and geometric views that are called STRCN-A and STRCN-G respectively.

Besides the above approaches, some researchers tried to utilize the apex frame and its related information for MER. Li et al. [26] used the Eulerian motion magnification [27] to amplify the apex frames and applied the VGG-Face model to fine-tune the weights of the network with small-scale data. Liong et al. [28] and Gan et al. [29] demonstrated that the apex frame provides sufficient information to recognize micro-expression. Gan et al. proposed OFF-ApexNet which extracts optical flow information between the onset and apex frames of each video. The horizontal and vertical components of the optical flow are fed into a two-stream network to learn the micro-expression features. Liu et al. [30] explored a deep learning method with antagonistic training and expression magnification, which achieved the best results in MEGC2019. Moreover, Liong et al. [31] presented STSTNet which learns features from the optical strain, the horizontal optical flow images, and the vertical optical flow images. Quang et al. [32] introduced a simple but effective CapsuleNet that exploits the knowledge from apex frames only without heavy and complicated computations when using all the frames in micro-expression sequences. Inspired by Inception, Zhou et al. [33] developed the Dual-Inception model that overcomes the challenge of the cross-dataset micro-expression recognition by feeding the optical flow features extracted from the onset and mid-position frames. Khor et al. [34] proposed an enriched long-term recurrent convolutional network (ELRCN) that encodes each frame into a feature vector through CNN and predicts the micro-expression bypassing the feature vector through the LSTM module.

3 Proposed Method

The overall process of our proposed MER method, which consists of three steps, is shown in Fig. 1. In Step ☒, the onset and apex frames are first identified from the image sequences. Then the optical flow information, including the horizontal optical flow component, vertical optical flow component, and optical strain, is extracted from the onset and apex frames in Step ▯. For Step ▯, the optical flow information is used as the input of the MFE-Net which is a robust model for successfully identifying the emotional labels of micro-expression.

Figure 2 illustrates the complete architecture of our proposed MFE-Net. Followed by two convolution layers and a max-pooling layer, the inputs are fed into three branches for extracting multi-scale features. Each of the branches in the network contains two SE-RES blocks and one MHSA-RES block. The three branches differ in the size of the convolution kernels, which are 3×3, 5×5, and 7×7 respectively. The receptive field is affected by the size of the convolution kernel. In general, since larger convolution kernels have larger receptive fields, more image information can be extracted. That means we can use larger convolution kernels to obtain large-scale features and smaller convolution kernels to get local features. Finally, features obtained from each branch are cascaded to form the multi-scale micro-expression features.

3.1 Preprocessing

Initially, the input color images are cropped to an appropriate size of 112×112 and normalized by inter-cubic interpolation, and then they are converted to grayscale before optical flow feature extraction. The optical flow represents the motion information of videos, which can effectively reduce the domain difference between different micro-expression datasets. Moreover, Khor et al. [34] demonstrated that optical flow features have a significant impact on improving the accuracy of micro-expression recognition. In this paper, we use the TV-L1 [35] to extract optical flow information which is obtained from the onset and apex frames of each micro-expression video. It was verified in literature [33] that using the middle frame instead of the apex frame can also achieve favorable recognition performance in datasets without the apex frame. Apex frames are labeled clearly in datasets CASME II and SAMM, but not in SMIC, so we take the middle frames instead of apex frames for optical flow feature extraction in SMIC.

We use u and v to denote the horizontal and vertical optical flow components, respectively. The optical strain is approximated by the deformation strength, which is defined as [31]:

$$\epsilon = \frac{1}{2}\left[\nabla \backslash\mathrm{varvec}u + (\nabla \backslash\mathrm{varvec}u)^{T}\right] \# (1)$$

where, $\backslash\mathrm{varvec}u = [u, v]^{T}$ is the displacement vector. Thus, the optical strain is formulated as:

$$\epsilon = \begin{bmatrix} \epsilon_{xx} = \frac{\partial u}{\partial x} & \epsilon_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \\ \epsilon_{yx} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) & \epsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix} \# (2)$$

where the diagonal strain components $(\epsilon_{xx}\text{▯}\epsilon_{yy})$ are normal strain components, $(\epsilon_{xy}\text{▯}\epsilon_{yx})$ are shear strain components, and (x, y) represents the position of the pixel.

Then, the optical strain of each pixel is calculated by taking the sum of the squares of the normal strain and shear strain components and is as follows:

$$\left|\epsilon_{x\text{▯}y}\right| = \sqrt{\frac{\partial u^{2}}{\partial x} + \frac{\partial v^{2}}{\partial y} + \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^{2}} \# (3)$$

In summary, we can obtain the horizontal optical flow u, the vertical optical flow v, and the optical strain ϵ as inputs from each micro-expression sequence, respectively.

3.2 SE-Res block

2D convolution only extracts features from a local neighborhood, and it is difficult to obtain relationship info among channels. To extract cross-channel features, channel attention SE [36] is employed to focus on the relationship among channels and automatically learn the importance of different channel features. The SE module performs the Squeeze operation on the feature map obtained by convolution to get the global features at the channel level. Then, Excitation is applied to the global features for learning the relationship between channels. Essentially, the SE module performs the attention operation among channels. This channel attention mechanism allows the model to pay more attention to the most informative channel features and suppress the unimportant ones.

