# Multi-view Geometry Consistency Network for Facial Micro-Expression Recognition From Various Perspectives

Yawen Lu, Student Member, IEEE
*Intelligent Vision and Sensing Lab*
*Rochester Institute of Technology*
Rochester, NY, USA
yl4280@rit.edu

Nikola Kasabov, FIEEE
*School of Eng., Comp. and Math Sc.*
*Auckland University of Technology*
Auckland, New Zealand
and University of Ulster, UK
nkasabov@aut.ac.nz

Guoyu Lu, Member, IEEE
*Intelligent Vision and Sensing Lab*
*Rochester Institute of Technology*
Rochester, NY, USA
luguoyu@cis.rit.edu

***Abstract*—Micro-expression can reveal underlying genuine emotions, but those rapid and subtle changes are hard to be captured by humans. Most existing research focuses on frontal face micro-expression recognition, which largely prevents the developed methods from the real applications and ignores the underlying geometry information. In this paper, we propose a multi-view geometry consistency framework to enable the same emotion to be recognized under different perspectives, which is difficult for existing systems. Based on the developed 3D face reconstruction network, the multi-view micro-expression recognition framework empowers the emotion recognition capability to learn from multiple perspectives of the 3D reconstructed faces based on view-consistency, and a spiking neural network is further applied to capture omitted tiny and detailed changes. With a sequence of images, we explore the subtle changes across frames through optical flow, which, as a clue, enhances the performance of our designated network for micro-expression recognition. Extensive experiments on benchmark micro-expression datasets CAS(ME)$^2$ and SMIC demonstrate the proposed method achieves promising results on novel-view micro-expression recognition where existing methods mainly fail.**

Keywords: Micro-expression Recognition; 3D Face Reconstruction; Multiple View Geometry; Spiking Neural Networks;

## I. Introduction

Facial expression, as a major human inner state reflection, has been applied in many applications such as product evaluation, mental health diagnosis, and criminal interrogation, etc. [1] [2]. In contrast of general macro- expression, facial micro-expression represents subtle expression changes that only last for very short time period (e.g., 50 ms) and is difficult to be observed by human eyes. From another perspective, the correct recognition of micro-expression can facilitate extensive services in real life, because it can reflect the true mental condition and emotion of a person which is very hard to hide.

There are quite limited efforts on micro-expression recognition compared to other recognition tasks due to its complexity and difficulties. Classic feature-based approaches include utilizing local binary pattern [3], local binary pattern on three orthogonal planes (LBP-TOP) [4], LBP-Six-Intersection-Points (LBP-SIP) [5] and directional mean optical flow [6].

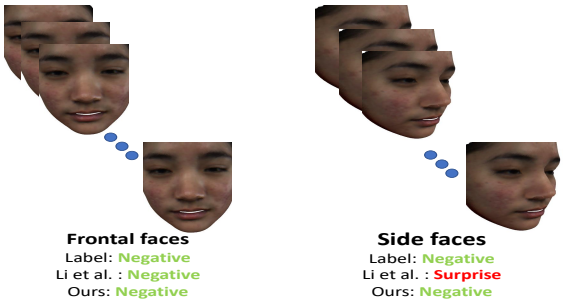


Fig. 1. An illustration of our view-consistent micro-expression recognition. Our framework is able to achieve accurate micro-expression recognition from different perspectives that existing methods [7] cannot deal with.

However, such methods which heavily rely on hand-crafted features and frontal images are insufficient to extract critical and subtle expression changing information from multi-views as illustrated in Fig. 1.

Recently, deep neural networks like convolutional neural networks (CNN) have been widely used to solve relevant tasks [8] [9] [10] [11] [12] [13] [14]. Compared to the handcraft-based methods, the deep learning-based methods can obtain better performance in most of the computer vision problems. However, those deep learning techniques proposed for facial micro-expression recognition are mostly trained and tested on exactly frontal faces by feeding video clips to CNN, RNN or a combination of them. This requirement is very strict and limited in the real-world applications.

To address this problem, we developed a method and a system to recognize micro-expressions from multiple views and perspectives of video clips. In case of a frontal image, we developed a method to generate 3D multiple views of the raw frontal facing face image inputs, which enables the designed micro-expression recognition network to learn to recognize the micro-expression from various perspectives. Besides commonly used CNN modules, we apply 3D CNN and bidirectional ConvLSTM networks to extract the temporal information between adjacent frames across the video clips to facilitate understanding continuous changes in facial expres-

