Micro-expression Action Unit Detection with Spatio-temporal Adaptive Pooling

Yante Li[1], Xiaohua Huang [1,2], Guoying Zhao[1]
[1] *Center for Machine Vision and Signal Analysis, Department of Computer Science and Engineering University of Oulu, Finland*
[2] *Nanjing Institute of Technology, China*
yante.li@oulu.fi, xiaohua.huang@live.cn, gyzhao@oulu.fi

***Abstract*—Action Unit (AU) detection plays an important role for facial expression recognition. To the best of our knowledge, there is little research about AU analysis for micro-expressions. In this paper, we focus on AU detection in micro-expressions. Micro-expression AU detection is challenging due to the small quantity of micro-expression databases, low intensity, short duration of facial muscle change, and class imbalance. In order to alleviate the problems, we propose a novel Spatio-Temporal Adaptive Pooling (STAP) network for AU detection in micro-expressions. Firstly, STAP is aggregated by a series of convolutional filters of different sizes. In this way, STAP can obtain multi-scale information on spatial and temporal domains. On the other hand, STAP contains less parameters, thus it has less computational cost and is suitable for micro-expression AU detection on very small databases. Furthermore, STAP module is designed to pool discriminative information for micro-expression AUs on spatial and temporal domains.Finally, Focal loss is employed to prevent the vast number of negatives from overwhelming the micro-expression AU detector. In experiments, we firstly polish the AU annotations on three commonly used databases. We conduct intensive experiments on three micro-expression databases, and provide several baseline results on micro-expression AU detection. The results show that our proposed approach outperforms the basic Inflated inception-v1 (I3D) in terms of an average of F1-score. We also evaluate the performance of our proposed method on cross-database protocol. It demonstrates that our proposed approach is feasible for cross-database micro-expression AU detection. Importantly, the results on three micro-expression databases and cross-database protocol provide extensive baseline results for future research on micro-expression AU detection.**

***Index Terms*—Micro-expression, Action unit detection, Neural networks, Video analysis**

I. Introduction

Emotion analysis is a meaningful and challenging task in our daily life. One of the most important ways to analyze emotion is through facial expressions. According to intensity and temporal change, facial expressions can be generally divided into two categories: macro-expression and micro-expression. Macro-expressions are the common and intentional expressions that we can see in our daily interactions typically last from 0.5s to 4s. Micro-expressions are involuntary and subtle facial muscle change which occurs within 0.5s [1], [2]. Micro-expressions can reveal people's hidden emotions and have many potential and emerging applications in different fields, such as clinical diagnosis, national security and interrogations [3].

Facial behavior consists of a set of Action Units (**AUs**) defined by Facial Action Coding System (**FACS**) [4]. FACS is a comprehensive, anatomically based system for describing all visually facial movements. In FACS, AUs are defined as the basic facial movements, which work as the building blocks to formulate multiple facial expressions [5]. Importantly, AU detection plays an indispensable role in analyzing complicated facial expressions [6], [7]. Currently, there are a number of researches on automatic AU detection of macro-expressions [8], [9]. Traditional AU detection approaches usually focus on geometric features based on facial landmarks or appearance features based on textures. In the recent years, various deep learning (**DL**) approaches have also been proposed to obtain discriminative facial representations and have achieved promising performance in AU detection of macro-expressions [6], [8]–[10].

TABLE I
Micro-expression databases (CASMEII [11], CASME [12] and SAMM [13])

	CASMEII	CASME	SAMM
Subjects	35	35	32
ME clips	247	195	159
FPS	200	60	200

However, to the best of our knowledge, there is little literature about AU analysis for micro-expressions. Compared with AU detection in macro-expressions, micro-expression AU detection becomes more challenging. Micro-expression AUs have much lower intensity and shorter duration. As observed in Table I, micro-expression databases have very small sample size, comparing with facial expression databases, such as BP4D [15] and DISFA [16]. For example, BP4D includes 328 videos (41 participants $\times$ 8 videos each, about 140,000 frames in total). For reducing the storage demands and processing time, each video consists of the most expressive segment (about one minute on average). This reduced the retention of frames in which little facial expression occurred [15]. So, the number of frames containing subtle AUs is also limited and it is not enough for micro-expression AU detection with static frames. Additionally, micro-expression AU detection suffers from the class imbalance, as shown in Table II. There is a big number gap between different AUs. For example, in CASMEII

TABLE II
AU DISTRIBUTION OF CASMEII, CASME AND SAMM DATABASES

database	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU14	AU15	AU17	AU18	AU20	AU24
CASMEII	26	21	129	2	13	58	13	17	34	27	16	25	1	1	2
CASME	23	17	69	0	1	4	40	3	9	23	14	13	11	2	4
SAMM	6	18	23	10	5	46	5	6	30	13	4	7	4	7	10

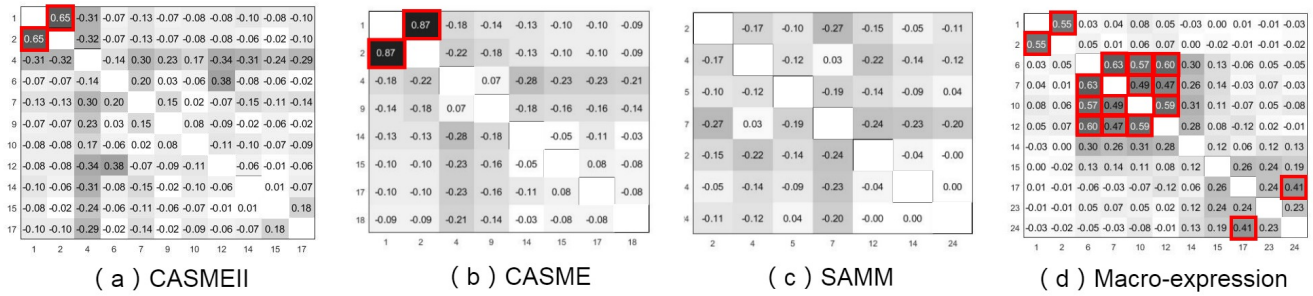

Fig. 1. Illustration of cross-correlation coefficients between different AUs. (a), (b) and (c) represent the AU cross-correlation coefficients in CASMEII, CASME and SAMM databases, respectively. Only common AUs with more than 10 samples are considered. (d) shows the macro-expression AU cross-correlation matrix studied on more than 350,000 valid frames with AU labels [14]. Red solid rectangles indicate the positive correlations between AUs.