Usually, SE can be inserted after the nonlinear layer followed by the convolution. Ref. [36] proposed the SE-Res block which is the combination of SE and the Res-block. In this paper, two SE-Res blocks are applied in each branch of our MFE-Net.

The detailed structure of the original Res-block and the SE-Res block are shown in Fig. 4(a) and Fig. 4(b). The Conv in the original Res-block uses a convolution kernel of size 3×3. For extracting multi-scale features, a different kernel is adopted in each of the three branches in our MFE-Net. In the experiments, the convolution kernel sizes are set as 3×3, 5×5, and 7×7, respectively.

3.3 MHSA-Res block

Transformer is initially proposed for natural language processing (NLP) [37] and attracts more and more concern from researchers in computer vision fields. Self-attention as a vital part of the Transformer is extremely helpful for computer vision tasks, so in this paper, we employ h = 4 parallel self-attention layers, or heads. Figure 4(c) illustrates the structure of the MHSA-Res block, in which the dashed box shows the specific operations of MHSA. From Fig. 4 (c), it is clear that the input tensor is transformed into three different representations, namely the query q, the key k, and the value v, through three linear transformation matrices Wq, Wk, and Wv. We can calculate the output of the self-attention module [38] as follows:

$$Attention\square q,k,v,r\square = Softmax\left(qk^T+qr^T\right)*v\#\left(4\right)$$

$$r_{x,y}=PE\left(x,featureMapSize\right)+PE\left(y,featureMapSize\right)\#\left(5\right)$$

$$PE\left(2i,d\right)=\sin\left(\frac{1}{10000^{\frac{2i}{d}}}\right)\#\left(6\right)$$

$$PE\left(2i+1,d\right)=\cos\left(\frac{1}{10000^{\frac{2i}{d}}}\right)\#\left(7\right)$$

where r denotes the position encoding (PE). x and y denote the pixel positions. i is the position and d means the dimension of the feature map size. The position encoding of a two-dimensional image is obtained by summing two one-dimensional sinusoidal position encoding. We compute the element-wise summation of the query-key matrix product and query-positional code matrix product. Then a softmax function is applied to obtain the weights on the values and we calculate the weights-value matrix product finally.

4 Experiment

4.1 Datasets

In the experiments, the performance of the proposed method is evaluated on three commonly used datasets, namely CASME II, SAMM, and SMIC. Table 1 shows the details of the three datasets.

Table 1
Introduction of the three commonly used datasets

Datasets	CASME II	SAMM	SMIC-HS
Size	280×340	400×400	190×230
Frame rate (fps)	200	200	100
Samples	255	159	164
AU	√	√	×
Apex	√	√	×
Classes	7	7	3

In CASME II, the camera records at a rate of 200 fps with a resolution of 640×480 and a facial resolution of 280×340. The total number of samples in CASME II is 255 and the emotion labels, apex frame labels, and AU labels are provided. In SAMM, which provides emotion labels, apex frame labels, and AU labels, the total number of samples is 159. The recording rate of the camera is 200 fps and the facial resolution is 400×400. As for SMIC, the camera records at 100 fps with a resolution of 640×480, and only emotion labels are provided. The total number of samples in SMIC is 164.

Table 2 details the cross-dataset from The Second Facial Micro-Expressions Grand Challenge (MEGC2019), which recombines CASME II, SAMM, and SMIC into 442 samples, including 68 subjects (16 from SMIC, 24 from CASME II, and 28 from SAMM).

Table 2
Cross-dataset and labels of each dataset

Class Dataset	Negative	Positive	Surprise	Total
CASME II	88	32	25	145
SMIC	70	51	43	164
SAMM	92	26	15	132
3DB-combined	250	109	83	442

4.2 Ablative analysis

To ensure the effectiveness of our model, we performed a series of ablation experiments. Referring to [39], all the experiments in ablation analysis are performed on CASME II with four classes. The aim is to find the optimal configuration of network parameters and structure on an ethnically homogeneous dataset (CASME II).

Table 3
Methods of CASME II merged into four classes

Four classes	Original class
Positive (32)	happiness (32)
Negative (69)	disgust (63) + sadness (4) + fear (2)
Surprise (28)	surprise (28)
Others (126)	others (99) + repression (27)

In the experiments, we evaluated our proposed method according to the leave-one-subject-out (LOSO) protocol. The results were evaluated in terms of accuracy and F1-score, which are calculated as follows:

$$Accuracy = \frac{T}{N} \times 100\% \# (8)$$

$$F1\text{-}\ score = \frac{2PR}{P+R} \# (9)$$

where $T$is the total number of correct predictions, N denotes the total number of test samples, P means precision, and R represents the recall rate, respectively.