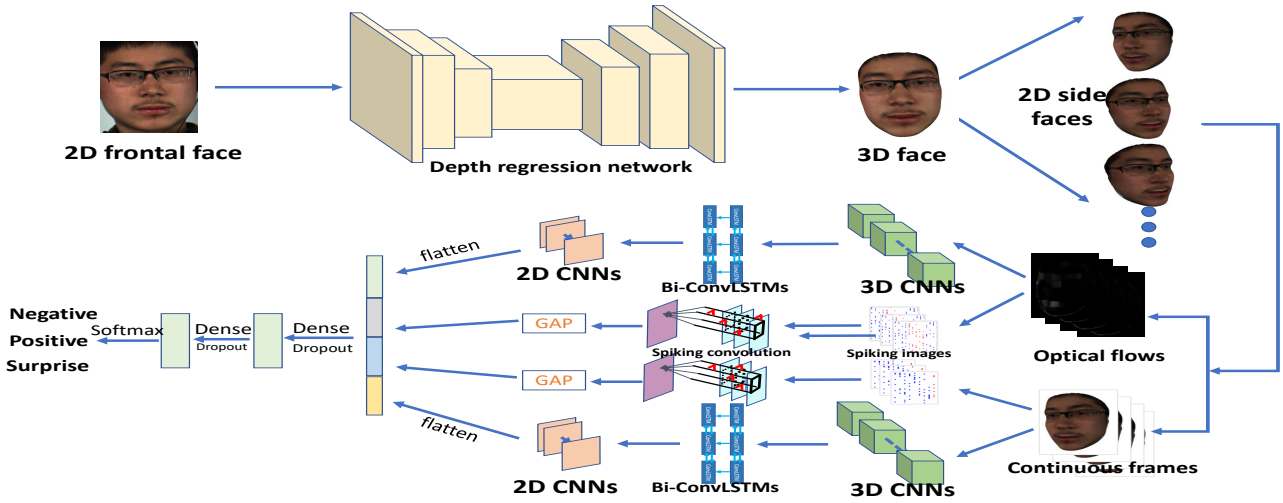


Fig. 2. Overview of the proposed multi-view consistency micro-expression recognition architecture. Following the multi-view expression generation (Top), continuous frames and the corresponding optical flows are fed simultaneously to output the final category (Bottom).

sion. A spiking neural network (SNN) is further applied to deal with the short and tiny changes represented and learned as spikes. In order to overcome the limitations that micro-expression may only occur on some key regions of interest that cannot be reflected on other views during the rotation, we introduce an interpretable CNN to analyze regions in frontal faces that play the most important role, and filter out those views which are invisible in the generated side face image sequences used for training. Therefore, our proposed pipeline is able to learn an interpretable and geometry-consistent micro-expression recognition system from different views.

The main contributions of the proposed method can be summarized as follows: 1. A view-consistent micro-expression recognition model is proposed to extract consistent spatial and temporal information from video clips to capture subtle and short expression changes, which is infeasible for existing methods only focusing on frontal faces. 2. We apply an interpretable CNN to visualize the crucial regions contributing for micro-expression recognition and filter out the generated face views that are invisible to the micro-expression, which helps to better understanding the multi-view micro-expression under various views. 3. In addition to existing facial micro-expression datasets with only frontal faces, we contribute with multi-view micro-expression datasets benefiting from the designed multi-view face generation framework and an efficient method to generate those datasets with only front-oriented face images. 4. A designated micro-expression recognition network is proposed which integrates for the first time optical change clues with spiking neural networks.

## II. Related Work

In this section, we first introduce some techniques for generating 3D face from images, which is important for generating our multi-view face data. Then existing approaches designed to recognize emotion categories are introduced. They are split into conventional methods using handcrafted visual features and deep methods to learn a classifier in an end-to-end network.

### A. 3D Face reconstruction methods

3D morphable model (3DMM) [15] is one of the most well-known model-based techniques in reconstructing 3D faces from images. Most early works aim to establish the correspondence between 2D images and 3D reconstructions using facial landmarks [16] [17] [18] or other local features (e.g., SIFT and HOG), and apply 3DMM coefficients to fit the full 3D face geometry, including occluded surfaces. The main shortcoming for 3DMM is that the results are over smooth and facial details (e.g., nuances of expression) are likely lost. Recent learning-based approaches explore to regress 3DMM coefficient with end-to-end network directly [18] [19]. Other single-view face reconstruction methods [20] [21] trained a CNN to generate 3D models fully supervised by ground truth 3D scans and achieve impressive results, and some weak-supervised methods [22] [23] proposed to decrease the dependence of high-quality 3D models during the training. However, a large amount of annotations are still in need, and the generalization abilities across dastasets are not good and stable. Unsupervised approaches are proposed to address the 3D face reconstruction issue by using images from multiple views. Sanyal et al. [22] enforced the shape consistency between the same identities and inconsistency among different people identities to constrain the captured multiple images from all participants. Wu et al. [24] proposed MVFNet to utilize photometric consistency between different viewpoints. More recent work [25] extends [24] to both geometric and photometric constraints to produce better results from a single 2D image. However, public micro-expression dastasets usually do not provide avaible multiview constraints or any ground truth face scans for training.

### B. Facial Micro-expression Estimation

In the past years, several works dealt with facial macro-expressions that use frontal face images only. The works on micro-expressions can be split into hand-crafted features based and deep learning methods. Conventional methods on facial

Micro-expression estimation focus on different feature extraction methods such as local binary pattern [3], local binary pattern on three orthogonal planes (LBP-TOP) [4], LBP-Six-Intersection-Points (LBP-SIP) [5] and directional mean optical flow [6]. Among them, Pfister et al. [26] proposed to use local binary patterns on three orthogonal planes for micro-expression feature extraction. LBP is a texture-based image feature extraction method and considers the size relationship between the intensity value of a pixel and surrounding neighboring pixels to encode. Liu et al. [6] proposed the main directional mean optical flow feature for expression recognition and then applyoptical flow on video clips to train a support vector machine (SVM) classifier to perform micro-expression estimation. Although these hand-crafted approaches can achieve good performance, it brings heavy cost in domain specificity and tedious parameter adjustment.