database, most AUs just have tens of samples. However, AU4 has 129 samples and AU18 only occurs once. Even though multiple AUs could appear on one micro-expression, much less number of micro-expression AUs co-exist compared with macro-expression AUs. Figure 1 demonstrates cross-correlation coefficients between AUs. Following the explicit rules defined by [14], AUs with over moderate positive correlations (correlation coefficient ≥ 0.40) are viewed as positive correlations. As seen in Figure 1(a)(b)(c), there is only one pair of AUs (AU1 and AU2) appears many cases on the same time in CASME and CAMSEII databases. It makes it difficult to use the AU correlations for micro-expression AU detection.

In this paper, we concern the aforementioned problems of micro-expression AU detection, and present a new method to resolve them. Firstly, multiple AUs may occur on one micro-expression, thus, it should be analyzed whether to train single AU or multi-label AUs on very small micro-expression databases [5]. Since various deep learning approaches have been proved to be a powerful tool for video analysis and achieve good performance on macro-expression AU detection, we detect micro-expression AUs based on the deep structure. Micro-expression AUs could co-exist. It would be difficult to train good model for multi-label classification [17]. For example, in Table II, there are 11 common AU labels in CASMEII database, thus making the high-dimensional exponential label space, which involves 2^{11} combinations. As shown in Figure 1, there are rare AUs occurring simultaneously in micro-expressions. Therefore, it is difficult to use correlation between micro-expression AUs to detect AUs. It is reasonable to train single AU detection model through one-vs-rest protocol.

On the other hand, recent researches demonstrate that local sparse structure with multi-scale filters in convolutional neural network can obtain multi-scale discriminative information with a small number of parameters. For example, the Inception module (shown as Figure 3) [18] and Inflated Inception-v1 (**I3D**) [19] are proposed to analyze videos in action recognition. The main advantage of Inception moudule is a significant quality gain with less computational requirements, compared to traditional deep networks [20]. Due to its less parameters, Inception module is the good basic module for micro-expression AU detection on very small micro-expression databases [11], [12], [21]. As well, this kind of module can reduce the time and storage cost during the training procedure. Therefore, we employ the Inception modules to detect AUs.

Secondly, according to the definition of AUs [4], AUs become active on sparse facial regions. Recently, region adaption learning representations for regions of interest have been presented to improve the results of AU detection [5], [6], [14]. For example, Zhao et al. [14] and Li et al. [5] learned AUs on cropped individual regions of interest (**ROI**) centered at facial landmarks. However, these methods seriously depends on the performance of facial landmark detection. Instead, Deep Region and Mulit-label Learning method (**DRML**) [5] was proposed to induce important facial regions by exploiting a region layer, such that it can capture facial structure information for obtaining better AU detection result with subtle movements. These studies indicate that facial ROI play an important role to AU detection.

Furthermore, since micro-expression AUs have low intensity, small regions related with AUs are occurred, while the most of regions are not changed [22]. It means that we can possibly detect micro-expression AUs on these ROI. Multiple-

instance learning (**MIL**) [23] was successfully implemented to locate the action region with the most information [24]. Motivated by the concept of MIL, we propose to use an attention mechanism with a maximum pooling operation to get representative micro-expression AU features.

Thirdly, micro-expression AUs are subtle, thus, the spatial information is not discriminative enough for detecting micro-expression AUs. Recent researches for facial AU detection demonstrate that considering the temporal information of AUs can improve the performance for AU detection in macro-expressions compared to static images [5], [25]. Facial muscle in micro-expressions does not only has low intensity, but also quickly changes in short duration. Therefore, temporal information plays an important role for micro-expression AU detection. In deep learning architectures, such as I3D [19], averaging pooling / maximum pooling has been proposed to aggregate the temporal information, e.g. in action recognition. However, this sort of mechanisms seriously ignores the different weights of frames contributing in temporal domain. Recent study of micro-expressions in [26] found that micro-expression frames can have unequal contributions for micro-expression recognition. Motivated by this finding [26], we consider the different contributions of frames for micro-expression AU detection. Adaptive pooling is a pooling operation to fuse different information by the importance without any additional inference steps or extra computation [27]. This adaptive pooling processing has also been implied for efficient human action recognition in videos [28] to pool informative frames. Considering the advantage of adaptive pooling, we design an adaptive pooling module to pool discriminative information for AUs on spatial and temporal spaces.

Lastly, as shown in Table II, it is seen that in the micro-expression databases, micro-expression AU detection suffers from severe class imbalance problem. For example, in CASMEII database, the 'AU4' category has 129 samples, while 'AU9' just has 13 samples. It leads to that the size of positive samples are significantly less than negative ones in micro-expression AU detection. In imbalanced databases, convolutional neural network tends to be biased toward the majority class with poor accuracy for the target class [29]. The classical strategies resolving class imbalance problem for deep learning are re-sampling, cost-sensitive training or weighted loss [30]. Amongst these methods, Focal loss [31] has proved to be a more effective alternative compared with other approaches for this problem. Focal loss is a dynamically scaled cross entropy loss which can automatically down-weight the contribution of vast easy examples and focus the model on hard examples rapidly. Their work inspires us to use Focal loss to alleviate class imbalance issue of micro-expression AU detection.