The SE-Res and MHSA-Res blocks are used in our network. To verify their effectiveness, we conducted extensive ablation experiments and Table 4 illustrates the comparative results. As we can see, the Accuracy and F1-score of the single-branch model only achieve 71.97% and 69.24%, respectively. The Accuracy of the single-branch model with MHSA and SE improves by 3.73–75.70% and the F1-score reaches 73.63% with an improvement of 4.39%. When we use a single-scale convolution kernel (3×3) in the three-branch network, the Accuracy and F1-score are improved by 2.19% and 2.66% compared to the single-branch model with MHSA and SE, respectively. For the three-branch network using a single-scale convolution kernel(7×7), the Accuracy is improved by only 0.04% and the F1-score declines by 6.15% on the contrary. However, our proposed MFE-Net achieves the best performance, and the Accuracy and F1-score arrive at 81.18% and 80.39%, respectively. According to the results of ablation experiments, it is not difficult to conclude that multi-scale features are more conducive to micro-expression recognition.

Table 4
Ablation experiments (with four classes on CASME II)

Methods	Kernel size	Accuracy	F1-score
single-branch	3×3	71.97%	69.24%
single-branch + MHSA	3×3	72.83%	73.05%
single-branch + SE	3×3	75.64%	73.04%
single-branch + MHSA + SE	3×3	75.70%	73.63%
three-branch + MHSA + SE + same kernel	3×3	77.89%	76.29%
three-branch + MHSA + SE + same kernel	5×5	78.52%	77.68%
three-branch + MHSA + SE + same kernel	7×7	75.74%	67.48%
MFE-Net (Ours)	3×3, 5×5, 7×7	81.18%	80.39%

4.3 Experiments on the CASME II dataset

We compared our MFE-Net with other existing methods on the CASME II dataset (with four classes), and the results are shown in Table 5. It can be seen that our method exceeds STRCN-G which has the best performance among these existing methods. The Accuracy of our MFE-Net reaches 81.18%, which is 0.88% higher than that of STRCN-G, and F1-score achieves 80.39% improved by 5.69%.

Table 5
The results of the comparison experiments on CASME II with four classes

Methods	Accuracy	F1-score	Published
MDMO [21]	51.00%	41.80%	2016
FDM [20]	41.70%	29.70%	2017
Im-based CNN [40]	44.40%	42.80%	2017
Bi-WOOF [28]	58.90%	61.00%	2018
Hier.STLBP-IP [18]	63.80%	61.10%	2018
STRCN-A [25]	56.00%	54.20%	2020
STRCN-G [25]	80.30%	74.70%	2020
Graph-tcn [41]	73.60%	-	2020
Ours	**81.18%**	**80.39%**	

To further verify the generalization ability of our proposed method, we utilized the best configuration of the network parameters in ablation analysis to directly carry out the 5-classification experiment on the CASME II dataset. The detailed experimental results are shown in Table 6. The results of our MFE-Net are second only to TSCNN, which shows that our method is very competitive. Notably, our F1-score is very close to that of TSCNN.

What's more, MFE-Net's Accuracy exceeds AU-GCN by 4.66–78.93%, and F1-score reaches 80.53% which is 7.56% higher than that of DSSN.

Table 6
Comparison experiments on CASME II with five classes

Methods	Accuracy	F1-score	Published
CNN + LSTM [24]	60.98%	-	2016
Bi-WOOF + Phase [42]	62.55%	65.00%	2017
MagGA [26]	63.30%	-	2018
Hier.STLBP-IP [18]	63.97%	61.25%	2018
Sparse MDMO [22]	66.95%	69.11%	2018
HIGO + Mag [43]	67.21%	-	2018
DiSTLBP-RIP [19]	64.78%	-	2019
ME-Booster [44]	70.85%	-	2019
SSSN [45]	71.19%	71.51%	2019
DSSN [45]	70.78%	72.97%	2019
TSCNN [46]	**80.97%**	**80.70%**	2019
Graph-tcn [41]	73.98%	72.46%	2020
AU-GCN [39]	74.27%	70.47%	2021
Ours	78.93%	80.53%	

4.4 Composite Database Evaluation (CDE)

To further prove the validity of our method, we performed a more robust evaluation with the LOSO protocol on the cross-dataset used in MEGC2019. UF1 and UAR employed in MEGC2019 are used as evaluation metrics. UF1 is determined by averaging the F1-score of per-class c (in class C) and we average all accuracy by the number of classes to obtain the final UAR score as below:

$$F1\text{-}score = \frac{2TP_c}{2TP_c+FP_c+FN_c} \# (10)$$

$$UF1 = \frac{1}{C}\sum_C F1_c \# (11)$$

$$UAR = \frac{1}{C}\sum_C \frac{TP_c}{n_c} \# (12)$$

where TP (True Positive) means the number of samples that will be predicted correctly as positive samples. FP (False Positive) refers to the number of negative samples predicted as positive samples. FN (False Negative) is the number of positive samples predicted as negative samples. n_c is the number of samples in class c.