Deep learning based methods like VGG-19 [27] and ResNet [28], have achieved remarkable results in computer vision area, to solve relevant tasks including expression classification [29], face detection [30], face segmentation [31], and face reconstruction [32] etc. Among such works, [8] embedded the landmark information together with the extracted features from RGB images into the CNN to increase the performance for micro-recognition. Liong et al. [10] designed a new feature descriptor to extract flow features from their off-apex network and then feed the extracted features to a normal CNN structure for classification output. [11] further inputted the extracted feature vectors through CNN modules to the long short-term memory (LSTM) to better learn the temporal information. [12] targeted at mining the data to solve the class imbalance issue to extract robust facial features from the processed motion magnified EAI images. However, existing deep learning based techniques proposed for facial micro-expression recognition all focused on recognizing exactly frontal faces, resulting in a impracticable in real-world scenarios, as strict front-facing images are very rare in normal lives.

## III. The Proposed Multi-view Micro-expression Recognition

In this section, we summarize the entire framework and introduce the major contributions, as shown in Fig. 2. To generate side face views, we first conduct 3D face reconstruction via a depth regression network based on a single front-facing image, which will be applied to project to various perspectives (Section III-A). Then we apply an interpretable deep learning model to analyze the key contributed regions in the face and filter out the occluded sequences (Section III-B). With the refined multiple view face images and the corresponding optical flows, a spatial-temporal network for micro-expression recognition is proposed (Section III-C), and a spiking neural network (SNN) (Section III-D) is further applied to capture subtle changes and handle temporal information.

### A. Multi-view Expression Generation

To obtain multi-view expressions, we first apply depth regression network to estimate an accurate depth map given a processed frontal facing image, and then recover a 3D face model with the known camera intrinsics. The 3D face models are able to be imposed to generate novel views by projecting the 3D face from different perspectives. To achieve high-quality and high-accurate reconstruction performance, we impose depth supervision on the frontal face image input to recover the face shape. However, as there is no ground truth depth labels for the existing micro-expression datasets, it is required to train the depth regression network on the 3D face dataset with provided face depth labels, and then transfer the model to the micro-expression recognition datatsets by constraining the two types of datasets to share same image properties (e.g., covering regions, image size etc), which leads to similar image appearance. We utilize a ResNet-18 based encoder structure and multiple upconvolutional operations to estimate the final depth. An overview of the network structure is given in Table I.

Fig. 3. First column: processed frontal face images from the CAS(ME)$^2$ micro-expression dataset. Rest columns: the corresponding multi-view face images ranging from -45 degree to 45 degree generated from 3D reconstruction face.

To guide the learning process of the depth regression network, we deploy the smoothness $L1$ loss function. Smoothness $L1$ term is widely applied in 2D and 3D bounding box regression in object detection task which has favorable robustness and low sensitivity to the outliers. We adopt it to get a balance both from $L1$ and $L2$ losses as:

$$L(D(p), \widetilde{D}(p)) = \frac{1}{M}\sum_{p=1}^{M}\Psi(D(p) - \widetilde{D}(p)) \tag{1}$$

where,

$$\Psi(x) = \begin{cases} \frac{1}{2}x^2, & |x| < 1 \\ |x| = 0.5, & else \end{cases} \tag{2}$$

To bridge the gap between the 3D face datasets and existing micro-expression datasets, we first project the reconstructed 3D face models in public 3D face datasets to generate 2D frontal faces, and apply MTCNN [33] to detect, crop and resize the face images to share the same region and size. Simultaneously, we processed the raw images from micro-expression dataset in the same manner to maintain the appearance commonality. After training, we are able to generate high-quality 3D faces on micro-expression datasets. Thus multi-view images can be created by projecting the reconstructed

| layer | kernel size | channel | function |
|---|---|---|---|
| Conv1 | 7×7 | 64 | ReLU |
| Max_pool | 3×3 | | – |
| Conv_2 | 3×3 | 64 | – |
| Conv_3 | 3×3 | 128 | – |
| Conv_4 | 3×3 | 256 | – |
| Conv_5 | 3×3 | 512 | – |
| Upconv_5 | 3×3 | 256 | ReLU |
| iconv5 | 3×3 | 256 | ReLU |
| Upconv_4 | 3×3 | 128 | ReLU |
| iconv4 | 3×3 | 128 | ReLU |
| Upconv3 | 3×3 | 64 | ReLU |
| iconv3 | 3×3 | 64 | ReLU |
| Upconv2 | 3×3 | 32 | ReLU |
| iconv2 | 3×3 | 32 | ReLU |
| Upconv1 | 3×3 | 16 | ReLU |
| iconv1 | 3×3 | 16 | ReLU |
| Disp | 3×3 | 1 | Sigmoid |

TABLE I
THE STRUCTURE OF FACE DEPTH REGRESSION NETWORK. THE LINE IN THE MIDDLE SEPARATES THE ENCODER AND DECODER.

