STCAM: Spatial-Temporal and Channel Attention Module for Dynamic Facial Expression Recognition

Weicong Chen, Dong Zhang, Ming Li, *Member, IEEE*, and Dah-Jye Lee, *Senior Member, IEEE*

Abstract— Capturing the dynamics of facial expression progression in video is an essential and challenging task for facial expression recognition (FER). In this paper, we propose an effective framework to address this challenge. We develop a C3D-based network architecture, 3D-Inception-ResNet, to extract spatial-temporal features from the dynamic facial expression image sequence. A Spatial-Temporal and Channel Attention Module (STCAM) is proposed to explicitly exploit the holistic spatial-temporal and channel-wise correlations among the extracted features. Specifically, the proposed STCAM calculates a channel-wise and a spatial-temporal-wise attention map to enhance the features along the corresponding feature dimensions for more representative features. We evaluate our method on three popular dynamic facial expression recognition datasets, CK+, Oulu-CASIA and MMI. Experimental results show that our method achieves better or comparable performance compared to the state-of-the-art approaches.

Index Terms— Dynamic facial expression recognition, 3D-Inception-ResNet, Channel attention, Spatial-temporal attention

1 Introduction

FACIAL expression is the most powerful and straightforward way for human to convey their emotion. In recent years, facial expression recognition (FER) has attracted research interests because of its prospects of application in human-computer interaction, driver monitoring and health care tasks. In the existing research on FER, several approachs were proposed to encode facial expression. Ekman et al. [1] defined six basic emotions based on cross-culture study, including anger, disgust, fear, happiness, sadness and surprise. The Facial Action Coding System (FACS) encodes facial expression into the combination of different Action Units (AUs) that represent the action parts of the human face [2]. Meanwhile, researchers used continuous model to encode human emotions, e.g. the valance and arousal model [3]. Although the FACS model and the continuous model represent a wide range of expressions, the category model based on the six basic emotions is still the most popular way to encode facial expression because of its unambiguous definition of each one of them.

Most early research on FER was image-based, which recognize facial expression from a single still image by extracting the spatial information of the image. However, the expression features extracted from a single image are easily intertwined with various confusion factors that do not correlate with the expression itself. Age, gender, and the identity of the subject are some of those confusion factors that affect the performance. Researchers have attempted to explicitly incorporate the subject identity information in their algorithms to minimize the impact of the inter-subject variance of the expression [4], [5], [6]. These methods attempted to exclude part of the factors that are not expression related which may confuse the classifier.

Some research have shown that human can recognize facial expression through the dynamics of facial proression from the neutral face to the expressive face [7], [8]. Inspired by this idea, optical flow was utilized to improve the accuracy in image-based FER task [9]. Methods were developed to model the correlation between the neutral face and the expression face based on deep generative networks [10], [11]. These networks are usually not trained end to end, which makes them unsuitable for practical applications. A better approach to capture the dynamics of facial progression is to recognize facial expression from video or image sequence. Some research focused on modeling the temporal correlation among facial expression frames in the video by modeling temporal correlation separately from spatial correlation [12], [13], [14], [15]. The biggest challenge with this approach is the ambiguity in the extracted spatial features. Others require auxiliary input, e.g. facial landmarks [13], [16], which may introduce the registration error.

To address the aforementioned issues, a 3D-Inception-ResNet-LSTM was proposed to extract the spatial-temporal features of the facial expression sequence [17]. It obtains a series of weight maps from the landmarks of each face and incorporate them into the feature maps to generate enhanced features that have better discriminability than the original features. This method has two limita-

- *Weicong Chen and Dong Zhang are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, 510006. E-mail: chenwc7@ mail2.sysu.edu.cn, zhangd@mail.sysu.edu.cn.*
- *Ming Li is with the Duke Kunshan University, Kunshan, China, 215316. E-mail: ming.li369@dukekunshan.edu.cn.*
- *Dah-Jye Lee is with the Department of Electrical and Computer Engineering, Brigham Young University, Provo, Utah, USA, 84602 E-mail: djlee@byu.edu*
- *Corresponding author: Dong Zhang, Email: zhangd@mail.sysu.edu.cn*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3027340, IEEE Transactions on Affective Computing

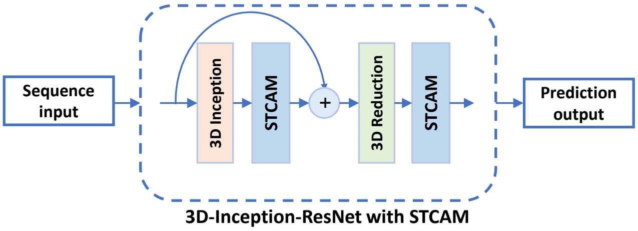

Fig. 1. The proposed framework for sequence-based facial expression recognition, including the baseline network called 3D-Inception-Resnet and the spatial-temporal attention module named STCAM. This figure presents only the key part of the whole network.

tions. Firstly, the weight map generated from the landmarks suppresses the features located far from the landmarks, regardless of their contributions to classification. Secondly, it employs only the weight maps along spatial dimension, treating the features from different times and channels equally.

Recently, attention mechanism has been widely used for deep learning-based image classification tasks. The Squeeze-and-Excitation Networks (SE-Nets) applied squeeze and excitation attention mechanism along the channel dimension to exploit the inter-channel relationship. It improved the recognition accuracy compared to using only the baseline networks [18]. Woo et al. achieved better result than the SE-Nets by applying additional spatial attention to the feature maps [19]. Both works showed that exploiting the global correlation within different dimensions helps with classification accuracy. Similarly, as a type of classification, FER can also take advantage of modeling and exploiting the inter-dimensional relation.

In this paper, we propose a novel framework to tackle the challenges of sequence-based FER. As shown in Fig. 1, we develop an effective 3D-Inception-ResNet as our baseline model to extract the spatial and temporal features simultaneously. The 3D-Inception-ResNet architecture has shown to achieve good performance [17]. We modify the original architecture to better extract the spatial-temporal features. We then utilize attention mechanism to generate enhanced spatial-temporal features in order to take the advantage of the global correlations among features and obtain a better representation of facial expression in image sequence. More specifically, we propose a learnable structure named Spatial-Temporal and Channel Attention Module (STCAM) to learn saliency maps along different dimensions of the spatial-temporal features. These attention maps are then multiplied with the original features to generate enhanced features. We integrate the STCAM into the baseline model as the feature extractor. Experimental results on popular dynamic facial expression datasets, including the Extended Cohn-Kanade Dataset (CK+) [20], Oulu-CASIA [21], and MMI [22], show that our method achieves better or comparable performance compared to the state-of-the-art approaches.