Table 7
Experiments on CDE with three classes

Methods	Full		CASME II		SAMM		SMIC		Published
	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1	
LBP-TOP [14]	0.5787	0.5882	0.7429	0.7026	0.4102	0.3954	0.5280	0.2000	2007
Bi-WOOF [28]	0.6227	0.6296	0.8026	0.7805	0.5139	0.5211	0.5829	0.5727	2018
OFF-ApexNet [29]	0.7096	0.7196	0.8681	0.8764	0.5392	0.5409	0.6695	0.6817	2019
CapsuleNet [32]	0.6506	0.6520	0.7018	0.7068	0.5989	0.6209	0.5877	0.5820	2019
Dual-Inception [33]	0.7278	0.7322	0.8560	0.8621	0.5663	0.5868	0.6726	0.6645	2019
STSTNet [31]	0.7605	0.7353	0.8686	0.8382	0.6810	0.6588	0.7013	0.6801	2019
EMR [30]	0.7824	0.7885	0.8209	0.8293	0.7152	**0.7754**	0.7530	0.7461	2019
FeatRef [47]	0.7832	0.7838	**0.8873**	**0.8915**	0.7155	0.7372	0.7083	0.7011	2021
GEME (Mutil-task) [48]	0.7303	0.7221	0.8790	0.8831	0.5455	0.5843	0.6387	0.6038	2021
AU-GCN [39]	0.7933	0.7914	0.8710	0.8798	0.7890	0.7751	0.7215	0.7192	2021
Ours	**0.8550**	**0.8549**	0.8858	0.8764	**0.7956**	0.7575	**0.7898**	**0.7966**	

Ten existing methods are compared with our proposed MFE-Net, among them LBP-TOP and Bi-WOOF are handcrafted feature extraction methods and others are deep learning approaches. Table 7 lists the scores of UF1 and UAR on the full cross-database and the separate parts including SMIC, CASME II, and SAMM. As shown in Table 7, our model obtains the scores of UF1 and UAR of 0.8549 and 0.8550, which exceed the scores of AU-GCN by 0.0617 and 0.0635, respectively. The performance of our model is the best on full cross-dataset and SMIC and the results on CASME II rank second only to FeatRef. For SAMM, our method exceeds AU-GCN by 0.066 to reach the highest one on UAR, while UF1 is second only to EMR. Comprehensive experimental results show the effectiveness of the proposed method which outperforms several powerful CNN models in MER.

4.5 Parameters of the model

The above comparative experiments demonstrate the effectiveness of our proposed method. In addition, we make a comparison with some current mainstream methods in terms of model parameters. As can be seen from Table 8, the number of parameters in our model is significantly reduced compared to the current mainstream models. Even though the number of parameters of our model is slightly higher than STSTNet, the UAR and UF1 scores achieved by our model on CDE are better than those of STSTNet.

Table 8
Comparison of model parameters

Model	Params
Off-ApexNet [29]	2.66M
STSTNet [31]	162051
Dual-Inception [33]	6.45M
MACNN [49]	70.57M
Micro-Attention [50]	53.38M
Ours	308998

5 Conclusion

In this paper, we propose a novel network MFE-Net, which achieves multi-scale feature extraction by three branches and makes feature extraction more adequate. The network also incorporates attention mechanisms to enhance the extraction of valid information and suppress the useless. What’s more, the number of model parameters is greatly reduced. The effectiveness of our method is verified on the publicly available datasets. In future work, we will explore more efficient architectures for MFE-Net and investigate effective ways to enrich the micro-expression samples.

Declarations

Author Contributions

All authors contributed to the study design. Yan Wang: Data Curation, Visualization, Validation, Review. Qingyu Zhang: Writing - review & editing, Software, Methodology. Xin Shu: Methodology, Supervision, Writing - review & editing. All authors read and approved the final manuscript.

Ethical Approval

Not applicable.

Competing Interest

All authors disclosed no relevant relationships and have no conflict of interest.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62276118.

Data Availability

Data are available from the authors upon reasonable request..

References

1. Kumar, R.J.R., Sundaram, M. & Arumugam, N. Facial emotion recognition using subband selective multilevel stationary wavelet gradient transform and fuzzy support vector machine. Vis Comput 37, 2315–2329 (2021). https://doi.org/10.1007/s00371-020-01988-1.
2. Z. Wang, Q. Ruan, G. An, Facial expression recognition using sparse local Fisher discriminant analysis, Neurocomputing. 174 (2016) 756–766. https://doi.org/https://doi.org/10.1016/j.neucom.2015.09.083.
3. Agarwal, S., Santra, B. & Mukherjee, D.P. Anubhav: recognizing emotions through facial expression. Vis Comput 34, 177–191 (2018). https://doi.org/10.1007/s00371-016-1323-z.
4. K.S. Haggard Ernest A. and Isaacs, Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy, in: Methods of Research in Psychotherapy, Springer US, Boston, MA, 1966: pp. 154–165. https://doi.org/10.1007/978-1-4684-6045-2_14.
5. P. Ekman, W. v Friesen, Nonverbal Leakage and Clues to Deception, Psychiatry. 32 (1969) 88–106. https://doi.org/10.1080/00332747.1969.11023575.
6. W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, X. Fu, How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions, Journal of Nonverbal Behavior. 37 (2013) 217–230. https://doi.org/10.1007/s10919-013-0159-8.
7. P. Ekman, Microexpression Training Tool (METT). San Francisco, CA, USA: University California, 2002.
8. Frank M, Herbasz M, Sinuk K, I see how you feel: Training laypeople and professionals to recognize fleeting emotions, in: The Annual Meeting of the International Communication Association, New York City, 2009: pp. 1–35.
9. X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A Spontaneous Micro-expression Database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013: pp. 1–6. https://doi.org/10.1109/FG.2013.6553717.
10. W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, X. Fu, 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), 2013: pp. 1–7. https://doi.org/10.1109/FG.2013.6553799.
11. W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, CASME II: an improved spontaneous micro-expression database and the baseline evaluation, PLoS One. 9 (2014) e86041. https://doi.org/10.1371/journal.pone.0086041.
12. F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, X. Fu, CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition, IEEE Transactions on Affective Computing. 9 (2018) 424–436. https://doi.org/10.1109/TAFFC.2017.2654440.
13. A.K. Davison, C. Lansley, N. Costen, K. Tan, M.H. Yap, SAMM: A Spontaneous Micro-Facial Movement Dataset, IEEE Transactions on Affective Computing. 9 (2018) 116–129. https://doi.org/10.1109/TAFFC.2016.2573832.
14. G. Zhao, M. Pietikainen, Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence. 29 (2007) 915–928.