Fig. 4. Sampled continuous frames from the generated multi-view dataset (first and third rows), and the corresponding optical flows (second and fourth rows).

3D face model onto 2D image planes. Generated multi-view images from CAS(ME)$^2$ micro-expression dataset are illustrated in Fig. 3.

After projection, the images projected from the 3D models may contain small offsets, which could affect the micro-expression recognition effect due to the influence of the large amount moving pixels caused by the misalignment. To overcome the misalignment, detected facial landmarks and ORB features are applied to register the images. The registration is completed by estimating the homography matrix between video frames based on corresponding matches after RANdom SAmple Consensus (RANSAC) [34] filtering. Once the images are aligned, the small movement on the face surface and organs will play a significant to the micro-expression recognition that cannot be easily observed by naked eyes. To learn more subtle information and fuse them into the recognition network, we compute the optical flow for each

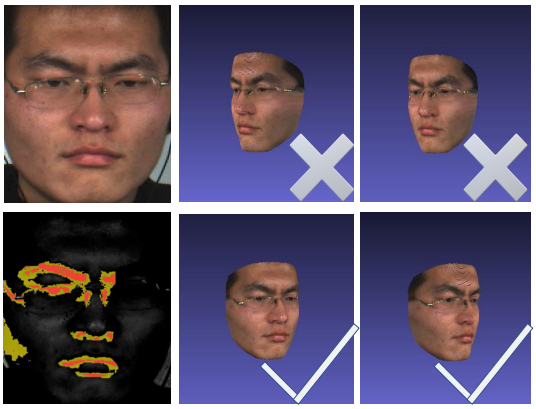

Fig. 5. An illustration of the attention generated from the interpretable CNN model, adn the refinement process on the generated side faces. Based on the attention map on the frontal face, we filter out the top two generated sequences, and select the bottom two generated sequences.

frame relative to the previous frame, generating a sequence of optical flows. Then we simultaneously input the raw face image sequences and the optical flows, and fuse them together in the flatten layers to output the final classification. The sampled processed continuous frames and the corresponding optical flows are illustrated in Fig. 4.

### *B. Data Refinement with Interpretable CNN*

The initial multi-view data generation covers the full angles from -45 to +45 degrees. However, it ignores the possible occlusion on the key regions in the multi-view images during the horizontal rotation. Therefore, we introduce an attention map to observe the activation maps localizing each micro-expression class, then observe the regions of interest from the final convolutional layer, and use the interpretable attention map to filter out those sequences with occlusion issues.

More specifically, we keep the entire network structure to be the same as Sec. III-A, except that on the output layer decisions. Similar as [35], we denote $c$ the classification class, $K$ representing the corresponding feature maps. The importance at a position $(i, j)$ of the attention map $A$ can be computed as a linear combination of the multiplication of the feature maps from the last convolutional layer $C$ and the corresponding weights $w_k^c$:

$$A(i,j) = \sum_k w_k^c \cdot C_{ij}^k \tag{3}$$

where the weights $w_k^c$ for each particular feature map can be computed by the following equation:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\vartheta(\sum_k w_k^c \cdot \sum_i \sum_j C_{ij}^k)}{\vartheta C_{ij}^k} \tag{4}$$

where $Z$ is a constant of the number pixels in the activation map. We illustrate the attention map in Figure 5. The top two side face views that do not contain the indicated important regions among all the generated views. Other examples of the visual result of the weighted attention maps from the interpretable CNN are provided in Figure 6.

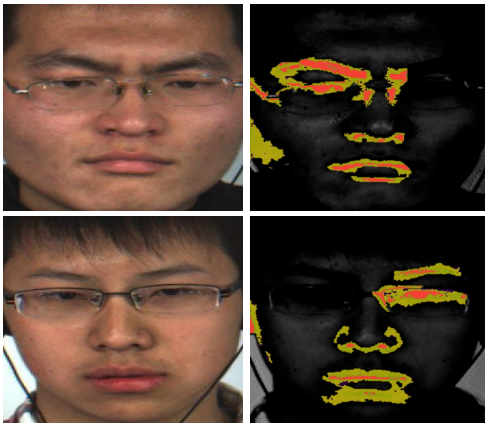

Fig. 6. Visual results of the attention map from the interpretable deep learning models on the micro-expression images. Darker color indicate more important.

### C. Spatial-temporal Micro-expression Recognition

3D CNN structure is proved to be effective in processing video clips for its spatial-and-temporal learning capability [36] [37], especially for short-term spatial-temporal relationships. In this work, we start with 3DCNN to learn the local spatial-temporal features from both raw color image sequences and optical flow sequences, respectively. As shown in the overview structure of Fig. 2, the proposed network applies 3D convolutional layers with a kernel size of 3×3×15 followed with a BatchNorm, a ReLu activation and a 3D max-polling layer sized in 3×3×3. The shallow temporal length is used to extract and learn only short and local temporal features from the continuous frames.

