SHALLOW OPTICAL FLOW THREE-STREAM CNN FOR MACRO- AND MICRO-EXPRESSION SPOTTING FROM LONG VIDEOS

Gen-Bing Liong[†] *John See*[‡⋆] *Lai-Kuan Wong*[†]

[†] Faculty of Computing and Informatics, Multimedia University, Malaysia
[‡] School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia
(⋆ Corresponding author)

ABSTRACT

Facial expressions vary from the visible to the subtle. In recent years, the analysis of micro-expressions— a natural occurrence resulting from the suppression of one's true emotions, has drawn the attention of researchers with a broad range of potential applications. However, spotting micro-expressions in long videos becomes increasingly challenging when intertwined with normal or macro-expressions. In this paper, we propose a shallow optical flow three-stream CNN (SOFTNet) model to predict a score that captures the likelihood of a frame being in an expression interval. By fashioning the spotting task as a regression problem, we introduce pseudo-labeling to facilitate the learning process. We demonstrate the efficacy and efficiency of the proposed approach on the recent MEGC 2020 benchmark, where state-of-the-art performance is achieved on CAS(ME)^2 with equally promising results on SAMM Long Videos.

Index Terms— Micro-expression, macro-expression, spotting, optical flow, shallow CNN

1. INTRODUCTION

In most naturalistic scenarios, spontaneous facial expressions could occur at varying degrees of intensities and brevity – from the visible to the subtle. These occurrences of *macro-expressions* and *micro-expressions* could coexist or occur in isolation. Micro-expressions, which typically lasts between 1/25 to 1/5 second at rather low intensities, occur when a person attempts to conceal his or her genuine emotions in a high-stake situation [1]. On the other hand, macro-expressions are easier to identify even without proper training as the duration is longer at higher intensities. Recent advances in deep learning have witnessed a widespread popularity in the recognition task while efforts in the spotting task, especially on long "untrimmed" videos, remain subdued [2]. As such, the micro-expression community has recently organized the 3rd MEGC Workshop (MEGC2020) [3] to challenge researchers towards spotting macro- and micro-expression in long videos. Generally, facial expressions undergo three distinct phases: *onset*, *apex*, and *offset*. As accurately described in [4], onset occurs when facial muscles begin contracting; apex is the phase where the facial action is at its peak intensity; offset signifies the muscles going back to neutral state. This paper highlights the task of spotting both macro- and micro-expression sequences, *i.e.* from onset to offset.

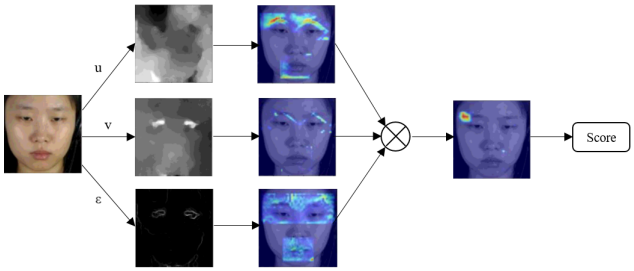

Fig. 1: Intuitively, the optical flow components used in each of the three streams capture different salient motion information for uncovering macro- and micro-expressions.

Early works by [5] and [6] notably laid the fundamental mechanism of the task; the latter in particular employed LBP as the feature descriptor with χ^2-distance used for feature difference (FD) analysis between two frames in a fixed duration. The micro-expression is determined if the frame's feature vector is above the threshold set for peak detection. Most works utilise established pre-processing techniques involving landmark detection [7, 8], region masking [5, 9], and emphasis on specific facial regions via ROI selection [10, 11, 12].

Motion-based approaches can characterize the subtle movements on the face. Shreve et al. [5] first introduced optical strain (a derivative of optical flow) to analyze subtle motion changes based on the elastic deformation of facial skin tissue. The amount of strain observed across time (by summing its magnitudes) at different facial regions is considered. The baseline method for MEGC2020, MDMD [13] encodes the maximal difference magnitude along the main direction of motion to predict if a macro- or micro-expression is present. Meanwhile, [10] constructed optical strain features for apex spotting. More recently, a few works [14, 12] have begun to adopt deep learning methods for spotting. [14] experimented with CNN and RNN models under an alternative benchmarking strategy while [12] fed pre-computed HOOF features