2 Related Work

2.1 Facial expression recognition based on dynamic image sequence

As a dynamic event, sequence-based FER is more reliable than image-based because temporal correlation among frames can be used to minimize ambiguity. Traditional FER methods extract spatial-temporal features by hand-crafted feature descriptors, such as LBP-TOP [23], HOG-TOP [24]. Liu et al. proposed a more powerful spatial-temporal descriptor STM-ExpLet based on a mid-level representation called expressionlet. It achieved better performance than traditional hand-crafted descriptors [25].

In recent years, deep learning has been used to tackle a wide range of computer vision tasks including sequence-based FER. A 3D CNN with deformable action parts constraints (3D CNN-DAP) was proposed to apply specific part filters to obtain partial dynamic features [26]. A 3D CNN without weight sharing along the time axis was proposed to capture the temporal appearance relation and incorporate the facial landmarks as the auxiliary data to model the geometrical motion information [16]. Although 3D CNN was employed in these two methods to capture the spatial-temporal relation, only one or two layers of 3D filters were used, which failed to fully describe the complicated spatial-temporal correlation among the frames in the whole facial expression sequence. A lightweight network called LBVCNN (Local Binary Volumn CNN) was proposed to extract spatial-temporal features from facial expression image sequences [27]. However, the representation power of the features extracted by LBVCNN was not as good as traditional CNNs.

Another way to process the facial expression sequence is to model the spatial and temporal correlation separately. A CNN-RNN framework was proposed to capture the spatial information with a multi-signal convolutional neural network (MSCNN) and to handle the facial geometrical evolution as the temporal information with a part-based hierarchical bidirectional recurrent neural network (PHRNN) [13]. 2D CNNs were used to extract spatial information and were followed by RNN (LSTM, GRU, BRNN) to model temporal correlation [12], [14], [15]. A spatial-temporal recurrent neural network (STRNN) was used to capture the spatial and temporal dependency from the features extracted by a pretrained CNN [28]. A frame attention module was proposed to aggregate the features extracted from each frame [29]. A framework consisting a spatial network and a temporal network was used to extracte spatial and temporal features saperately, and then aggregate the features by a BiLSTM [30]. Yu et al. proposed a global-local framework to capture global and local spatial features, and modeled the temporal information with LSTM [31]. These methods improved classification performance by using the temporal information of image sequence in addition to the spatial information of individual frames. They tend to intertwining factors such as age, gender, and subject identity in the spatial features, which may confuse the classifier.

A C3D network was proposed to extract spatial-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3027340, IEEE Transactions on Affective Computing

temporal features directly [17]. Specifically, a 3D-Inception-ResNet was used to extract spatial-temporal features, followed by a LSTM to further model the temporal relation. Weight masks obtained from facial landmarks were employed to enhance the extracted features. One downside of this approach is that useful information may be filtered out by the weight masks obtained from the facial landmarks.

2.2 Attention mechanism

Attention mechanism was first introduced in machine translation [31], [32]. The alignment between the target word and parts of the source sentence was learned to predict the most relevant target word [31]. Luong et al. proposed a global/local attention to respectively attend to all/parts of the source sentence [32]. More recently, Vaswani et al. introduced a transformer model with self-attention and achieved great success in machine translation [33].

Attention mechanism was also applied to many computer vision tasks. Similar to [32], Xu et al. proposed soft/hard attention for image captioning [34]. Hu et al. proposed to apply channel-wise attention in squeeze-and-excitation operation for image classification [18]. Woo et al. went further by additionally integrating spatial attention [19]. An attention map for each video frame based on soft attention was learned for video action analysis [35]. Saliency mask was generated for spatial attention and temporal attention, respectively, to enhance the discriminability of the features extracted by CNN and LSTM [36]. A three-stream network was developed for low resolution video analysis [37], which integrated multi-head self-attention to capture the temporal saliency among features extracted from different video clips. Non-local block was proposed as an extension of self-attention to capture the spatial-temporal dependency among the features extracted by CNN [38].

As a video analysis task, sequence-based FER can also benefit from attention mechanism. Inspired by [18], [19], we propose a structure named Spatial-Temporal and Channel Attention Module (STCAM) for dynamic FER. Different from the aforementioned attention methods in video analysis, which are mostly designed for 2D CNN + RNN network and operate spatial and temporal attention separately, our STCAM is designed for C3D network to handle spatial-temporal attention jointly. In addition, we integrate channel-wise attention to fully exploit the inter-channel dependencies.

2.3 Facial expression recognition with attention mechanism

The Facial Action Coding System (FACS) [2] has indicated the strong relationship between facial expression and specific parts of human face. In other words, distinct parts of the face contribute differently to facial expression. Therefore, adapting attention mechanism to FER has potential to enhance the salient part of feature and suppress redundant information. An AU-aware Deep Networks (AUDN) was proposed to automatically select the receptive fields corresponding to the facial expression [39]. Salient patches of facial image were selected based on facial landmarks and extracted features from these patches were used to classify facial expression [40]. A visual attention map was generated by aggregating CNN features of different channels and used to enhance the salient part of features [41]. A Facial-Motion Mask Generator (FMG) was proposed to highlight the facial motion parts while presenting expression [42]. A salient mask obtained from facial landmarks was used to enhance the features close to the landmarks [17]. The aforementioned methods adapted attention mechanism to successfully improve the FER performance. However, these works either concentrated on image-based FER [39], [40], [41], [42], or only applied attention along spatial dimension [17].

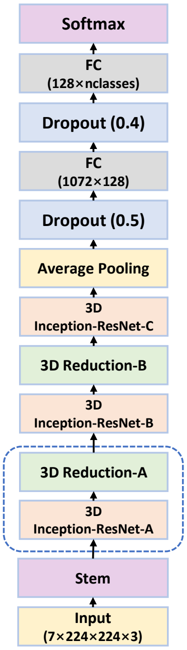

Fig. 2. The architecture of the baseline model (3D-Inception-ResNet) The network includes a stem structures, three 3D-Inception-ResNet structures, and two 3D-Reduction structures. Two fully connected layers serve as a classifier at the end of the network.

In order to take full advantage of the features extracted by deep spatial-temporal networks, we propose a learnable structure named Spatial-Temporal and Channel Attention Module (STCAM) for FER. The proposed network is able to automatically learn attention maps from the global correlation among channel and spatial-temporal dimensions of features and enhance the features along the corresponding dimensions.