Following the shallow 3DCNN, we further apply a bi-directional ConvLSTM network to learn long-term feature representations across the entire sequences. The ConvLSTM replaces the fully connected layer in the normal LSTM blocks with convolutional layers to retain the spatial information and learn holistic and long-term information from the video. The ConvLSTM network structure can be expressed as:

$$
\begin{aligned}
\mathrm{i}_t &= Sigmiod(Conv_{xi} * x_t + Conv_{hi} * h_{t-1} + b_i) \\
\mathrm{f}_t &= Sigmiod(Conv_{xf} * x_t + Conv_{hf} * h_{t-1} + b_f) \\
\mathrm{o}_t &= Sigmiod(Conv_{xo} * x_t + Conv_{ho} * h_{t-1} + b_o) \\
\mathrm{C}_t &= f_t \odot C_{t-1} + i_t \odot tanh(Conv_{xc} * x_t + Conv_{hc} * h_{t-1} + b_c) \\
\mathrm{h}_t &= o_t \odot tanh(C_t)
\end{aligned}
\tag{5}
$$

where '*' denotes the operator of convolution, $\odot$ denotes the Hadamard product, and $Conv(\cdot)$ is the 2D Convolution kernel that applies to the input and hidden state respectively. The bi-direction ConvLSTM is an enhanced version to the ConvLSTM in which features from two directions extracted from hidden layers are utilized for each LSTM cell (one for a forward sequence and the other for a backward sequence). For each sequence, the features from the forward and backward hidden states are stacked together and fed into the convolutional layer to generate the final feature representations. Fig. 2 illustrates a general structure of the applied Bi-direction ConvLSTM in the middle part of our recognition network.

Due to the large spatial dimension from the Bi-ConvLSTM, a vanilla 2D-CNN structure is applied to reduce the dimensionality for the final micro-expression classification. The 2D-CNN structure is composed of two same blocks including a 2D convolutional layer, a batch normalization, a ReLu activation, and a pooling layer sequentially. Finally, a flatten layer is connected with the 2D-CNN blocks for both original face image sequence input and optical flow sequence input, concatenated together to one vector, and proceeded with dense and dropout layers to output the final classification with SoftMax function. Compared with existing 3D-CNN based structures, the introduced recognition network can learn both long-term global and short-term local information, and extract higher-level 2D features map simultaneously from the original sequences and optical flow maps, leading to highly accurate and adequate learning. Cross-entropy loss based on each category (surprise, positive and negative) is used to guide the training procedure as: $-\sum_{i=1}^{3} y_i \cdot log(\widetilde{y}i)$, where $y_i$ is the target labels from the dataset, and $\widetilde{y}_i$ is the corresponding output from the recognition model.

### D. Late Fusion with Spiking Neural Network

To further capture the subtle changes of facial micro-expression, and better handle the temporal information like videos and sounds, we propose to apply a spiking neural network (SNN) on the same inputs as part of the designed spatial-temporal recognition network, as depicted in Fig. 2. SNN are suitable for detecting fast and small changes from streaming data [39] [40].

The input is a 3D tensor with a shape of $(u, v, t)$ where $(u, v)$ denotes the image coordinate and $t$ is the defined time resolution within a image sequence. The sampled tensor is then fed into the SNN and convolved with a 3D spiking convolution kernel to generate spiking neuronal feature maps. Different from the conventional feature maps generated from CNN, the information at each coordinate of a spiking feature map is represented by a number of neuronal spike trains, which is more suitable to capture subtle details in the micro-expression videos. The final recognition is output from a global average pooling layer and two dense layers following the spiking feature maps.

To effectively fuse features extracted from different neural networks, and output the class probabilities of each video clips, we concatenate the feature maps from both diverse networks and inputs together following the late fusion structure to improve the accuracy of the predicted classes. The weights of various feature maps will be optimized in training.

## IV. Experiments

In this section, we first briefly introduce the public datasets and the generated multi-view datasets. Then we introduce the implementation process of the proposed network for multi-view consistent micro-expression recognition. Finally, we evaluate the proposed method on the processed multi-view dastaset in comparison with both classic and learning-based methods to demonstrate the advantages of our method on micro-expression recognition under diverse perspectives.