In this work, we propose a new deep spatio-temporal adaptive pooling network with Focal loss for micro-expression AU detection. The architecture is shown in Figure 2. Different from traditional CNN frameworks which consist of convolutional layers followed by fully connected layers, our work is primarily consisted of Inception modules which contain multi-scale convolutional filters. In this way, we can obtain the multi-scale spatial and temporal information of micro-expression AUs with effective computational cost. The attention based adaptive pooling module can pool discriminative and informative representation for micro-expression AU detection. Specifically, for spatial domain, we detect the AU based on the ROI (the region including AUs) with a maximum pooling operation, while for temporal domain, we employ a softmax weighted pooling to adjust the temporal weight for aggregating the temporal information. Lastly, we use the Focal loss [31] to resolve the classification problem of category imbalance.

Our main contributions can be concluded as followed:

- To the best of our knowledge, this is the first work to detect micro-expression Action Units. To alleviate the problems of micro-expression AU detection , we propose a new spatio-temporal adaptive pooling network with Focal loss to efficiently pool discriminative and informative frame regions. According to intensive experiments, our proposed network has been demonstrated to achieve the promising performance, comparing with hand-crafted features and the state-of-the-art I3D.
- In this paper, we provide the baseline results of AU detection in micro-expressions on three publicly available micro-expression databases. Different from the existing works in micro-expression recognition, we polish the AU annotations of CASME, CASMEII and SAMM databases and make the intensive experiments on these AU annotations. The polished annotation will provide basic information for further study in micro-expression analysis, especially in exploring the relationship between micro-expression AU and micro-expression emotion category.
- Besides the experiments on intra-database, we conduct the experiments by using our proposed method and one hand-crafted feature on cross-database micro-expression AU detection. It is also the first time to conduct this kind of study in micro-expression AU detection. The experiments show that our approach can achieve promising performance on cross-database micro-expression AU detection.

II. Spatio-temporal Adaptive Pooling Network with Focal Loss

A common CNN structure for macro-expression AU detection is designed as region-based feature extraction followed by multiple fully connected (**FC**) layers and multi-sigmoid cross entropy loss to recognize multiple AUs [5], [6], [14]. Especially, FC layer needs lots of parameters and multi-sigmoid cross entropy is based on the predicted results for the multiple AUs, which makes it prone to extreme loss explode [5]. However, the micro-expression AU refers to subtle intensity and quickly change. As well, the micro-expression databases are very small, and with severe class imbalance, it makes it far away from satisfying requirements of FC and multi-sigmoid cross entropy. To solve these problems, we construct a Spatio-Temporal Adaptive Pooling network (**STAP**) with Focal loss for micro-expression AU detection.

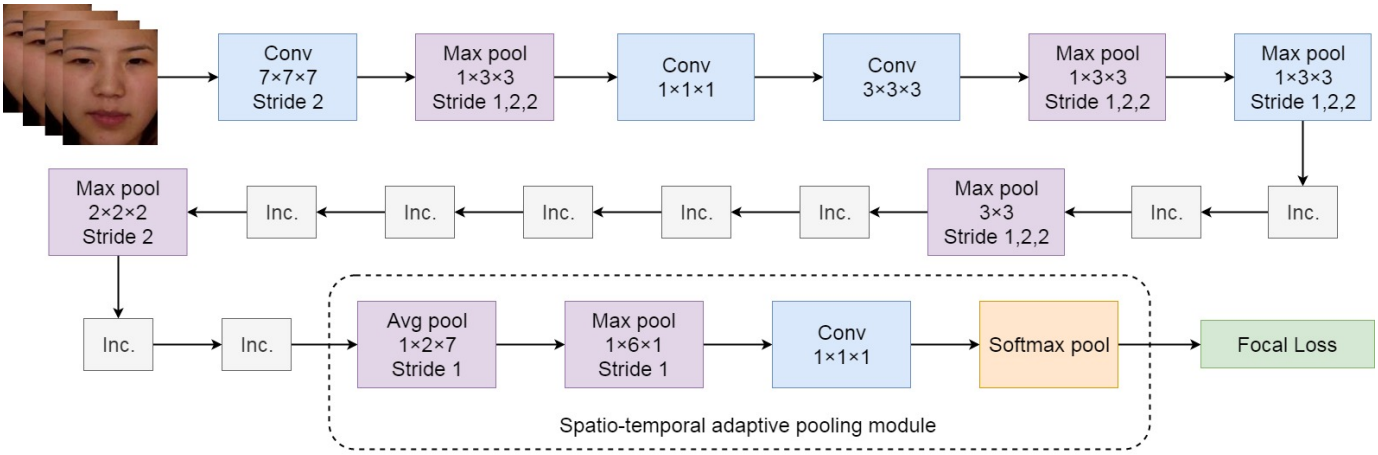

Fig. 2. Framework of the proposed Spatio-temporal adaptive pooling neural network with Focal loss (Inc. represents the Inception module).

The outline of the proposed STAP network architecture is shown as Figure 2. This network consists of three important components: micro-expression AU multi-label detection based on Inception modules, spatio-temporal adaptive pooling module and Focal loss. In this section, we firstly discuss multi-label AU detection, and then illustrate the effectiveness of spatio-temporal adaptive pooling module. Finally, we describe the Focal loss which impedes the mis-detection caused by AU imbalance.