https://doi.org/10.1109/TPAMI.2007.1110.

15. X. Shu, Z. Song, J. Shi, S. Huang, X.-J. Wu, Multiple channels local binary pattern for color texture representation and classification, Signal Processing: Image Communication. 98 (2021) 116392. https://doi.org/https://doi.org/10.1016/j.image.2021.116392.
16. J. and P.R.C.-W. and O.Y.-H. Wang Yandan and See, LBP with Six Intersection Points: Reducing Redundant Information in LBP-TOP for Micro-expression Recognition, in: I. and S.H. and Y.M.-H. Cremers Daniel and Reid (Ed.), Computer Vision – ACCV 2014, Springer International Publishing, Cham, 2015: pp. 525–537.
17. X. Huang, S.-J. Wang, G. Zhao, M. Piteikäinen, Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern with Integral Projection, in: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015: pp. 1–9. https://doi.org/10.1109/ICCVW.2015.10.
18. Y. Zong, X. Huang, W. Zheng, Z. Cui, G. Zhao, Learning From Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition, IEEE Transactions on Multimedia. 20 (2018) 3160–3172. https://doi.org/10.1109/TMM.2018.2820321.
19. X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, M. Pietikäinen, Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection for Spontaneous Facial Micro-Expression Recognition, IEEE Transactions on Affective Computing. 10 (2019) 32–47. https://doi.org/10.1109/TAFFC.2017.2713359.
20. F. Xu, J. Zhang, J.Z. Wang, Microexpression Identification and Categorization Using a Facial Dynamics Map, IEEE Transactions on Affective Computing. 8 (2017) 254–267. https://doi.org/10.1109/TAFFC.2016.2518162.
21. Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition, IEEE Transactions on Affective Computing. 7 (2016) 299–310. https://doi.org/10.1109/TAFFC.2015.2485205.
22. Y.-J. Liu, B.-J. Li, Y.-K. Lai, Sparse MDMO: Learning a Discriminative Feature for Micro-Expression Recognition, IEEE Transactions on Affective Computing. 12 (2021) 254–261. https://doi.org/10.1109/TAFFC.2018.2854166.
23. D. Patel, X. Hong, G. Zhao, Selective deep features for micro-expression recognition, in: 2016 23rd International Conference on Pattern Recognition (ICPR), 2016: pp. 2258–2263. https://doi.org/10.1109/ICPR.2016.7899972.
24. D.H. Kim, W.J. Baddar, Y.M. Ro, Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations, in: Proceedings of the 24th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2016: pp. 382–386. https://doi.org/10.1145/2964284.2967247.
25. Z. Xia, X. Hong, X. Gao, X. Feng, G. Zhao, Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-Expressions, IEEE Transactions on Multimedia. 22 (2020) 626–640. https://doi.org/10.1109/TMM.2019.2931351.
26. Y. Li, X. Huang, G. Zhao, Can Micro-Expression be Recognized Based on Single Apex Frame?, in: 2018 25th IEEE International Conference on Image Processing (ICIP), 2018: pp. 3094–3098. https://doi.org/10.1109/ICIP.2018.8451376.
27. H. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian Video Magnification for Revealing Subtle Changes in the World, in: 2012.

28. S.-T. Liong, J. See, K. Wong, R.C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Processing: Image Communication. 62 (2018) 82–92. https://doi.org/https://doi.org/10.1016/j.image.2017.11.006.

29. Y.S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, OFF-ApexNet on micro-expression recognition system, Signal Processing: Image Communication. 74 (2019) 129–139. https://doi.org/https://doi.org/10.1016/j.image.2019.02.005.

30. Y. Liu, H. Du, L. Zheng, T. Gedeon, A Neural Micro-Expression Recognizer, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019: pp. 1–4. https://doi.org/10.1109/FG.2019.8756583.

31. S.-T. Liong, Y.S. Gan, J. See, H.-Q. Khor, Y.-C. Huang, Shallow Triple Stream Three-dimensional CNN (STSTNet) for Micro-expression Recognition, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019: pp. 1–5. https://doi.org/10.1109/FG.2019.8756567.