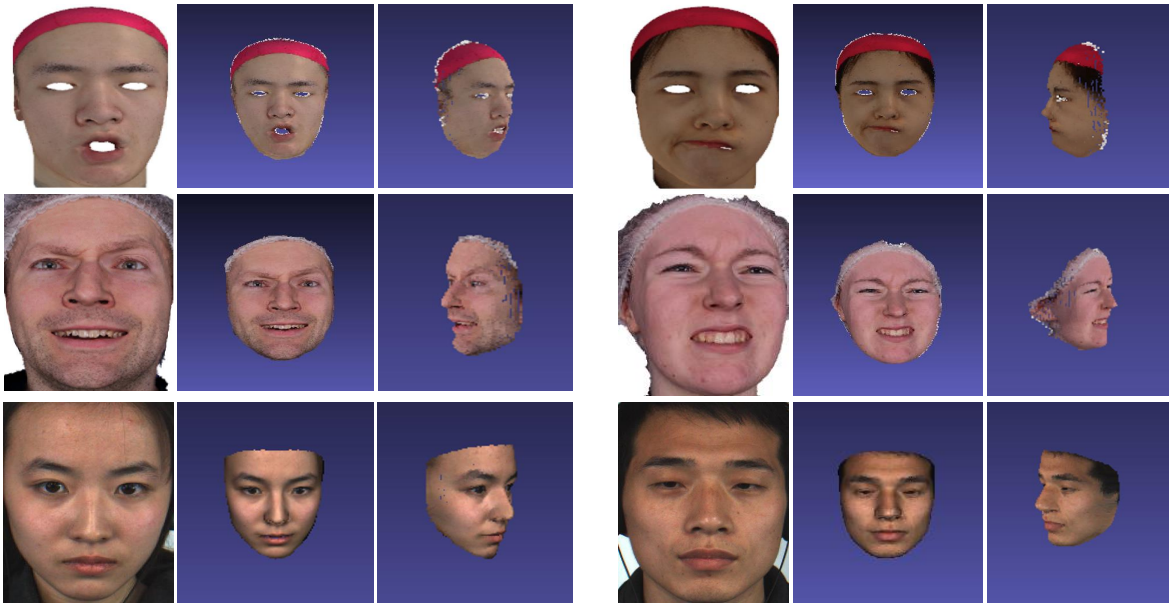

Fig. 7. Given a single color face image of frontal view, our method can produce a high-quality mesh (depicted in two different perspectives) that contains correct global shape and detailed textures on organs. Top to bottom: Examples from test split of Facescape dataset; Examples from test split of Stirling ESRC 3D face dataset; Examples from facial micro-expression CAS(ME)$^2$ dataset;

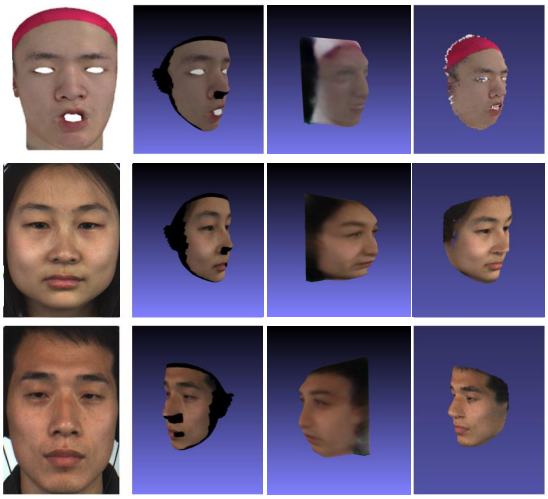

Fig. 8. Qualitative comparisons on facescape dataset (the first row) and on CAS(ME)$^2$ (the second and third rows) between our proposed method and recent state-of-the-art methods. Left to right: 2D face input; 3D reconstructed faces from [32]; 3D faces from [38]; Our 3D face result.

### A. Datasets

CAS(ME)$^2$ dataset [41] is the second version of the original CASME for facial micro expression. We split the entire dataset containing 206 video clips into three categories based on the labels: Surprise, Positive and Negative. 80% of data is applied to generate the multi-view images and then the optical flows calculated from the multi-view sequential frames. Generated multi-view frames are then sampled every ten degrees ranging from 45 degree to 135 degree horizontally. The processed face images and the optical flows will then be used for training. Similarly, the rest 20% will be used to build the testing set.

Spontaneous Micro Expression Database (SMIC) [42] is a dataset composed of spontaneous micro-expressions. Similarly to CAS(ME)$^2$ dataset, we use a total of 156 videos clips composed of the same three categories to generate the train data, maintaining the ratio between training and testing splits as CAS(ME)$^2$ dataset.

For 3D face datasets, Stirling ESRC 3D face dataset [43] is composed of 101 scanned subjects (male: 47, female: 54) in the format of wavefront objects. Facescape [44] is a large-scale dataset that contains a considerable number of high-quality 3D face subjects. The entire dataset is composed of 847 subjects with 20 expressions (totally 16490 models). We split 90% of all available 3D face models for training, and the rest 10% for validation. Blender 2.80 is used to project 3D models into 2D images to generate side-view face images together with their depth map. After our pre-processed introduced in Section III-A, we are able to train the depth regression network and utilize the trained network on the front facing images from the micro-expression datasets to generate multi-view face images.

### B. Experimental settings

The network for multi-view face generation is trained for 30 epochs with Adam optimizer [45] and a batch size of 8 on two NVIDIA 1080Ti GPUs. The learning rate for the first half epochs is 10e-4 and then gradually decreases to 10e-5 for the rest half. To make the dataset more robust, we perform a 50% possibility for random adjustness of brightness, contrast, saturation and hue in a range of $\pm 0.2$, $\pm 0.2$, $\pm 0.2$ and $\pm 0.1$.