A. Micro-expression AU Multi-label Detection based on Inception Modules

The principle of designing the basic network is inspired by the Inflated Inception-V1 (I3D) network for action detection with Inception modules [18], [19]. Unlike conventional CNN frameworks [32], where the same convolutional filters are shared within the same layer, the Inception module [18], [19] aggregates 2 different sizes of filters ($1 \times 1 \times 1, 3 \times 3 \times 3$) to compute multi-scale spatial information and assembles $1 \times 1 \times 1$ convolutional filters for dimension and parameter reduction, shown in Figure 3.

As shown in Table II, for most AUs, there are less than 40 samples. The limited quantity cannot guarantee the convergency when training multiple AUs detection models simultaneously for each micro-expression. Moreover, the relationship information between different micro-expression AUs is not strong enough to identify multiple AUs [33]. In our paper, the one-vs-rest protocol is used to transform the multi-label micro-expression AU detection to multiple binary classification problem. For each AU detection, the samples of target AU are viewed as positive samples and all other samples as negatives.

The shallow layers of STAP network are consisted of two convolutional layers with two max-pooling operation and 9 Inception modules following the I3D structure. The input is N aligned RGB micro-expression sequential frames. The output feature maps of the last Inception module corresponds to facial micro-expression spatial structure. Different from I3D which average pools all the feature maps leading to information lose, we design spatio-temporal adaptive pooling module to pool discriminative information for micro-expression AUs. Finally, the Focal loss is employed to train the imbalanced micro-expression AUs.

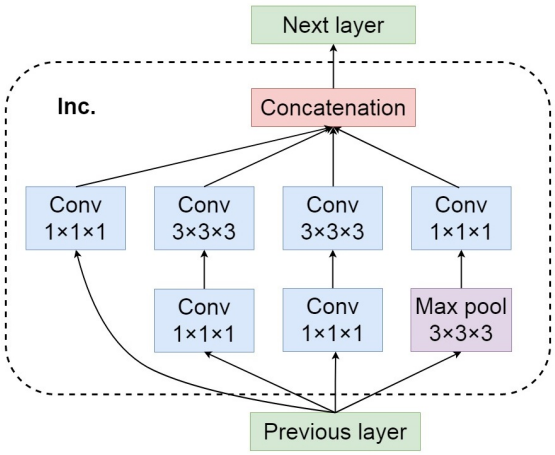

Fig. 3. Inception module [19]

B. Spatio-temporal Adaptive Pooling Module

One key aspect of STAP network is to design a spatio-temporal adaptive pooling module (shown in Figure 4), which can capture the discriminative local appearance change which is subtle and quick, according to the micro-expression AU regions of interest in spatio-temporal domain.

Most of deep learning methods utilize standard convolutional layers shared over the image. However, for structured face images, the spatial stationarity assumption can not hold. So, different facial regions abide by different local statistics and have different contribution for AU detection. Zhao et.al [6] proposed a region layer for learning weights to capture the information from the important patches. The region layer

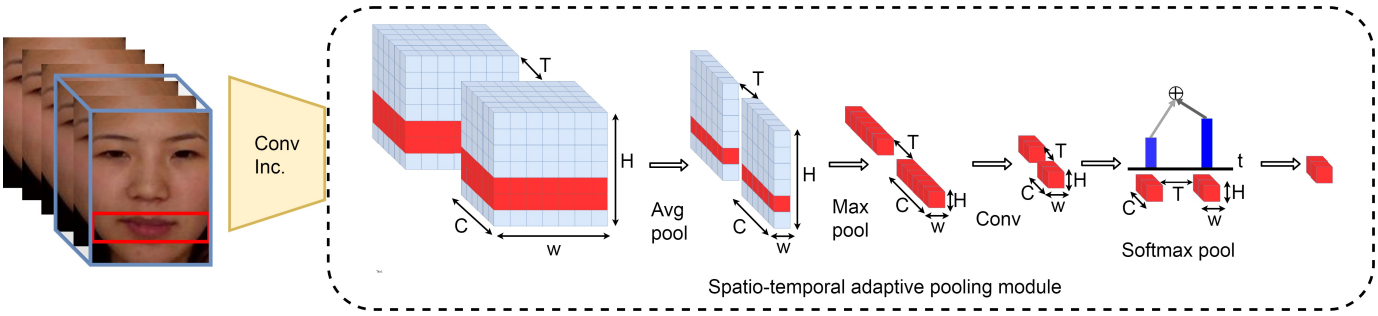

Fig. 4. The illustration of spatio-temporal adaptive pooling module. The input of spatio-temporal adaptive pooling module is 4D feature maps ($T \times H \times W \times C$) which are obtained through 3D convolutional layers and Inception modules. They are viewed as T 3D features maps ($H \times W \times C$) along temporal dimension. T, H, W and C represent the time, height, wide, channel of the feature maps, respectively. The red solid rectangle indicates the ROI in micro-expressions.

can obtain better performance, especially for subtle AUs, compared with other CNN approaches [14]. Motivated by above observations [6], we firstly design a spatial adaptive pooling operation to learn the micro-expression AUs information from the region of interest. As one micro-expression AU is detected for each training process and face is symmetrical, AUs lie in only one or two facial regions. How to decide the region containing information about micro-expression AUs is a problem. Multiple instance learning is a kind of weakly supervised leaning method which can train classifiers with a maximum operation on bag-level labeled data [34]. For object detection task, MIL can realize object detection in a series of candidate region boxes [35]. For micro-expression AU detection, the face can also be divided into several equal candidate regions, and identify AU based on the ROI through a maximum pooling operation. These findings encourage us to design a multi-instance learning scheme to pool AU features on the ROI automatically.

On the other hand, unlike AUs in macro-expressions, micro-expression AU changes in very short duration. The temporal information plays an important role for micro-expressions. A popular framework for extracting temporal information in videos is using a temporal pooling operation to squeeze different frame information to a summary vector, such as average and max pooling [36], [37]. Compared with other temporal information fusion solutions like Latent Variable Models [38] or RNN [39], temporal pooling methods do not need additional inference steps during learning or intermediate hidden states.