3 Proposed Method

3.1 Baseline model

C3D has achieved great success in video-based vision tasks, e.g. action recognition [43], [44], [45]. By additionally integrating the temporal dimension into traditional 2D convolution, C3D is capable of capturing spatial-temporal features from image sequences. As opposed to 2D CNNs methods which extract spatial features of individual frames and then fuse them to model their temporal rela-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3027340, IEEE Transactions on Affective Computing

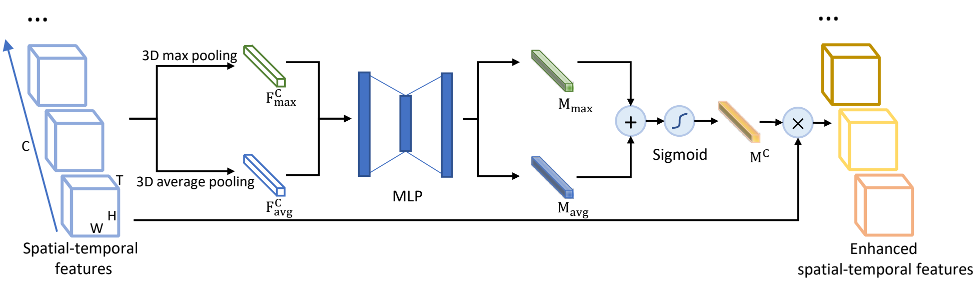

Fig. 3. The proposed 3D Channel Attention Module (3D-CAM). Two pooling operations are firstly applied to the spatial-temporal features. Two weight vectors M_{avg} and M_{max} are calculated by a multi-layer perceptron (MLP) and then they are added together and passed through a sigmoid function to generate the final attention map M^C.

tion, C3D handles spatial-temporal relation directly. In other words, C3D is able to describe the dynamic facial expression progression in the video with low-level and high-level features. This unique characteristic makes it more efficient than 2D CNNs methods for sequence-based facial expression recognition.

Similar to [17], we choose the 3D-Inception-ResNet as our baseline model to extract spatial-temporal features because of its efficient structure. Considering the small size of currently available sequence-based FER datasets, we design a relatively shallow network that includes three 3D-Inception-ResNet structures and two 3D-reduction structures, as shown in Fig. 2. Since the dimensions of the spatial and temporal domains of the input sequence are different (e.g. the spatial dimension of the network input is 224 × 224 while the temporal dimension is 7), the kernel size and stride in the baseline network are set differently in the spatial and temporal dimensions. For example, in the first convolution layer of the stem module, the kernel size (denoted as *time* × *height* × *width*) is 3 × 7 × 7 and the stride (denoted as ($stride_{temporal}$, $stride_{spatial}$)) is (1,2). After the average pooling layer, two fully connected layers and a softmax layer serve as a classifier which projects the 1072-dimension features into *n* classes probability values.

3.2 Spatial-Temporal and Channel Attention Module (STCAM)

Although the baseline model can effectively extract spatial-temporal features, the extracted features are simply the combination of information along the spatial-temporal and channel dimensions within local receptive fields. In fact, features extracted from different locations of the face contribute differently to the classification result, and features from specific channels may be more important than features from other channels with respect to a specific expression class. In other words, the importance of different channels and different spatial-temporal locations are variant in the extracted feature maps. Works have shown that the attention mechanism helps the network to focus on the important features by assigning higher weight to the corresponding locations in the feature maps [18], [19].

Inspired by this idea, we propose a Spatial-Temporal and Channel Attention Module (STCAM) to generate more powerful representation of dynamic facial expression. The attention weight maps are computed by modelling the interdependencies inside the features and assigned to the original feature maps to highlight the different importance among features during feature extraction. STCAM includes two sub-modules, called 3D Channel Attention Module (3D-CAM) and Spatial-Temporal Attention Module (STAM). Details of our design are illustrated in Sections 3.2.1 and 3.2.2.

3.2.1 3D Channel Attention Module

To explore the channel dependency among spatial-temporal features, a 3D Channel Attention Module (3D-CAM) is constructed to aggregate information across spatial-temporal dimensions and generate a channel-wise attention map.

As shown in Fig. 3, given a spatial-temporal feature map $\mathbf{F} \in \mathbb{R}^{C \times T \times H \times W}$, two 3D pooling operations are used to aggregate spatial-temporal information and generate two feature vectors, denoted as $\mathbf{F}^{C}_{avg} \in \mathbb{R}^{C}$ and $\mathbf{F}^{C}_{max} \in \mathbb{R}^{C}$:

$$\mathrm{F}^{C}_{avg}(c) = \frac{1}{T \times H \times W} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{F}(c,t,h,w) \qquad (1)$$
$$\mathrm{F}^{C}_{max}(c) = \max_{t,h,w} \mathbf{F}(c,t,h,w)$$

where $\mathbf{F}(c) \in \mathbb{R}^{T \times H \times W}$ refers to the c^{th} channel of $\mathbf{F}$. We use both 3D average pooling and 3D max pooling because they can gather complementary information for each other [19].

After the pooling operations, the two feature vectors simultaneously pass through a Multi-Layer Perceptron (MLP) with one hidden layer and one output layer to generate two weight vectors: M_{avg} and M_{max}. The weights of the hidden layer and output layer are denoted as $\mathbf{W}_1 \in \mathbb{R}^{(C/r) \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times (C/r)}$, where r is the reduction ratio used to avoid overfitting. The hidden layer is followed by a ReLU activation function for nonlinearity. Then the information contained in two weight vectors are united by element-wise addition and activated by a sigmoid function. In summary, the channel-wise attention map is generated by:

$$\mathrm{M}^{C} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathrm{F}^{C}_{avg}) + \mathbf{W}_2 \delta(\mathbf{W}_1 \mathrm{F}^{C}_{max})) \qquad (2)$$

where σ and δ indicate the sigmoid and ReLU function.