For deep micro-expression recognition network with our multi-view data, we trained it from scratch on the CAS(ME)$^2$ and SMIC datasets, as stated in the Section III-A. SGD optimization is applied with a batch size of 32 and learning rate of 0.01. The recognition network is trained for 100 epochs and the best model with the lowest validation error is saved for testing. The spatial size of the input raw and optical flow

| Method | Type | CAS(ME)$^2$ | SMIC |
|---|---|---|---|
| LBP-TOP [42] | Hand-crafted | 0.33 | 0.30 |
| Takalkar et al. [46] | Deep Learning | 0.42 / 0.59 | 0.33 / 0.47 |
| Li et al. [7] | Deep Learning | 0.39 / 0.56 | 0.33 / 0.51 |
| MicroExpSTCNN [47] | Deep Learning | 0.52 / 0.72 | 0.39 / 0.56 |
| Our full | Deep Learning | **0.62 / 0.83** | **0.47 / 0.68** |

TABLE II
A COMPARISON WITH OTHER METHODS IN MICRO-EXPRESSION RECOGNITION USING HAND-CRAFTED FEATURES AND DEEP NEURAL NETWORKS. THE TESTING ACCURACY IS REPORTED ON BOTH CAS(ME)$^2$ AND SMIC DATASETS. FOR DEEP LEARNING BASED METHODS, THE RESULTS ON THE LEFT SIDE OF "/" REPRESENT TESTING OUTCOMES THAT THE METHODS TESTED ON MULTI-VIEW DATASETS BUT ONLY TRAINED ON THE ORIGINAL MICRO-EXPRESSION DASTASET; THE RIGHT SIDE OF "/" MEANS THAT THE METHOD TESTED ON THE MULTI-VIEW DATA IS ALSO TRAINED ON THE SPLIT. THE BEST RESULT FOR EACH DATASET AND SETTING IS HIGHLIGHTED IN BOLD.

| | Negative | Positive | Surprise |
|---|---|---|---|
| Negative | 46 | 0 | 2 |
| Positive | 3 | 28 | 0 |
| Surprise | 4 | 1 | 36 |

TABLE III
THE CONFUSION MATRIX OF OUR METHOD TESTING RESULT ON CAS(ME)$^2$ DATASET. VERTICAL AXIS INDICATES THE TRUE LABEL AND HORIZONTAL AXIS INDICATES OUR OUTPUTS.

images are resized to be 128 $\times$ 128. GPU version and memory space maintained the same as the 3D face generation.

### C. Experimental results

We demoed the performance of our reconstruction on Facescape and STIR 3D datasets (with ground truth) and the facial micro-expression dataset (without ground truth), as depicted in Fig. 7. By sharing the same image property such as detected region and image size, our method recovers good global shape on face surface as well as details on organs like nose and mouth on micro-expression images even without ground truth labels or multi-view constraints for training.

A comparison of the 3D faces reconstructed by different approachs from one single facial image is depicted in Fig. 8. Compared with [32], our reconstruction is able to perform more correct prediction in the depth estimation on nose and cheek, with full face surface and texture. Compared with [38], our reconstruction appear more real effect, and prevent the distortion in forehead and cheek in [38].

First, we evaluate the proposed view-consistent micro-expression recognition method both on CAS(ME)$^2$ and SMIC datasets. A comparison of testing accuracy in percentage with other recent methods [42] [46] [7] [47] is shown in Table II. As shown in Table II, we first can observe that for current

| | Negative | Positive | Surprise |
|---|---|---|---|
| Negative | 44 | 0 | 4 |
| Positive | 4 | 8 | 3 |
| Surprise | 2 | 2 | 23 |

TABLE IV
THE CONFUSION MATRIX OF OUR METHOD TESTING RESULT ON SMIC DATASET. VERTICAL AXIS INDICATES THE TRUE LABEL AND HORIZONTAL AXIS INDICATES OUR OUTPUTS.

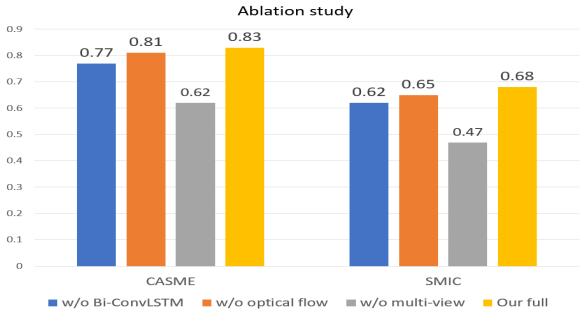


Fig. 9. Ablation study on core components of our method in terms of accuracy.

existing works focusing on frontal faces, there is a large gap in accuracy between the results on frontal faces and the side-view faces, which demonstrates the benefits of our proposed method in real application. Meanwhile, it can be noticed that the proposed deep architecture both achieves the state-of-the-art accuracy on the testing split of the generated multi-view with or without training on multi-view data. Specifically, the proposed method achieves an 11% and 12% improvement in recognition accuracy compared with the latest approach [47]. The confusion matrices with our proposed recognition model on CAS(ME)$^2$ and SMIC datasets are also provided in Table III and IV respectively, to avoid potentially misleading results in accuracy because of the unbalanced data. It can be observed that our method distinguishes positive and negative expressions very well, but may confuse the surprise and negative expressions.