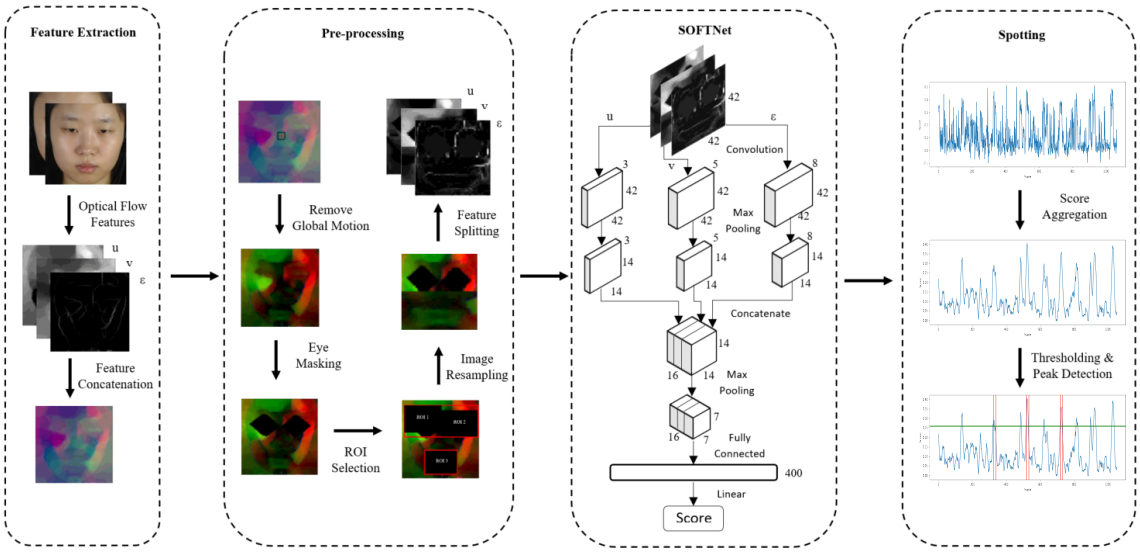

Fig. 2: Framework of the proposed approach and its four phases.

into an RNN which is fashioned to spot short intervals with likely micro-movements. To address issues such as shortage of micro-expression samples and over-complex models, [15] proposed to use shallow convolutional networks with multiple streams of input information. Their work however, was catered for the recognition task.

Inspired by [15], we hypothesize that such kind of models can be trained to alleviate the insufficiency of data, but concurrently harnessing the benefits of motion information. To achieve this, we fashion the spotting task as a *regression* problem that predicts how likely a frame belongs to a macro- or micro-expression. In the core, we build a shallow optical flow three-stream CNN (SOFTNet) to capture the relevant features from different optical flow components. The contributions of this paper are summarized as follows:

1. We propose a multi-stream shallow network infused with optical flow inputs to regress a score for spotting.
2. We present a new automatic way of pseudo-labeling frames to enable the training of a regression model.
3. We demonstrate the efficacy of the proposed approach in terms of F1-score and computational time on the MEGC2020 benchmark, achieving state-of-the-art results on CAS(ME)^2.
4. We re-explore the viability of a detection metric which provides a fairer and more consistent measure for locating both macro- and micro-expression occurrences.

2. PROPOSED FRAMEWORK

The proposed framework is illustrated in Figure 2. This section discusses the four phases of the framework: initial feature extraction of optical flow components, a series of pre-processing steps, feature learning with the SOFTNet regression network, and finally the expression spotting procedure.

2.1. Feature Extraction

The prevalent use of optical flow features in several works in micro-expression analysis [5, 15, 16] have shown the usefulness of spatio-temporal motion information. To normalize the face resolution, the facial region in each frame is cropped and resized to 128×128 pixels. Cropping was performed using the Dlib toolbox [17] after the 68 landmark points were detected from the first (reference) frame of each raw video.

Subsequently, *optical flow* features are computed from two frames, *i.e.* current frame F_i and the k-th frame from i, F_{i+k}, where k is half of the average length of an expression. Horizontal and vertical components, $\mathbf{u}$ and $\mathbf{v}$ respectively are computed using TV-L1 optical flow method [18]. Additionally, *optical strain*, which is adopted from infinitesimal strain theory, captures the subtle facial deformation from optical flow components [5]. It can be defined as follows:

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

where ϵ_{xx} and ϵ_{yy} indicate normal strain components while ϵ_{xy} and ϵ_{yx} indicate shear strain components. The optical strain magnitude ϵ, can be computed as:

$$|\epsilon| = \sqrt{\epsilon_{xx}^2 + \epsilon_{yy}^2 + \epsilon_{xy}^2 + \epsilon_{yx}^2}. \quad (2)$$

These three components ($\mathbf{u}$, $\mathbf{v}$, and ϵ) represent the input data for the model learning phase.