But, the normally used pooling approaches usually consider that all frames in the video make the equal contribution. Recently, research in micro-expression recognition [40] shows that micro-expressions recorded with high-speed camera contain redundancy information. Therefore, not all frames have useful information. Some works on micro-expression recognition [26] further demonstrates that the method using just the onset, apex and offset frames can also achieve promising performance as the approaches based on video sequence. These state-of-the-art works indicate that different frames in micro-expression make different contribution to micro-expression recognition. Therefore, simple averaging all frames or choosing the maximum frame (Apex frame) is not appropriate. Motivated by [26], [40], STAP adds adaptive weights on the micro-expression frames by using softmax weighted average pooling operation. In this way, the micro-expression AU temporal information can be fused efficiently without extra learning cost. In STAP framework, the output of the last Inception module is a sequence of 1024 7×7 feature maps maintaining the ME spatial and temporal information. Following the face structure, every two rows of the feature map can be viewed as a ROI candidate. The average pooling with $1 \times 2 \times 7$ kernel and stride 1 is employed to pool features representing ROI change between frames. In this way, we do not need to detect ROI regions by facial landmark in advance. Then a max pooling is followed to pool the feature of ROI with a maximum operator over $1 \times 6 \times 1$. As so far, we obtain 1024D feature sequences to represent the micro-expression AU information along the temporal domain. Unlike common CNNs employing multiple FC layers to compute the category information, STAP uses a $1 \times 1 \times 1$ convolutional filter to convolve the 1024D micro-expression AUs features to 2D vectors $\hat{P}_t = [\hat{p_1}, \hat{p_1}, ..., \hat{p_t}]$, corresponding to AU classes. The flowchat is shown in Figure 4.

Finally, soft-max weighed pooling [27] is employed to fuse the AUs information on time dimension. The equation of soft-max weighed average pooling is defined as followed:

$$\hat{P}_t = \sum_{t \in T} \hat{p}(t) \left(\frac{\exp(\hat{p}(t))}{\sum_{u \in T} \exp(\hat{p}(u))} \right) \tag{1}$$

where $\hat{p}(t)$ represents the 2D feature vector corresponding to AU classes probability. This operation can increase the weight of important temporal information.

C. Focal loss for Data Imbalance

In this paper, our STAP network simplifies the multi-label AU detection problem to multiple binary classification problems with one-vs-rest protocol. However, the AU distribution is imbalanced. Especially, in our framework, we treat the target AU as positive samples and all other AUs as negative samples for each AU detection. The number of negative samples is much larger than the positive samples which lead to bias information learning and misclassification.

In order to detect small amount of subtle AUs in micro-expression databases, the Focal loss focusing training on a sparse set of hard examples is employed, defined as followed:

$$FL(p_t) = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot log(p_t), \quad (2)$$

where the p_t represents micro-expression AU probability. The relative loss for correctly classified instances can be reduced by setting $\gamma > 0$ and the focus is training hard, easily misclassified samples. α_t is the factor to adjust the ratio of positive and negative quantity.

III. Experiments

In this section, we present the results of LBP-TOP [41], LPQ-TOP [42], I3D [19], I3D with Focal loss (**I3DF**) and our proposed method STAP, STAP with Focal loss (**STAPF**) in micro-expression AU detection and the result analysis. Firstly, we introduce the settings of our experiments. Then, we demonstrate and discuss the results of baseline algorithms, STAP network and other comparing approaches. Finally, we intensively discuss the results of cross-database training with our proposed STAP network.

A. Settings

Database and annotation. We evaluated STAP network on three spontaneous micro-expression databases: CASMEII, CASME and SAMM. In our experiments, we only consider the common AUs with more than 10 samples in each micro-expression database.

CASMEII consists of 247 micro-expression elicited from 26 participants with a camera of 200 fps. It includes the cropped face with high resolution (640×480 pixels). We used 243 videos with common 11 AUs: 'AU1', 'AU2', 'AU4', 'AU5', 'AU7', 'AU9', 'AU10', 'AU12', 'AU14', 'AU15', 'AU17' occurred in disgust, happy, surprise, angry and others emotions.

CASME database is spontaneous micro-expression clips including frames from onset to offset. It contains 195 spontaneous micro-expression clips from 19 subjects and recorded with frame rate of 60 fps. There are 171 samples related with four common emotions: disgust, repression, surprise and tense, including 8 AUs: 'AU1', 'AU2', 'AU4', 'AU9', 'AU14', 'AU15' and 'AU17'.

SAMM collects 159 micro-expression samples from 32 participants with 13 ethnicities with recording rate 200 fps. 'AU2', 'AU4', 'AU5', 'AU7', 'AU12', 'AU14' and 'AU24' were evaluated in our experiments.

Metrics. AU detection is a multi-label binary classification problem. For a binary classification task especially when samples are not balanced, F1-score can better interpret the algorithm performance better. In our evaluation, F1-scores are computed for 11 AUs in CASMEII, 8 AUs in CASME and 7 AUs in SAMM according to the AU samples quantity and importance. The overall performance of the algorithm is described by the average F1-score.

Implementation. In our experiments, we all employ the cropped face images provided by the databases. The input is aligned RGB micro-expression sequential images, which are interpolated into 10 through the Temporal interpolation model [43]. All models were pre-trained on ImageNet [44]. During training, the learning rate is set as 0.01. A momentum of 0.9 and drop out of 0.5 was used. All implementations were based on the Tensorflow.

Following common experimental settings for AU detection, we use two-fold cross validation. The databases are splited to two folds based on subject IDs. Each time one fold is used for training. The other fold is divided to two sub-folds for testing and validation. For Linear SVM classifiers, we use one fold to train and the other fold to test for each time.