An ablation study on key components of our method is conducted and results presented in Fig. 9. As the additional bi-direction ConvLSTM and the SNN are main differences from 3DCNN-based approaches, they are considered to contribute to the extraction of long-term holistic features from video clips. We observe 4% and 6% improvements on CAS(ME)$^2$ and SMIC datasets. The proposed methods for multi-view image data generation and micro-expression recognition result in a significant improvement on the recognition accuracy, 21% for both two datasets. In addition, optical flow inputs contribute with 2% and 3%.

## V. CONCLUSION

This work proposes a multi-view geometry micro-expression recognition framework from videos clips based on global and local spatial-temporal networks. Benefiting from our generated multi-view faces, we are able to recognize micro-expression accurately under different perspectives, which is difficult for other existing methods. The introduced recognition network leverages both local short-term and global long-term feature representations and incorporates both intensity information and optical flow information to achieve higher recognition accuracy. Ablation analysis verified the effectiveness of each core component designed in our full method, and extensive experiments on the two benchmark datasets demonstrate that the proposed method outperforms recent methods by a large margin on micro-expression recognition tasks, particularly for side-view faces.

## REFERENCES


[1] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces," in *IEEE FG*, 2013.

[2] W.-J. Yan, S.-J. Wang, Y.-J. Liu, Q. Wu, and X. Fu, "For micro-expression recognition: Database and suggestions," *Neurocomputing*, 2014.

[3] X. Huang, S.-J. Wang, G. Zhao, and M. Piteikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *ICCVW*, 2015.

[4] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE TPAMI*, 2007.

[5] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *ACCV*, 2014.

[6] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, 2015.

[7] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3d flow convolutional neural network," *Pattern Analysis and Applications*, 2019.

[8] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, "Multi-modality network with visual and geometrical information for micro emotion recognition," in *IEEE FG*, 2017.

[9] D. Y. Choi, D. H. Kim, and B. C. Song, "Recognizing fine facial micro-expressions using two-dimensional landmark feature," in *ICIP*. IEEE, 2018.

[10] S.-T. Liong, Y. Gan, W.-C. Yau, Y.-C. Huang, and T. L. Ken, "Off-apexnet on micro-expression recognition system," *arXiv preprint arXiv:1805.08699*, 2018.

[11] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *IEEE FG*, 2018.

[12] A. J. R. Kumar, R. Theagarajan, O. Peraza, and B. Bhanu, "Classification of facial micro-expressions using motion magnified emotion avatar images." in *IEEE CVPRW*, 2019.

[13] S. Miao, H. Xu, Z. Han, and Y. Zhu, "Recognizing facial expressions using a shallow convolutional neural network," *IEEE Access*, vol. 7, pp. 78 000–78 011, 2019.

[14] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "Learnet: Dynamic imaging network for micro expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 1618–1627, 2019.

[15] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[16] E. Richardson, M. Sela, and R. Kimmel, "3d face reconstruction by learning from synthetic data," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 460–469.

[17] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[18] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5163–5172.

[19] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7346–7355.

[20] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3d face reconstruction from a single image via direct volumetric cnn regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1031–1039.

[21] X. Zeng, X. Peng, and Y. Qiao, "Df2net: A dense-fine-finer network for detailed 3d face reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2315–2324.

[22] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3d face shape and expression from an image without 3d supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7763–7772.

[23] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[24] F. Wu, L. Bao, Y. Chen, Y. Ling, Y. Song, S. Li, K. N. Ngan, and W. Liu, "Mvf-net: Multi-view 3d face morphable model regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 959–968.

[25] J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, and L. Quan, "Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency," *arXiv preprint arXiv:2007.12494*, 2020.

[26] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *2011 international conference on computer vision*. IEEE, 2011, pp. 1449–1456.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] Z. Li, S. Han, A. S. Khan, J. Cai, Z. Meng, J. O'Reilly, and Y. Tong, "Pooling map adaptation in convolutional neural network for facial expression recognition," in *ICME*, 2019.

[30] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," in *AAAI*, 2019.

[31] Y. Zhao, F. Tang, W. Dong, F. Huang, and X. Zhang, "Joint face alignment and segmentation via deep multi-task learning," *Multimedia Tools and Applications*, 2019.

[32] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018.

[33] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE SPL*, 2016.

[34] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, 1981.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[37] H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen *et al.*, "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview," in *ICPR*, 2016.

[38] S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1–10.

[39] N. K. Kasabov, *Time-space, spiking neural networks and brain-inspired artificial intelligence*. Springer, 2019.

[40] C. Tan, G. Ceballos, N. Kasabov, and N. Puthanmadam Subramaniyam, "Fusionsense: Emotion classification using feature fusion of multimodal data and deep learning in a brain-inspired spiking neural network," *Sensors*, vol. 20, no. 18, p. 5328, 2020.

[41] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas (me) ˆ2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, 2017.

[42] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE FG*, 2013.

[43] P. Hancock and B. Tiddeman, "Stirling/esrc 3D face database," 2011.

[44] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *CVPR*, 2020.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *DICTA*, 2017.

[47] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks," in *IJCNN*, 2019.