2.2. Pre-processing

Prior to the learning phase, a series of pre-processing steps are introduced to ensure consistency of the data before model learning. Motivated by the work of [19], we take the landmark position of the nose region with five pixels margin to eliminate the global head motion for each frame.

Then, we omit the left and right eye regions since optical flow features are highly sensitive to eye blinking [9]. Fol-

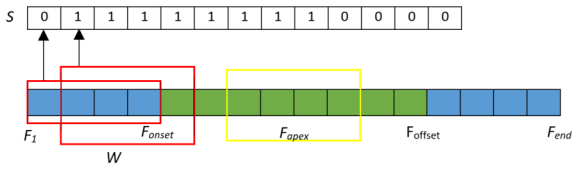

Fig. 3: Pseudo-labeling in a video using sliding window approach. Blue frames are not part of the expression interval while green frames are within the interval.

lowing this work, a polygon (with extra margin of 15 pixels along the width and height) is applied to mask out these eye regions. Next, the area bounded by three regions: (1) left eye and left eyebrow; (2) right eye and right eyebrow; (3) mouth, are selected on the basis that they typically contain significant movements [10, 9]. The area bounded by these regions are re-sampled (to 21×21, 21×21 for the left and right eye regions and 21×42 for mouth region) and pieced together (in that same order) to form a square 42×42 image which retains the important features while removing non-informative areas.

2.3. Shallow Optical Flow Three-stream CNN

SOFTNet. Motivated by the architecture in [15], we propose SOFTNet with these further considerations: (1) The convolutional layer applies a 5×5 filter rather than 3×3 to increase the receptive field coverage to accommodate macro-expressions; (2) A regression output layer added to predict the score of each frame corresponding to its likelihood of being involved in the interval of expression. Intuitively, optical flow features are typically not as significant at the middle frames of a sequence (yellow color window in Fig. 3) as compared to frames nearer to the F_{onset} and F_{offset} hence it is desirable to regress a high score for peak detection.

SOFTNet is a three-stream (very) shallow architecture where each stream consisting of a single convolutional layer with 3, 5, and 8 filters respectively, followed by a max-pooling layer to reduce the feature map size. The feature maps from each stream are then stacked channel-wise to combine the features, with another max-pooling layer thereafter. Finally, it flattens out to a 400-node layer, fully connected to a single output score via linear activation. Specifically, the learned model $\mathcal{M}$ takes in the three optical flow components $\mathbf{u}$, $\mathbf{v}$, and ϵ of the i-th frame as input to each stream and predicts a spotting confidence score $\hat{s}_i$. Separate models are learned to spot macro-expressions (α) and micro-expressions (β), *i.e.* $\hat{s}_{i,\phi} = \mathcal{M}_\phi(\mathbf{u}_i, \mathbf{v}_i, \epsilon_i)$ where $\phi = \{\alpha, \beta\}$.

Pseudo-labeling. Since ground-truth labels only provide the onset, offset and apex frame indices, to realize a sliding window mechanism, we need to create labels for each window position. To label the frame in videos, the sliding window, W_j at the j-th position with length k[1] corresponding to interval $[F_i, F_i + k - 1]$, is scanned across each video. We impose a pseudo-labeling function g (for Heaviside step function, $g(IoU) = 0$ if $IoU \leq 0$, else $g(IoU) = 1$), which to determine the score s for each j-th window calculated from the IoU between W and $\mathcal{E}$:

$$IoU = \frac{W \bigcap \mathcal{E}}{W \bigcup \mathcal{E}} \quad \text{where } \mathcal{E} = [F_{onset}, F_{offset}] \tag{3}$$

Finally, the pseudo-label set $S = \{s_{i,\phi} \text{ for } i = 1, \ldots, F_{end} - k\}$ which represents the labels (of ϕ) for the SOFTNet inputs, is obtained, as illustrated in Figure 3 example. Other pseudo-labeling functions such as linear function and step function were found to be less desirable after experiments.

Training configurations. In our experiments, we applied SGD with learning rate 5 x 10^{-4} with the number of epochs set to 10. Since the dataset is highly imbalanced, we opt to sample 1 of every 2 non-expression frames, similar to the strategy in [11]. Data augmentation including horizontal flip, Gaussian blur (7×7), and adding random Gaussian noise ($\mathcal{N}(0, 1)$), is performed during micro-expression training only to address the small sample size problem.