Finally, the weighted feature map is obtained by multiplying the spatial-temporal feature map with the channel-wise attention map:

$$\mathbf{F}^{'}(c) = \mathrm{M}^{C}(c)\mathbf{F}(c) \qquad (3)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3027340, IEEE Transactions on Affective Computing

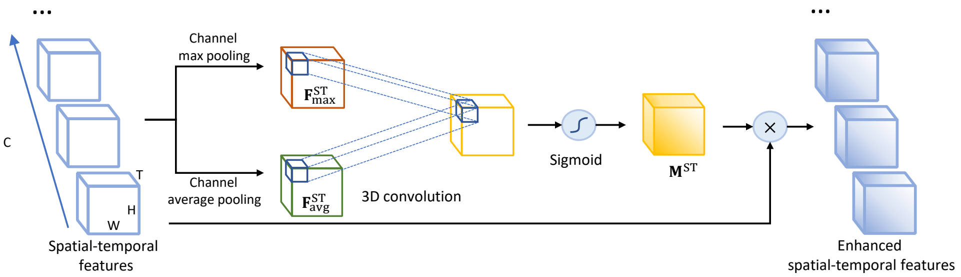

Fig. 4. The proposed Spatial-Temporal Attention Module (STAM). Two channel-wise pooling operations are firstly applied to the spatial-temporal features to aggregate information from different channels. Then, a 3D convolution kernel and a sigmoid function are used to generate the final spatial-temporal attention map M^{ST}.

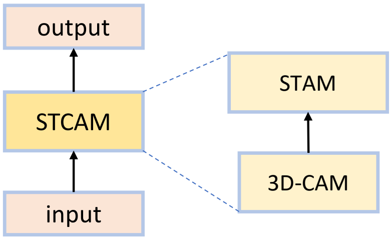

Fig. 5. The arrangement of STCAM. Specifically, we place 3D-CAM (3D Channel Attention Aodule) and STAM (Spatial-Temporal Attention Module) of the STCAM in a sequential manner.

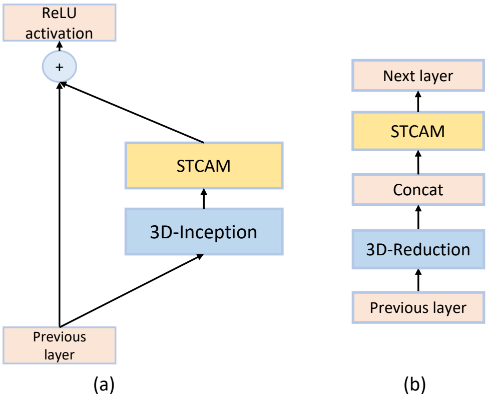

Fig. 6. (a) The integration of STCAM with 3D-Inception-ResNet structure. (b) The integration of STCAM with 3D-Reduction structure.

3.2.2 Spatial-Temporal Attention Module

Except for channel dependency, there are also global spatial-temporal correlations among the extracted features at different positions, indicating “where” and “when” the features are salient. A Spatial-Temporal Attention Module (STAM) is employed to capture these global spatial-temporal correlations.

Similarly, given a spatial-temporal feature map, $\mathbf{F} \in \mathbb{R}^{C\times T\times H\times W}$, the information from different channels are first gathered by average pooling and max pooling. Different from channel attention, the pooling operations are applied along the channel dimension to generate two spatial-temporal descriptors, denoted as $\mathbf{F}^{\mathrm{ST}}_{\mathrm{avg}} \in \mathbb{R}^{T\times H\times W}$ and $\mathbf{F}^{\mathrm{ST}}_{\mathrm{max}} \in \mathbb{R}^{T\times H\times W}$ (see Eq. (4)).

$$\mathbf{F}^{\mathrm{ST}}_{\mathrm{avg}}(t,h,w)=\frac{1}{C}\sum_{c=1}^{C}\mathbf{F}(c,t,h,w) \qquad (4)$$
$$\mathbf{F}^{\mathrm{ST}}_{\mathrm{max}}(t,h,w)=\max_{c}\mathbf{F}(c,t,h,w)$$

After the pooling operations, the spatial-temporal descriptors are concatenated along the channel axis, and a 3D convolution layer followed by a sigmoid function is used to compute the spatial-temporal attention map. The 3D convolution layer has two input channels (average pooling and max pooling) and one output channel, aggregating information into the final attention map, $\mathbf{M}^{\mathrm{ST}} \in \mathbb{R}^{T\times H\times W}$:

$$\mathbf{M}^{\mathrm{ST}}=\sigma(f([\mathbf{F}^{\mathrm{ST}}_{\mathrm{avg}},\mathbf{F}^{\mathrm{ST}}_{\mathrm{max}}])) \qquad (5)$$

where f denotes the 3D convolution. Note that every element in $\mathbf{M}^{\mathrm{ST}}$ is the weight of the corresponding spatial-temporal position of $\mathbf{F}$.

Finally, the feature map $\mathbf{F}$ is multiplied by the spatial-temporal attention map $\mathbf{M}^{\mathrm{ST}}$ to generate the weighted feature:

$$\mathbf{F}^{'}(c,t,h,w)=\mathrm{M}^{\mathrm{ST}}(t,h,w)\mathbf{F}(c,t,h,w) \qquad (6)$$

3.3 Attention modules integration

In this section, we explain the placements of 3D-CAM and STAM, and the integration of STCAM with the baseline model.

3.3.1 Combination of attention modules

For a spatial-temporal feature map $\mathbf{F}$, neither 3D-CAM nor STAM changes the shape of the original feature map. Therefore, the proposed attention modules can be easily combined in a sequential or a parallel manner. In this work, we put the two attention modules in sequence, as shown in Fig. 5, because the output of 3D-CAM helps the performance of STAM. In other words, the feature map modified by the first attention module (3D-CAM) provides useful information for computing the next attention map. An experiment that explores the order of the two attention modules is reported in Section 4.3.

3.3.2 Integration with baseline

As mentioned before, the shape retaining property of STCAM makes it easy to integrate the STCAM with any C3D-based architectures, including our baseline model. As discussed in Section 3.1, our baseline model contains two kinds of sub-structures, the 3D-Inception-ResNet structure and the 3D-reduction structure. By integrating these structures with the STCAM, we are able to improve

TABLE 1
CONCLUSION OF DATASETS

Dataset	Subjects	Sequences	Manner [a]
CK+	118	327	N→P
Oulu-CASIA	80	480	N→P
MMI	30	208	N→P→N

[a]*N: Neutral, P: Peak*

their ability to extract spatial-temporal features.

The integration of the two network sub-structures with STCAM is shown in Fig. 6. For the 3D-Inception-ResNet structure, the STCAM is placed after the 3D-Inception in a separate branch. After the STCAM, the modified features are added to the features from the previous layer (shortcut connection). For the 3D-reduction structure, the STCAM is inserted directly into the path of the 3D-reduction structure.

4 EXPERIMENTS

4.1 Datasets

Experiments in this paper were conducted on three popular dynamic facial expression datasets, including the Extended Cohn-Kanade Dataset (CK+) [20], Oulu-CASIA [21] and MMI [22].