TABLE III
F1-score (%) on CASMEII database. Bold numbers indicate our proposed methods; Underlined numbers indicate the best performance.

AU	LBP-TOP	LPQ-TOP	I3D	I3DF	STAP	STAPF
1	41.55	31.02	73.40	76.19	**71.90**	**73.01**
2	43.71	41.67	67.31	67.31	**67.31**	**75.64**
4	86.44	83.51	94.22	96.85	**96.48**	**95.71**
6	16.67	0.00	86.36	86.36	**83.33**	**86.36**
7	11.44	20.52	85.58	88.93	**87.5**	**84.86**
9	0.00	0.00	73.33	81.67	**91.67**	**90.00**
10	0.00	0.00	86.36	92.86	**82.86**	**74.78**
12	40.91	36.37	79.49	83.71	**83.71**	**78.88**
14	39.27	36.37	79.39	83.33	**77.50**	**87.50**
15	42.86	45.19	79.63	78.57	**79.63**	**89.29**
17	17.65	26.32	100.0	96.67	**96.67**	**100.0**
Avg	29.76	29.23	82.30	84.77	**83.51**	**85.09**

B. Results

Tables 3, 4 and 5 show the results of 11 AUs for CASMEII, 8 AUs for CASME and 7 AUs for SAMM, respectively. The results are discussed from six perspectives: handcrafted vs. deep features, multi-label vs. single label learning, spatio-temporal adaptive pooling vs. average pooling, data imbalance, cross databases and computational and storage cost.

Handcrafted vs. deep features. This is the first work for micro-expression AU detection. We provide the baseline of the micro-expression AU detection. For the baseline, the frames are divided to 5×5 blocks. LBP-TOP features [41] and LPQ-TOP features [42] are extracted on blocks as handcrafted features. For LBP-TOP, the radii were set to (3,3,3). One-vs-rest Linear SVM ($c = 1000$) is used to train a single classifier per AU class. LBP-TOP achieves better results compared to LPQ-TOP. LBP-TOP is chosen as the baseline of micro-expression AU detecion. The baselines of CASMEII, CASME and SAMM databases based on LBP-TOP are 29.76%, 24.41% and 8.62%, respectively. Some of the AU F1-scores are 0.00%, e.g. AU9 and AU10 in CASMEII. This is caused by the class imbalance. Comparing the results of LBP-TOP, LPQ-TOP and deep learning methods (I3D, STAP and STAPF) in Table IV on CASME database, all of 8 AUs in F1-score are higher for DL methods. The improvements of DL are larger on CASMEII and SAMM databases by showing in Tables III and V, respectively. It illustrates that DL features has generalizability on different databases. Amongst all the DL methods, our STAPF has the best performance on all three databases. As shown in

Tables III, IV and V, STAPF obtains about 55%, 36% and 49% improvement in terms of average F1-score on CASMEII, CASME and SAMM databases, respectively. It is worth to notice that the feature dimensions for Linear SVM with LBP-TOP, LPQ-TOP, I3D, STAP and STAPF network are 4425, 19200, 1024, 1024 and 1024, respectively. We can see that the performance of I3D, STAP and STAPF network surpass handcrafted features, such as LBP-TOP and LPQ-TOP, and also the learning features of I3D, STAP and STAPF network have lower dimension than the engineered features. We can infer that the learned features can capture more discriminative characteristics for micro-expression AU detection, compared to the handcrafted features.

Multi-label vs. single label learning. Multi-label learning could improve AU detection for macro-expressions through taking AU correlations into account [6]. In our experiments, we also try to realize multi-label AU detection of micro-expressions with multi-label sigmoid cross-entropy loss, shown by Equation 3, which considers the correlations between AUs.

$$Loss = -\sum (l \cdot log(p) + (1-l) \cdot log(1-p)), \quad (3)$$

where l and p represent the ground truth and score for one class, respectively. $1-l$ and $1-p$ represent for other classes, respectively. In our implementation, we see that the loss function of [6] cannot converge well in micro-expression AU detection. The training accuracy just can achieve 55%. It may be explained by that the micro-expression databases are much smaller compared with macro-expressions databases. For macro-expression AUs in BP4D database, the 10 multi-AU detection model is trained on more than 140,000 face images. However, for CASMEII, it only contains 247 micro-expression clips with 17,101 frames. That is too difficult to identify 11 AUs simultaneously with so little data, especially for the subtle and quick micro-expression movements. The number of subtle AUs in macro-expression databases [15] is also small which is not enough for micro-expression AU detection on static frames. As for the limited quantity, it is also difficult to dig statistical relationship between different AUs. The I3D, STAP and STAPF network results in Tables III, IV and V show that approaches with one-vs-rest protocol can get better performance on CASMEII, CASME and SAMM databases. The average F1-score of STAPF even achieves more than 85.09% on CASMEII database.

Spatio-temporal adaptive vs. average pooling. Here we discuss the effectiveness of the spatio-temporal adaptive pooling module. The shallow layers of I3D, STAP and STAPF network share the same structure. However, I3D uses average pooling for final classification. Observing the results of I3D, STAP in Tables III, IV and V, we can find that spatio-temporal adaptive pooling module improves the average F1-score by 1.21%, 0.28% and 1.95% on CASMEII, CASME and SAMM databases compared with average pooling. It validates the observation that STAP module learns more discriminative information than average pooling. The results show that the proposed STAP module helps micro-expression AU detection by fusing the spatial and temporal information effectively.

TABLE IV
F1-SCORE (%) ON CASME DATABASE. BOLD NUMBERS INDICATE OUR PROPOSED METHODS; UNDERLINED NUMBERS INDICATE THE BEST PERFORMANCE.