2.4. Spotting

The predicted score of each frame is aggregated as:

$$\hat{s}_{i,\phi} = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} \hat{s}_{j,\phi} \text{ for } i = F_1 + k, \ldots, F_{end} - k \tag{4}$$

whereby the predicted scores from k frames before until k frames after the current i-th frame are averaged for smoothing purpose. Intuitively, each frame now represents a potential interval of expression by accumulation of confidence scores.

Finally, we employ the standard threshold and peak detection technique of [6] to spot the peaks in each video where the threshold is defined as:

$$T = \hat{S}_{mean} + p \times (\hat{S}_{max} - \hat{S}_{mean}) \tag{5}$$

where $\hat{S}_{mean}$ and $\hat{S}_{max}$ are the average and maximum predicted score over the entire video, and p is a tuning parameter in the range of $[0, 1]$. As shown in Figure 2 spotting phase, the green line (bottom row) is the threshold and red lines indicate a few intervals of expressions. A peak frame $s_{P,\phi}$ is spotted by finding a local maxima (with minimum distance of k between peaks) and extending by k frames to obtain the spotted interval $\hat{\mathcal{E}}_\phi = [s_{P,\phi} - k, s_{P,\phi} + k]$ for evaluation.

3. EXPERIMENTS

To demonstrate the effectiveness of the proposed framework, we conduct extensive experiments on the MEGC 2020 spotting benchmark. It is important to note that the SOFTNet models are implemented separately (*i.e.* training and inference) for both macro- and micro-expressions. To encourage community usage, the code is publicly available [2].

3.1. Evaluation Details

Datasets. Two benchmark datasets, namely CAS(ME)^2 [22] and SAMM Long Videos [21] are used. Briefly, CAS(ME)^2

[1] $k = (N + 1)/2$ is half the average length of expression in each dataset.

[2] Link: `https://github.com/genbing99/SoftNet-SpotME`

Table 1: Comparison between the proposed approach against baseline and state-of-the-art methods in F1-score

Dataset	CAS(ME)^2			SAMM Long Videos		
Methods	Macro	Micro	Overall	Macro	Micro	Overall
Baseline [20]	0.1196	0.0082	0.0376	0.0629	0.0364	0.0445
Gan et al [3]	0.1436	0.0098	0.0448	-	-	-
Pan [3]	-	-	0.0595	-	-	0.0813
Zhang et al. [19]	0.2131	0.0547	0.1403	0.0725	0.1331	0.0999
Yap et al. [21]	-	-	-	**0.4081**	0.0508	**0.3299**
Ours (W/o SOFTNet)	0.1615	**0.1379**	0.1551	0.1463	0.1063	0.1293
Ours (With SOFTNet)	**0.2410**	0.1173	**0.2022**	0.2169	**0.1520**	0.1881

Table 2: Detailed result of the proposed SOFTNet approach

Dataset	CAS(ME)^2			SAMM Long Videos		
Expression	Macro	Micro	Overall	Macro	Micro	Overall
Total	300	57	357	333	159	492
TP	90	20	110	68	38	106
FP	357	264	621	226	303	529
FN	210	37	247	265	121	386
Precision	0.2013	0.0704	0.1505	0.2313	0.1114	0.1669
Recall	0.3000	0.3509	0.3081	0.2042	0.2390	0.2154
F1-Score	0.2410	0.1173	0.2022	0.2169	0.1520	0.1881
AP@[.5:.95]	0.0168	0.0112	0.0140	0.0117	0.0103	0.0110

contains 98 long videos consisting of 300 macro-expressions and 57 micro-expressions captured from 22 subjects; SAMM Long Videos is an extension of SAMM [23], one of the most culturally diverse datasets in this domain, with 147 long videos (343 macro-movements, 159 micro-movements) elicited from 32 subjects. However, a small number (10) of macro-expression samples were discarded due to the ambiguous onset annotation. In spite of that, both datasets were fully annotated with onset, apex, and offset by professional coders.

Performance Metric. We benchmark our proposed approach against recent works from MEGC 2020 [3], adopting the similar F1-score metric for both macro- and micro-expression spotting. Besides, we propose the use of Average Precision over different Intersection over Union (IoU) thresholds from 0.5 to 0.95 with a step size of 0.05 (denoted as **AP@[.5:.95]**), a popular metric used in MS COCO [24], to provide a more consistent measure of the quality of the spotting result.