CK+: The CK+ dataset is the most popular sequence-based facial expression dataset to assess the performance of FER. It includs 593 facial expression sequences from 123 subjects. Among these sequences, 327 facial expression sequences from 118 subjects are labeled with seven basic expressions (Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise). Each sequence begins with the neutral face and ends with the peak expression. Since CK+ does not provide official splits of training, validation and test sets, a 10-fold cross-validation protocol was adopted to assess the performance. More specific, we sampled the subjects in an ID ascending order with a step size of 10 to construct 10 subject-independent subsets, as in the previous works [13], [14], [16], [25].

Oulu-CASIA: The Oulu-CASIA dataset consists of 80 subjects, presenting 6 basic expressions (Anger, Disgust, Fear, Happiness, Sadness, Surprise) under 3 illumination conditions of 2 types of imaging system, i.e., near-infrared (NIR) and visible light (VIS). Only the 480 sequences taken under the strong illumination condition of the VIS were used in our experiment. Similar to CK+, the sequences in Oulu-CASIA begin with the neutral face and end with the peak expression. And the same 10-fold cross-validation protocol as CK+ was employed in our experiment on Oulu-CASIA.

MMI: The MMI is a small dataset, containing 236 sequences from 32 subjects. 208 sequences of the front view of 30 subjects are labeled as one of six basic expressions. Unlike CK+ and Oulu-CASIA, each sequence in MMI begins with the neutral face, reaches the peak expression in the middle, and finally ends with the neutral face. Similar to the previous 2 datasets, a subject-independent 10-fold cross-validation protocol was employed in MMI dataset evaluation.

TABLE 2
EVALUATION OF STCAM ON CK+

Model	Accuracy
Baseline	98.47%
Baseline + 3D-CAM	98.78%
Baseline + STAM	98.47%
Baseline + STAM + 3D-CAM	98.47%
Baseline + 3D-CAM + STAM	99.08%

TABLE 3
EVALUATION OF STCAM ON OULU-CASIA

Model	Accuracy
Baseline	89.16%
Baseline + 3D-CAM	90.21%
Baseline + STAM	89.79%
Baseline + STAM + 3D-CAM	90.62%
Baseline + 3D-CAM + STAM	91.25%

TABLE 4
EVALUATION OF STCAM ON MMI

Model	Accuracy
Baseline	77.40%
Baseline + 3D-CAM	78.84%
Baseline + STAM	78.37%
Baseline + STAM + 3D-CAM	77.89%
Baseline + 3D-CAM + STAM	82.21%

The number of subjects, available number of sequences, and the expression manner of the three datasets are summarized in Table 1.

4.2 Implementation details

Data preprocessing: Instead of using the original frames directly, data preprocessing is necessary to improve the FER performance. In our experiments, facial landmarks were used to determine the bounding box of the face area. For CK+, we utilized the provided landmarks label directly. Since the Oulu-CASIA and MMI datasets do not provide facial landmarks, the famous Dlib library [46] was used to detect the frontal faces and facial landmarks. After that, all faces were cropped by the bounding box and resized to 240 × 240.

Data augmentation: In order to avoid overfitting, we applied on-the-fly data augmentation during training. Specifically, random horizontal flip and random crop of four corners and center were utilized during training. The random cropping operation crops the images from the size of 240 × 240 to 224 × 224 as the input to our model. During the testing state, only the center crop of the size 224 × 224 was applied.

Model pretraining: Before training on our target datasets, i.e. CK+, Oulu-CASIA and MMI, we pretrained our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3027340, IEEE Transactions on Affective Computing

TABLE 5
COMPARISON WITH STATE-OF-THE-ART ON CK+

Method	Accuracy
STM-ExpLet [25]	94.19%
3DCNN-DAP [26]	92.40%
DTAGN [16]	97.25%
PHRNN-MSCNN [13]	98.50%
CNN-GRU [14]	98.47%
STRNN [28]	95.40%
LBVCNN [27]	97.38%
FAN [29]	99.69%
MGLN-GRU [31]	99.08%
Baseline	**98.47%**
Baseline + STCAM	**99.08%**

TABLE 6
COMPARISON WITH STATE-OF-THE-ART ON OULU-CASIA

Method	Accuracy
STM-ExpLet [25]	74.59%
DTAGN [16]	81.46%
PHRNN-MSCNN [13]	86.25%
CNN-GRU [14]	91.67%
LBVCNN [27]	82.41%
DSN + DTN + BiLSTM [30]	91.07%
MGLN-GRU [31]	90.40%
Baseline	**89.16%**
Baseline + STCAM	**91.25%**

model on a large-scale action recognition dataset, Kinetics [44], for a better initialization of the model weights, aimed to accelerate training and avoid overfitting. We pretrained the model using stochastic gradient descent (SGD) with Nesterov momentum of 0.9. The pretraining learning rate was 0.001, and the weight decay was 5e-4. The model was pretrained for 8 epochs and the batch size was 32. While training on target datasets, the convolution layers were initialized by the pretrained model weights, and the fully connected layers were initialized by Xavier [47].

Frames selection: Our goal is to capture the progression from the neutral face to expressive face. However, using all frames from neutral face to expressive face will include redundant information or very subtle differences between two successive frames and require more computation resources. We down sampled every sequence into 7 frames as the input to the model. More specific, we selected the neutral face (the first frame of the sequence) as the first frame and the peak expression (the last frame of CK+ and Oulu-CASIA sequences, and the middle frame of MMI) as the last frame, and then sampled evenly between the neutral frame and the peak frame to obtain another 5 frames. It should be noted that sample jittering was applied to augment the training data during the training state.

Parameters settings: We optimized our network on CK+, Oulu-CASIA and MMI using stochastic gradient descent (SGD) with Nesterov momentum of 0.9. The network was trained for 100 epochs. The learning rate was initially set to 0.001 and then divided by a factor of 10 on 40th and 90th epoch. The batch size was 16 and the weight decay was 0.01. The reduction ratio in 3D-CAM was set to 16 according to [18] and [19], and the kernel size of the 3D convolution in STAM was set to 3 × 7 × 7 for 3D-Inception-ResNet-A module and 3 × 5 × 5 for other C3D modules.