AU	LBP-TOP	LPQ-TOP	I3D	I3DF	STAP	STAPF
1	62.50	43.64	66.67	66.67	**40.88**	**60.00**
2	38.33	23.81	66.67	55.56	**49.05**	**63.77**
4	47.07	56.85	49.75	58.83	**71.15**	**62.19**
9	0.00	2.56	32.70	56.94	**56.94**	**56.94**
14	21.98	0.00	52.03	50.71	**52.03**	**50.71**
15	11.11	22.62	64.10	64.10	**64.10**	**64.10**
17	0.00	0.00	62.50	62.50	**62.50**	**62.50**
18	14.29	0.00	63.33	63.33	**63.33**	**63.33**
Avg	24.41	18.68	57.22	59.83	**57.50**	**60.44**

TABLE V
F1-SCORE (%) ON SAMM DATABASE. BOLD NUMBERS INDICATE OUR PROPOSED METHODS; UNDERLINED NUMBERS INDICATE THE BEST PERFORMANCE.

AU	LBP-TOP	LPQ-TOP	I3D	I3DF	STAP	STAPF
2	19.44	32.16	59.34	64.10	**64.10**	**59.34**
4	4.76	0.00	51.98	51.98	**58.56**	**51.98**
5	14.29	30.95	66.67	66.67	**66.67**	**66.67**
7	39.97	39.31	43.96	35.56	**48.26**	**46.89**
12	7.69	22.18	30.39	41.73	**28.38**	**46.96**
14	0.00	0.00	65.15	65.15	**65.15**	**65.15**
24	0.00	0.00	64.10	64.10	**64.10**	**64.10**
Avg	8.62	14.72	54.51	55.61	**56.46**	**57.30**

Data imbalance. As the micro-expression AUs are difficult to learn and imbalanced, the Focal loss is employed to learn the hard micro-expression AU information and avoid bias learning, compared to common softmax cross entropy loss. From Table III, we can see that I3D with Focal loss (I3DF) reached higher F1-score in 7 out of 11 AUs and STAPF can improve the average F1-score by 1.58% in comparision wiht I3D and STAP, respectively. In CASME database (Table IV), I3DF and STAPF outperformed I3D and STAP on F1-score by 2.61% and 2.94%, respectively. In SAMM database (Table V), on average, I3DF outperformed I3D with 1.10% and STAPF outperformed STAP with 0.84% higher in terms of F1-score. The results justify that Focal loss can learn the hard micro-expression AU information from imbalanced AU samples.

Cross databases. We evaluate our STAPF on cross-database scenarios. For each run, we choose one database for training, one for validation and the other for test. Considering the AU distribution of the three databases, AU1, AU2, AU4, AU9, AU12, AU14 and AU17 are chosen to be evaluated. In this way, the CASMEII, CASME and SAMM database contains 228, 154 and 78 samples, respectively. The results are shown in Table VII. STAPF achieves promising performance on cross-database scenarios. The F1-score of CASMEII trained on CASME database even obtains 63.91%. The F1-score of CASME trained on SAMM database is the worst which is still more than 43%. In general,the results training on CASMEII and CASME databases are better than those from training on SAMM database. That is because the size of CASMEII

TABLE VI
F1-SCORE (%) OF CROSS-DATABASE MICRO-EXPRESSION ACTION UNIT DETECTION WITH LBP-TOP (BASELINE).

Train Data	CASMEII	CASMEII	CASME	CASME	SAMM	SAMM
Test Data	CASME	SAMM	CASMEII	SAMM	CASME	CASMEII
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00
4	62.50	44.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00
Avg	8.90	6.29	0.00	0.00	0.00	0.00

TABLE VII
F1-SCORE (%) OF CROSS-DATABASE MICRO-EXPRESSION ACTION UNIT DETECTION WITH STAPF.

Train Data	CASMEII	CASMEII	CASME	CASME	SAMM	SAMM
Test Data	CASME	SAMM	CASMEII	SAMM	CASME	CASMEII
1	63.83	60.00	42.86	66.67	61.22	50.00
2	66.67	59.57	63.16	60.87	53.57	41.67
4	62.18	54.55	65.79	30.59	45.28	41.67
9	49.35	54.55	66.67	40.00	49.35	66.67
12	25.80	45.28	66.67	45.28	6.67	53.84
14	51.85	64.86	68.57	52.63	40.00	62.50
17	61.11	72.72	73.68	66.67	50.00	58.82
Avg	54.40	58.79	63.91	51.82	43.72	53.60

and CASME database is larger than the size of SAMM database. In addition, the CASMEII and CASME databases only contain subjects from Asia, while the SAMM database includes subjects from various ethnicities. The baseline results with LBP-TOP on three databases were shown in Table VI. In most cases, the one-vs-rest Linear SVM classifies all the AU samples to the negative class with large sample size, and leads to 0.00% in terms of F1-score. Compared with the baseline results, we can conclude that the STAPF is feasible for cross-database micro-expression AU detection. The database size and diversity has influence on the detection performance. The larger database and unified appearance can improve the micro-expression AU detection accuracy.

Computational and storage cost. The STAP network experiments were running on a NVIDIA Tesla K80c GPU with 12 GB memory. For each step in the training process, we evaluated the running time for each iteration. When the batch size is set as 6. The average training time of each step is 0.92 sec. As the micro-expression databases are small. The number of training step for each database is no more than 1000, which means each model can be finished in 920 sec (15.3 min). The size of parameter for STAPF network model is about 32.4M on average. Although we train one model for each micro-expression AU, the computational and storage cost for each model is rather small.