32. N. van Quang, J. Chun, T. Tokuyama, CapsuleNet for Micro-Expression Recognition, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019: pp. 1–7. https://doi.org/10.1109/FG.2019.8756544.

33. L. Zhou, Q. Mao, L. Xue, Dual-Inception Network for Cross-Database Micro-Expression Recognition, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019: pp. 1–5. https://doi.org/10.1109/FG.2019.8756579.

34. H.-Q. Khor, J. See, R.C.W. Phan, W. Lin, Enriched Long-Term Recurrent Convolutional Network for Facial Micro-Expression Recognition, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018: pp. 667–674. https://doi.org/10.1109/FG.2018.00105.

35. T. and B.H. Zach C. and Pock, A Duality Based Approach for Realtime TV-L1 Optical Flow, in: C. and J.B. Hamprecht Fred A. and Schnörr (Ed.), Pattern Recognition, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007: pp. 214–223.

36. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence. 42 (2020) 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372.

37. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Lukasz Kaiser, I. Polosukhin, Attention is All You Need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017: pp. 6000–6010.

38. F. and Z.A. Zhang Jiahao and Liu, Off-TANet: A Lightweight Neural Micro-expression Recognizer with Optical Flow Features and Integrated Attention Mechanism, in: T. and G.G. and L.F. Pham Duc Nghia and Theeramunkong (Ed.), PRICAI 2021: Trends in Artificial Intelligence, Springer International Publishing, Cham, 2021: pp. 266–279.

39. L. Lei, T. Chen, S. Li, J. Li, Micro-expression Recognition Based on Facial Graph Representation Learning and Facial Action Unit Fusion, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021: pp. 1571–1580. https://doi.org/10.1109/CVPRW53098.2021.00173.

40. M.A. Takalkar, M. Xu, Image Based Facial Micro-Expression Recognition Using Deep Learning on Small Datasets, in: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2017: pp. 1–7. https://doi.org/10.1109/DICTA.2017.8227443.

41. L. Lei, J. Li, T. Chen, S. Li, A Novel Graph-TCN with a Graph Structured Representation for Micro-Expression Recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2020: pp. 2237–2245. https://doi.org/10.1145/3394171.3413714.
42. S.-T. Liong, K. Wong, Micro-expression recognition using apex frame with phase information, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017: pp. 534–537. https://doi.org/10.1109/APSIPA.2017.8282090.
43. X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, M. Pietikäinen, Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods, IEEE Transactions on Affective Computing. 9 (2018) 563–577. https://doi.org/10.1109/TAFFC.2017.2667642.
44. W. Peng, X. Hong, Y. Xu, G. Zhao, A Boost in Revealing Subtle Facial Expressions: A Consolidated Eulerian Framework, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019: pp. 1–5. https://doi.org/10.1109/FG.2019.8756541.
45. H.-Q. Khor, J. See, S.-T. Liong, R.C.W. Phan, W. Lin, Dual-stream Shallow Networks for Facial Micro-expression Recognition, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019: pp. 36–40. https://doi.org/10.1109/ICIP.2019.8802965.
46. B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, L. Zhao, Recognizing Spontaneous Micro-Expression Using a Three-Stream Convolutional Neural Network, IEEE Access. 7 (2019) 184537–184551. https://doi.org/10.1109/ACCESS.2019.2960629.
47. L. Zhou, Q. Mao, X. Huang, F. Zhang, Z. Zhang, Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition, Pattern Recognition. 122 (2022) 108275. https://doi.org/https://doi.org/10.1016/j.patcog.2021.108275.
48. X. Nie, M.A. Takalkar, M. Duan, H. Zhang, M. Xu, GEME: Dual-stream multi-task GEnder-based micro-expression recognition, Neurocomputing. 427 (2021) 13–28. https://doi.org/https://doi.org/10.1016/j.neucom.2020.10.082.
49. Z. Lai, R. Chen, J. Jia, Y. Qian, Real-time micro-expression recognition based on ResNet and atrous convolutions, Journal of Ambient Intelligence and Humanized Computing. (2020). https://doi.org/10.1007/s12652-020-01779-5.
50. C. Wang, M. Peng, T. Bi, T. Chen, Micro-attention for micro-expression recognition, Neurocomputing. 410 (2020) 354–362. https://doi.org/https://doi.org/10.1016/j.neucom.2020.06.005.