TABLE 7
COMPARISON WITH STATE-OF-THE-ART ON MMI

Method	Accuracy
STM-ExpLet [25]	75.12%
3DCNN-DAP [26]	63.40%
DTAGN [16]	70.24%
PHRNN-MSCNN [13]	81.18%
DSN + DTN + BiLSTM [30]	80.71%
Baseline	**77.40%**
Baseline + STCAM	**82.21%**

4.3 Evaluation of STCAM

We evaluated our proposed method on three dynamic facial expression datasets, CK+, Oulu-CASIA and MMI. All experiments were conducted on a subject-independent 10-fold cross-validation protocol.

We evaluated the two sub-modules of STCAM, i.e., 3D-CAM and STAM separately. The results on the three datasets are shown in Tables 2, 3 and 4. Compared to the baseline model, utilizing 3D-CAM or STAM individually obtained a slight recognition accuracy improvement on the three datasets. The performance of 3D-CAM is slightly better than STAM, indicating that the global channel-wise correlation can provide more information than the holistic spatial-temporal correlation.

We then put the 3D-CAM and STAM in sequence in different orders to compare their performances. Tables 2, 3 and 4 show the comparison results on the three datasets. Utilizing STAM and 3D-CAM together in sequence improved the recognition accuracy, comparing to the baseline model. However, different orders led to different results. As shown in our experiments, placing 3D-CAM before STAM performed significantly better than placing STAM first. On all three datasets, 3D-CAM + STAM achieved the best results in our experiments.

Our analysis on the difference in performance is that the 3D-CAM enhances some important channels of the feature map by channel attention, and the STAM aggregates information across the channel dimension by the channel pooling operations to generate the spatial-temporal descriptors. In other words, the channel pooling operations can focus more on the important channels enhanced by the 3D-CAM. Therefore, the spatial-temporal descriptors generated from the feature maps enhanced by 3D-CAM preserve more useful information than that generated from the original feature maps, leading to the better result when placing 3D-CAM first. Also notice that the STAM + 3D-CAM placement led to more significant improvement on Oulu-CASIA than on CK+ and MMI. This was because part of the subjects in the Oulu-CASIA dataset wear glasses that reflect the computer screen. The reflection on the glasses occluded the face. The placement of STAM + 3D-CAM coped with the occlusion on Oulu-CASIA better than the other two datasets.

TABLE 8
COMPARISON WITH NON-LOCAL ATTENTION METHOD

	CK+	Oulu-CASIA	MMI
Baseline	98.47%	89.16%	77.40%
NL-Gaussian	98.78%	90.21%	78.37%
NL-Embedded Gaussian	98.78%	90.00%	78.84%
STCAM	99.08%	91.25%	82.21%

TABLE 9
CONFUSION MATRIX OF BASELINE + STCAM ON CK+

	An	Co	Di	Fe	Ha	Sa	Su
An	**0.978**	0	0.022	0	0	0	0
Co	0	**1**	0	0	0	0	0
Di	0	0	**1**	0	0	0	0
Fe	0	0	0	**0.960**	0.040	0	0
Ha	0	0	0	0	**1**	0	0
Sa	0	0.036	0	0	0	**0.964**	0
Su	0	0	0	0	0	0	**1**

TABLE 10
CONFUSION MATRIX OF BASELINE + STCAM ON OULU-CASIA

	An	Di	Fe	Ha	Sa	Su
An	**0.875**	0.05	0.0125	0.0125	0.05	0
Di	0.0875	**0.8625**	0	0	0.05	0
Fe	0	0	**0.9375**	0.025	0.0125	0.025
Ha	0	0	0.0125	**0.975**	0.0125	0
Sa	0.075	0.0375	0.0125	0.0125	**0.8625**	0
Su	0	0	0.0375	0	0	**0.9625**

TABLE 11
CONFUSION MATRIX OF BASELINE + STCAM ON MMI

	An	Di	Fe	Ha	Sa	Su
An	**0.76**	0.03	0.03	0.03	0.15	0
Di	0.0625	**0.8125**	0	0.0625	0.0625	0
Fe	0	0.036	**0.643**	0.036	0.036	0.25
Ha	0	0.024	0	**0.976**	0	0
Sa	0.125	0.0625	0.0625	0	**0.75**	0
Su	0	0.049	0	0.024	0.024	**0.902**

4.4 Comparison with state-of-the-art methods

We also compared our method with the existing sequence-based FER methods on CK+, Oulu-CASIA and MMI. Specifically, we denote our 3D-CAM + STAM model as STCAM.

Comparison result on CK+: Table 5 shows the comparison result on the CK+ dataset. The hand-crafted feature, STM-ExpLet [25] achieved an accuracy of 94.19%, and outperformed the deep learning-based 3DCNN-DAP [26]. STRNN [28] captured the spatial and temporal dependency from the features and achieved an accuracy of 95.40%. DTAGN [16] and PHRNN-MSCNN [13] reported performance improvement by jointly utilizing the appearance information extracted from the image and the geometrical information extracted from facial landmarks, reaching 97.25% and 98.50% on CK+. Kuo et al. extracted spatial features with CNN and temporal features with GRU [14], which reached an accuracy comparable to [13]. MGLN-GRU [31] achieved 99.08% accuracy by utilizing a complex multitask model and GRU. FAN [29] reported an accuracy of 99.69% by using frame attention in addition to a deep network. As shown in Table 5, our baseline model achieved a competitive result comparing to most state-of-the-art methods. The proposed STCAM applies attention mechanism on various dimensions of the spatial-temporal features to achieve an accuracy of 99.08% on CK+, which is better than most state-of-the-art methods and competitive with FAN [29] and MGLN-GRU [31].

Comparison result on Oulu-CASIA: For the Oulu-CASIA dataset, the comparison result is shown in Table 6. Our baseline model achieved an accuracy of 89.16%, which exceeded other methods by a large gap except the CNN-GRU [14], MGLN-GRU [31] and DSN + DTN + BiLSTM [30]. We checked our misclassified samples in the Oulu-CASIA dataset and found that some subjects present different expression in a very similar way, e.g. anger and disgust, which is nearly impossible for human to classify them all correctly. The local dynamic facial expression progression in this situation may confuse our baseline model, which extracts spatial-temporal features simultaneously and hierarchically from local to holist. The proposed STCAM computes a weight map for distinct channels of features, which suppresses irrelevant features and therefore promote the classification accuracy. Finally, the accuracy of our method was just 0.42% lower than the accuracy reported by [14], and it was better than all other state-of-the-art methods.