IV. CONCLUSION

Micro-expression AU detection is an important and challenging task, as the micro-expressions have samll databases, subtle and quick facial muscle change and data imbalance. In this paper, we propose deep Spatio-temporal Adaptive Pooling network (STAP) with Focal loss for micro-expression AU detection. STAP is an end-to-end trainable network and able to identify subtle and quick micro-expression AUs on specific regions with efficient temporal information. To this end, we introduce a spatio-temporal adaptive pooling module to capture discriminative AU information based on regions of interest and weighted temporal information fusion. Furthermore, Focal loss is employed to solve the mis-detection caused by data imbalance. To the best of our knowledge, this is the first work concentrating on micro-expression AU detection. We presented the baseline resluts of micro-expression AU detection. The re-organized annotations of CASME, CASMEII and SAMM databases are going to be released for future study. Experiments conducted on within- and cross-database scenarios demonstrate the effectiveness of STAP network with focal loss. The future work includes detecting multi-label micro-expression AUs by multi-task learning. The proposed STAP module introduces potential applications to general facial AU detection and facial expression recognition.

REFERENCES

[1] Paul Ekman and Wallace V Friesen, "Nonverbal leakage and clues to deception," *Study Interpers*, vol. 32, pp. 88–106, 1969.

[2] Paul Ekman, "Lie catching and microexpressions," *Phil. Decept.*, pp. 118–133, 2009.

[3] Paul Ekman and Wallace V Friesen, "Constants across cultures in the face and emotion," *Personal. Soc. Psychol.*, vol. 17, no. 2, 1971.

[4] Wallace V Friesen and Paul Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[5] Li Wei, Abtahi Farnaz, and Zhu Zhigang, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1841–1850.

[6] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3391–3399.

[7] Wiggers Michiel, "Judgments of facial expressions of emotion predicted from facial behavior," *Journal of Nonverbal Behavior*, vol. 7, no. 2, pp. 101–116, 1982.

[8] Kaili Zhao, Wen-Sheng Chu, and Aleix Martinez, "Learning facial action units from web images with scalable weakly supervised clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2090–2099.

[9] Shizhong Han, Zibo Meng, Zhiyuan Li, O'Reilly James, Jie Cai, Xiaofeng Wang, and Yan Tong, "Optimizing filter size in convolutional neural networks for facial action unit recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5070–5078.

[10] Yong Zhang, Weiming Dong, Bao Gang Hu, and Qiang Ji, "Classifier learning with prior probabilities for facial action unit recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5108–5116.

[11] WenJing Yan, Xiaobai Li, SuJing Wang, Guoying Zhao, YongJin Liu, YuHsin Chen, and Xiaolan Fu, "Casme ii: an improved spontaneous micro-expression database and the baseline evaluation," *Plos One*, vol. 9, no. 1, pp. e86041, 2014.

[12] WenJing Yan, Qi Wu, YongJin Liu, SuJing Wang, and Xiaolan Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *Automatic Face and Gesture Recognition*, 2013, pp. 1–7.

[13] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2018.

[14] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang, "Joint patch and multi-label learning for facial action unit and holistic expression recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3931–3946, 2016.

[15] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[16] Mavadati S Mohammad, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[17] Wenjie Zhang, Junchi Yan, Xiangfeng Wang, and Hongyuan Zha, "Deep extreme multi-label learning," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, 2018, pp. 100–107.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[19] Carreira Joao and Zisserman Andrew, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[20] Parkhi M. Omkar, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, pp. 41.1–41.12.

[21] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–6.

[22] A Freitas-magalhães, *The Face of Psychopath-Brain and Emotion*, Leya, 2018.

[23] Cha Zhang, John C Platt, and Paul A Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2006, pp. 1417–1424.

[24] Georgia Gkioxari, Ross Girshick, and Jitendra Malik, "Contextual action recognition with r* cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1080–1088.

[25] Michel Valstar, Maja Pantic, and Ioannis Patras, "Motion history for facial action detection in video," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*. IEEE, 2004, vol. 1, pp. 635–640.

[26] Sze Teng Liong, John See, Kok Sheik Wong, and Raphael C. W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[27] Brian McFee, Justin Salamon, and Juan Pablo Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 11, pp. 2180–2193, 2018.

[28] Amlan Kar, Nishant Rai, Karan Sikka, and Gaurav Sharma, "Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3376–3385.

[29] Haibo He and Edwardo A Garcia, "Learning from imbalanced data ieee transactions on knowledge and data engineering v. 21 n. 9," 2009.

[30] Chen Huang, Yining Li, Change Loy Chen, and XiaoouTang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[31] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 2999–3007, 2017.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Ji Qiang, "Capturing global semantic relationships for facial action unit recognition," in *IEEE International Conference on Computer Vision*, 2014.

[34] Songhe Feng, Congyan Lang, and De Xu, "Beyond tag relevance: integrating visual attention model and multi-instance learning for tag saliency ranking," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 288–295.

[35] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell, "On learning to localize objects with minimal supervision," *arXiv preprint arXiv:1403.1024*, 2014.

[36] Naila Murray and Florent Perronnin, "Generalized max pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2473–2480.

[37] Chen-Yu Lee and Patrick W Gallagherand Zhuowen Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," in *Artificial intelligence and statistics*, 2016, pp. 464–472.

[38] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid, "Activity representation with motion hierarchies," *International journal of computer vision*, vol. 107, no. 3, pp. 219–238, 2014.

[39] Ng Joe Yue-Hei, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[40] Yante Li, Xiaohua Huang, and Guoying Zhao, "Can micro-expression be recognized based on single apex frame?," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3094–3098.

[41] Guoying Zhao and Matti Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[42] Juhani Päivärinta, Esa Rahtu, and Janne Heikkilä, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Scandinavian Conference on Image Analysis*. Springer, 2011, pp. 360–369.

[43] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen, "Towards a practical lipreading system," in *Proceedings of the IEEE conference on computer vision and pattern recognition 2011*. IEEE, 2011, pp. 137–144.

[44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.