Figures

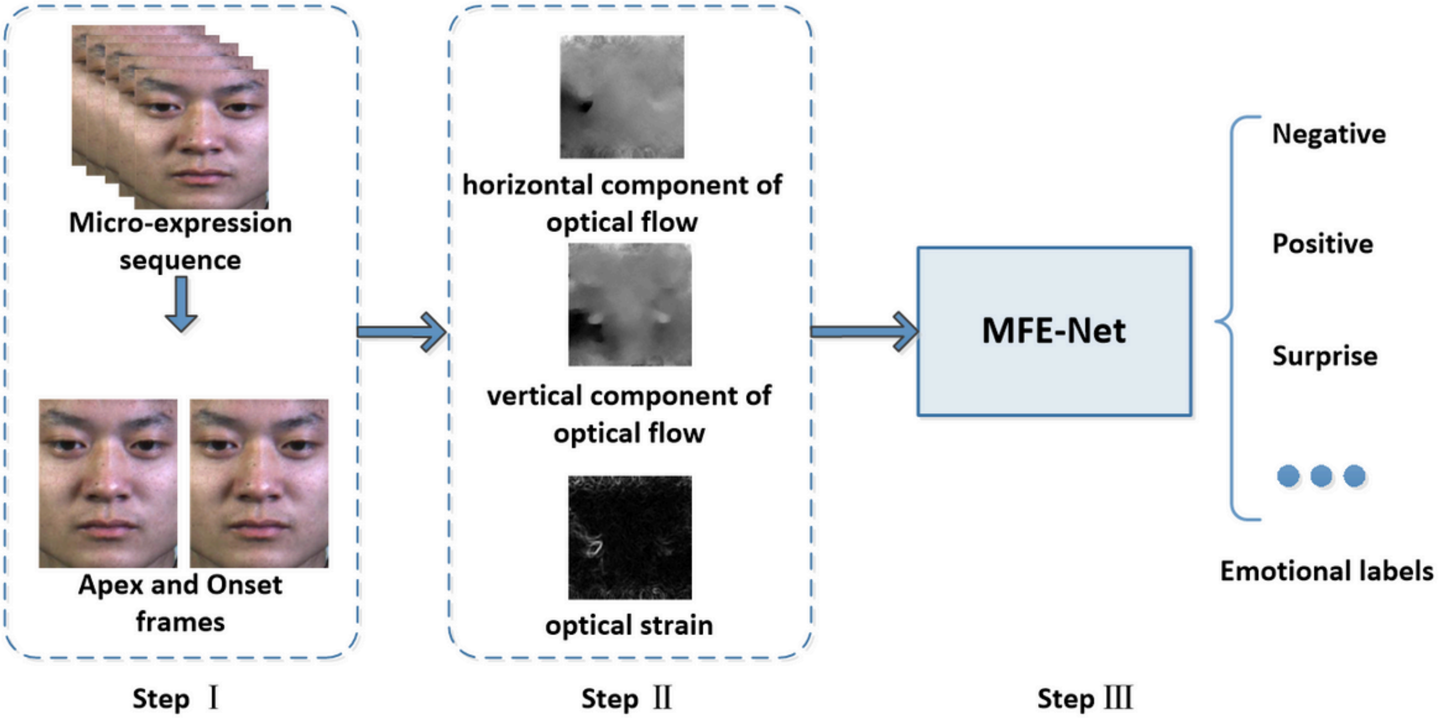

Figure 1

The overall flow chart of our proposed MER method. The micro-expression samples in Step are from CASME dataset. The images in Step are the visualized optical flow information of the chosen samples. The robust model MFE-Net is trained in Step III.