Comparison result on MMI: The MMI dataset is more challenging than the other two datasets because some subjects in this dataset present an expression with and without facial accessories such as glasses or present an expression with very different intensity. Also, the size of MMI is much smaller than CK+ and Oulu-CASIA, which makes it harder for deep learning-based method to recognize facial expression. The comparison result of MMI is shown in Table 7. PHRNN-MSCNN [13] achieved 81.18% on MMI which is the highest accuracy among the existing sequence-based methods. Our proposed STCAM outperformed PHRNN-MSCNN [13] and other methods by utilizing spatial-temporal and channel attention mechanism, which eliminates the confusing factors and generate more representative features.

4.5 Comparison with Non-local attention method

We compared the proposed STCAM with the Non-local attention method [38] using the CK+, Oulu-CASIA and MMI datasets. Similar to STCAM, the Non-local block doesn’t change the shape of feature maps. Therefore, we simply replaced STCAM with the Non-local block in our network and compared them under the same experimental settings. We followed the same subject-independent 10-fold cross-validation for accuracy evaluation.

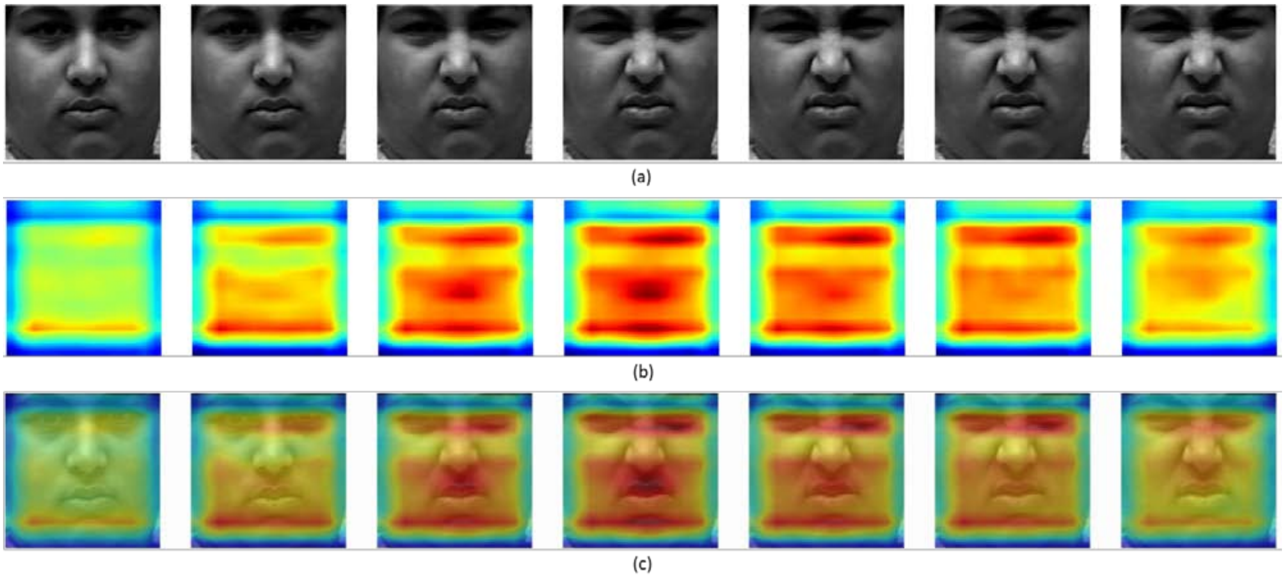

Fig. 7. Visualization of the learnt attention maps (a) The selected frames of an example video. (b) The learnt spatial-temporal attention maps. (c) The overlap of (a) and (b).

We compared STCAM with two different kernels of Non-local block, including the Gaussian kernel (denoted as NL-Gaussian) and the Embedded Gaussian kernel (denoted as NL-Embedded Gaussian). The comparison result is shown in Table 8. Both NL-Gaussian and NL-Embedded Gaussian achieved better performance than the baseline network without the attention module because the Non-local operation models the holistic spatial-temporal dependency and rebuilds the spatial-temporal features. STCAM outperformed both NL-Gaussian and NL-Embedded Gaussian by applying spatial-temporal attention and channel attention together.

4.6 Confusion matrix analysis

The confusion matrixes of the proposed baseline + STCAM method on CK+, Oulu-CASIA and MMI datasets are shown in Tables 9, 10, and 11.

Confusion matrix on CK+: The confusion matrix of our method on the CK+ dataset is shown in Table 9. Our method made no mistake on contempt, disgust, happy and surprise, while only misclassified one sample for each class on anger, fear and sadness. This impressive result demonstrated the effectiveness of our method.

Confusion matrix on Oulu-CASIA: Table 10 presents the confusion matrix on the Oulu-CASIA dataset. Our method performed well on fear, happy and surprise. Notice that anger is easily mixed with disgust and sadness because some subjects present these expressions indistinctively.

Confusion matrix on MMI: The confusion matrix on MMI is shown in Table 11. Our method achieved the best performance on happy and surprise. It is interesting that one fourth of the samples from fear were misclassified as surprise. This was because the training samples of fear were relatively fewer than other classes, and some of its samples were with different intensity, making them very similar to surprise.

4.7 Visualization

A visualization of the learnt attention maps that are used to explore the relation between the attention maps and the input video is shown in this section. Including the visualization of spatial-temporal attention maps is helpful because the spatial and temporal dimensions are straightforward for human to understand.

Fig. 7(a) presents the selected frames of an expression video from the CK+ dataset. These frames were input into our network to extract the spatial-temporal features. Fig. 7(b) shows the spatial-temporal attention maps corresponding to the extracted features generated by the proposed STCAM. For visualization, the spatial-temporal attention maps were split along the time axis and interpolated to the same spatial size as the input frame. The warm tone of the attention maps corresponds to the high weight, while the cold tone corresponds to the low weight. As shown in Fig. 7(c), we overlapped the input frames with the attention maps to explore the correlation between them. In the aspect of spatial dimensions, the attention maps assigned higher weight to the important parts of the face that are correlated to the expression, e.g. the eyes and the mouth. From the perspective of temporal dimension, the attention maps concentrated more on the frames in which the expression changes rapidly. Fig. 7 shows that STCAM is able to enhance the important parts of the features and supress the less relevant ones.

5 Conclusions

In this paper, we propose a novel framework to address the dynamic facial expression recognition task. In order to capture the dynamic progression of facial expression, we develop a C3D-based architecture called 3D-Inception-ResNet as our baseline network to extract the spatial-temporal features.