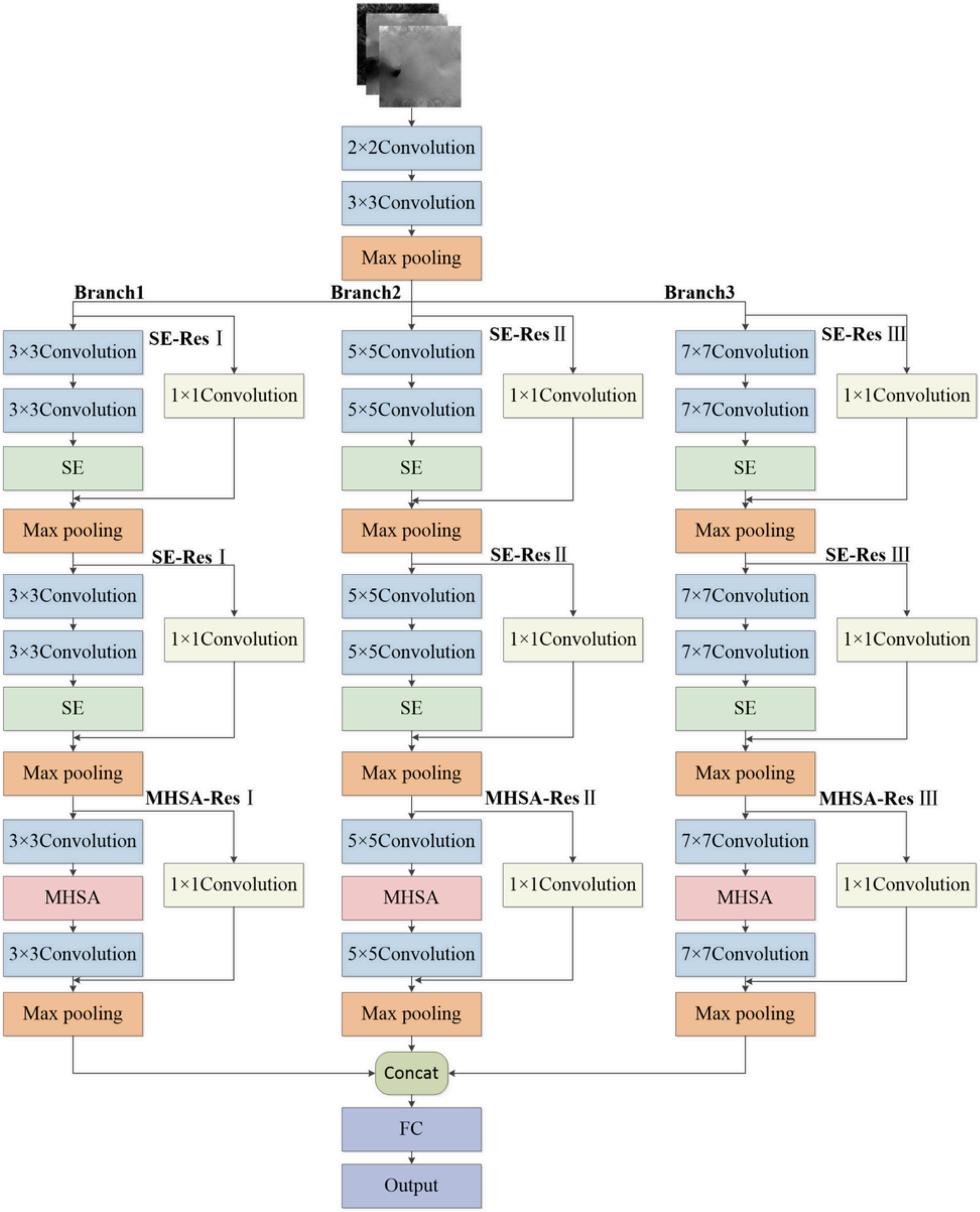

Figure 2

The global architecture of MFE-Net. The kernel size for all max pooling is 2 x 2. The features obtained from the three branches are fused by the contact operation.

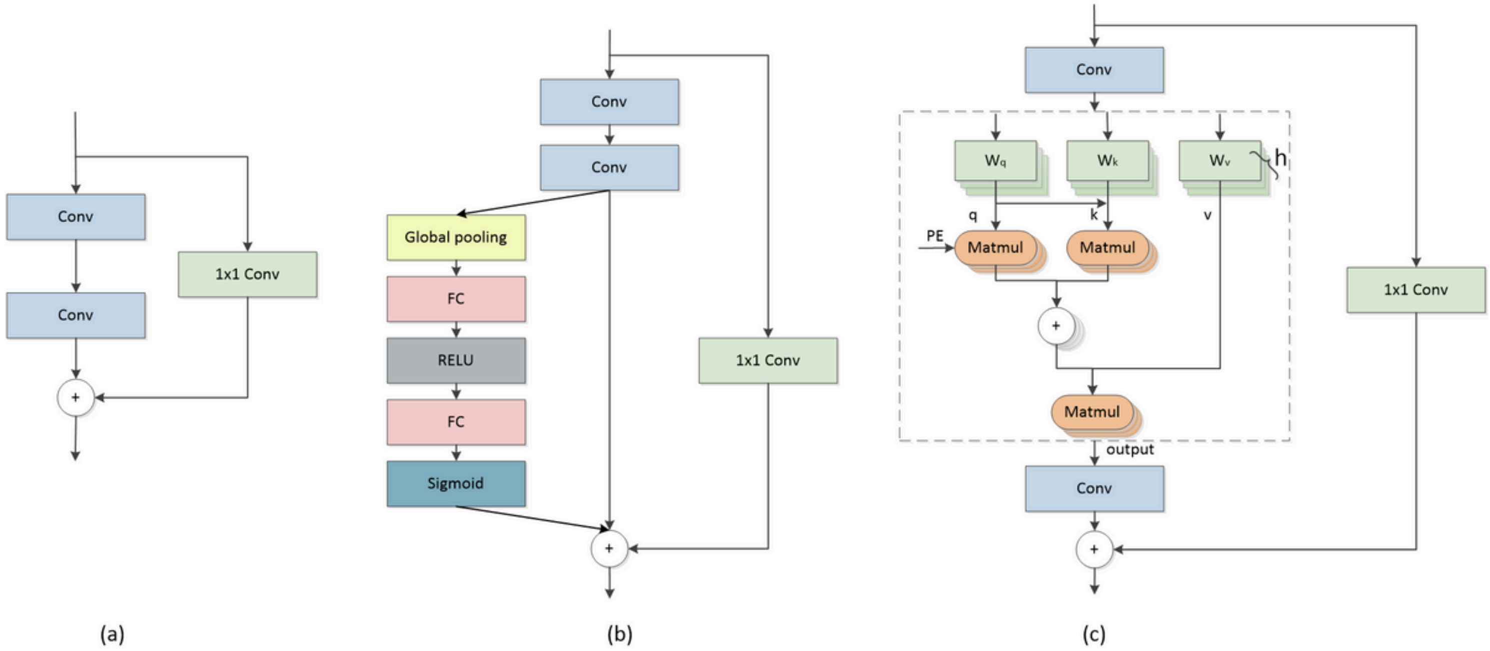

Figure 3

The structures of SE-Res and MHSA-Res blocks. (a) is the structure of the original Res-block, (b) shows the structure of the SE-Res block, and (c) is the structure of the MHSA-Res block. The dashed box in (c) shows the specific operations of MHSA.