We proposed a spatial-temporal and channel attention module (STCAM) to explore and utilize the global correlations among channel and spatial-temporal dimensions. STCAM includes two sub-modules, 3D Channel Attention Module (3D-CAM) and Spatial-Temporal Attention Module (STAM). 3D-CAM aggregates information across spatial-temporal dimensions of the extracted features to explore the channel dependency. It also generates a channel-wise attention map to enhance the class-specific features according to the channel dependency. STAM explores the spatial-temporal correlations among features by aggregating information across channel dimension and generates a spatial-temporal attention map to highlight "where" and "when" the features are important. These two sub-modules are placed in sequence to form the STCAM, which is integrated into the baseline model as the feature extractor.

We evaluated our method on three widely used sequences-based FER datasets, CK+, Oulu-CASIA, and MMI. The experimental results demonstrated that our method achieved better or comparable performance compared with the state-of-the-art approaches.

Acknowledgment

This work was supported by Guangzhou Municipal People's Livelihood Science and Technology Plan (201903010040), National Natural Science Foundation of China (61773413), and the Science and Technology Planning Project of Guangdong Province of China (2019B070702004).

References

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[2] P. Ekman and W. V. Friesen, "Facial action coding system: a technique for the measurement of facial movement," 1978.

[3] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[4] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-Aware Convolutional Neural Network for Facial Expression Recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, DC, USA, 2017, pp. 558–565.

[5] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, 2017, pp. 522–531.

[6] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Identity-Free Facial Expression Recognition using conditional Generative Adversarial Network," *ArXiv190308051 Cs*, Mar. 2019.

[7] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "The distributed human neural system for face perception," *Trends Cogn. Sci.*, vol. 4, no. 6, pp. 223–233, Jun. 2000.

[8] A. J. Calder and A. W. Young, "Understanding the recognition of facial identity and facial expression," *Nat. Rev. Neurosci.*, vol. 6, no. 8, pp. 641–651, Aug. 2005.

[9] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognit. Lett.*, vol. 119, pp. 49–61, Mar. 2019.

[10] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," *ArXiv170307140 Cs*, Mar. 2017.

[11] H. Yang, U. Ciftci, and L. Yin, "Facial Expression Recognition by De-expression Residue Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 2168–2177.

[12] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 223–236, Apr. 2017.

[13] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[14] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A Compact Deep Learning Model for Robust Facial Expression Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, 2018, pp. 2202–22028.

[15] Y. Cai, W. Zheng, T. Zhang, Q. Li, Z. Cui, and J. Ye, "Video Based Emotion Recognition Using CNN and BRNN," in *Pattern Recognition*, vol. 663, Singapore: Springer Singapore, 2016, pp. 679–691.

[16] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2983–2991.

[17] B. Hasani and M. H. Mahoor, "Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 2278–2288.

[18] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Computer Vision – ECCV 2018*, Cham, 2018, vol. 11211, pp. 3–19.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, 2010, pp. 94–101.

[21] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[22] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, 2005.

[23] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[24] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial Expression Recognition in Video with Multiple Feature Fusion," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 38–50, Jan. 2018.

[25] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning Expression-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3027340, IEEE Transactions on Affective Computing

lets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749–1756.

[26] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis," in *Computer Vision -- ACCV 2014*, Singapore, 2015, vol. 9006, pp. 143–157.

[27] S. Kumawat, M. Verma, and S. Raman, "LBVCNN: Local Binary Volume Convolutional Neural Network for Facial Expression Recognition From Image Sequences," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 207–216.

[28] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–Temporal Recurrent Neural Network for Emotion Recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019.

[29] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame Attention Networks for Facial Expression Recognition in Videos," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3866–3870.

[30] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," *Vis. Comput.*, vol. 36, no. 3, pp. 499–508, Mar. 2020.

[31] M. Yu, H. Zheng, Z. Peng, J. Dong, and H. Du, "Facial expression recognition based on a multi-task global-local network," *Pattern Recognit. Lett.*, vol. 131, pp. 166–171, Mar. 2020.

[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ArXiv14090473 Cs Stat*, Sep. 2014.

[33] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," *ArXiv150804025 Cs*, Aug. 2015.

[34] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[35] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2048–2057.

[36] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action Recognition using Visual Attention," *ArXiv151104119 Cs*, Nov. 2015.

[37] L. Meng *et al.*, "Interpretable Spatio-temporal Attention for Video Action Recognition," *ArXiv181004511 Cs Stat*, Oct. 2018.

[38] D. Purwanto, R. R. A. Pramono, Y. Chen, and W. Fang, "Three-Stream Network With Bidirectional Self-Attention for Action Recognition in Extreme Low Resolution Videos," *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1187–1191, Aug. 2019.

[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[40] M. Liu, S. Li, S. Shan, and X. Chen, "AU-aware Deep Networks for facial expression recognition," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[41] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[42] W. Sun, H. Zhao, and Z. Jin, "A visual attention based ROI detection method for facial expression recognition," *Neurocomputing*, vol. 296, pp. 12–22, Jun. 2018.

[43] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial Motion Prior Networks for Facial Expression Recognition," *ArXiv190208788 Cs*, Feb. 2019.

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 4489–4497.

[45] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4724–4733.

[46] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6546–6555.

[47] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *J. Mach. Learn. Res.*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[48] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

Weicong Chen received his B.S. degree from Sun Yat-sen University, China, in 2018. He is currently a graduate student at the School of Electronics and Tenology, Sun Yat-sen University. His research interests include facial expression recognition and computer vision.

Dong Zhang received his B.S.E.E. and M. S. degrees from Nanjing University, China, in 1999 and 2003, respectively, and Ph.D. degree from Sun Yat-sen University, China, in 2009. He is currently an associate professor in the School of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, pattern recognition and information hiding.

Ming Li received his Ph.D. in Electrical Engineering from University of Southern California in May 2013. He is currently an associate professor of Electrical and Computer Engineering at Duke Kunshan University, a research scholar at the ECE department of Duke University, and the adjunct professor at Wuhan University. His research interests are in the areas of speech processing and multimodal behavior signal analysis with applications to human centered behavioral informatics notably in health, education and security.

Dah-Jye Lee received his B.S. degree from National Taiwan University of Science and Technology in 1984, M.S. and Ph.D. degrees in electrical engineering from Texas Tech University in 1987 and 1990, respectively. He also received his MBA degree from Shenandoah University, Winchester, Virginia in 1999. He worked in the machine vision industry for eleven years prior to joining BYU in 2001. He is currently a Professor in the Department of Electrical and Computer Engineering at Brigham Young University. His research work focuses on object recognition, hardware implementation of real-time vision algorithms and machine vision applications.