Settings. Leave-one-subject-out (LOSO) cross-validation is applied to ensure all samples are evaluated. For peak detection, we empirically select $p = 0.55$ for SOFTNet and $p = 0.5$ for without SOFTNet. Parameter k is computed to be $\{6, 18\}$ for CAS(ME)^2 and $\{37, 174\}$ for SAMM (smaller value for micro, larger value for macro).

3.2. Results and Discussions

Table 1 compares the performance of our proposed approach with the accepted submissions in MEGC 2020 [3] on both datasets. Our best approach is capable of outperforming other methods on CAS(ME)^2 while on the SAMM Long Videos, we achieve the highest F1-score for micro-expressions and is second best for macro-expressions, behind the original dataset authors [21]. The control experiment (without SOFTNet and image resampling in pre-processing) determines the spotting score by the sum of the feature map for each frame as suggested in [5]. By examining in detail our SOFTNet approach in Table 2, the amount of TP that we obtained is comparable with other approaches whilst with a much lower FP. The FN is less problematic, but this is proven to be an obstacle in the SAMM Long Videos. The AP@[.5:.95] metric offers a way of equalizing the impact of the half-window length k by considering different IoU levels for matching the intervals.

Table 3: Performance comparison between various network backbones (for CAS(ME)^2 macro-expressions)

Network	F1-score	AP@ [.5:.95]	Inference Time (s)	Parameter (Million)
SOFTNet	**0.2410**	**0.0168**	**2.7826**	**0.3148**
MobileNetV2	0.2152	0.0160	10.1651	2.2631
ResNet-18	0.2147	0.0150	9.3175	11.2877
ResNet-50	0.1155	0.0039	22.2178	23.5960
VGG-16	0.1724	0.0095	8.2692	14.7152

Ablation Studies & Insights. Table 3 compares the SOFTNet against a few popular architectures, showing its superiority across various aspects from accuracy to efficiency. Another ablation study on the choice of pseudo-labeling function g shows the unit step (0.2410) performing better than linear (0.2269) and step (0.2092) functions in F1-score. To offer insights into the predictions, we used timeline plots and GradCAM [25] heatmaps to visualize the spotted temporal and spatial locations. Fig. 1 shows an example of how each stream contributes towards the final outcome.

4. CONCLUSION

This paper proposes a new regression-based strategy towards macro- and micro-expression spotting in long videos by means of a three-stream shallow network based on optical flow information. On the MEGC 2020 benchmark, our approach achieved promising results on both CAS(ME)^2 and SAMM Long Videos. No less importantly, we re-introduce the AP@[.5:.95] metric (from object detection) which measures more consistently across both expression types. We surmise through findings in this paper that spotting both micro- and macro-expressions demands for innovative modeling of the localized facial transitions and robust peak detection.

Acknowledgement: This work is supported in part by Malaysia Ministry of Education FRGS Research Grant (Project No: FRGS/1/2018/ICT02/MMU/02/2).

5. REFERENCES

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

[2] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: databases, methods, and challenges," *Frontiers in Psychology*, vol. 9, p. 1128, 2018.

[3] J. Li, S. Wang, M. H. Yap, J. See, X. Hong, and X. Li, "Megc2020-the third facial micro-expression grand challenge," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pp. 234–237.

[4] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 1, pp. 28–43, 2011.

[5] M. Shreve, J. Brizzi, S. Fefilatyev, T. Luguev, D. Goldgof, and S. Sarkar, "Automatic expression spotting in videos," *Image and Vision Computing*, vol. 32, no. 8, pp. 476–486, 2014.

[6] A. Moilanen, G. Zhao, and M. Pietikäinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *2014 22nd Int. Conf. on Pattern Recognition*, 2014, pp. 1722–1727.

[7] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *Bmvc*, vol. 1, no. 2. Citeseer, 2006, p. 3.

[8] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. of IEEE CVPR*, 2013, pp. 3444–3451.

[9] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Automatic micro-expression recognition from long video using a single spotted apex," in *Asian Conf. on Computer Vision (ACCV)*. Springer, 2016, pp. 345–360.

[10] S.-T. Liong, J. See, K. Wong, A. C. Le Ngo, Y.-H. Oh, and R. Phan, "Automatic apex frame spotting in micro-expression database," in *3rd IAPR Asian Conf. on Pattern Recognition (ACPR)*, 2015, pp. 665–669.

[11] J. Li, C. Soladie, and R. Seguier, "Ltp-ml: Micro-expression detection by recognition of local temporal pattern of facial movements," in *13th IEEE FG*, 2018, pp. 634–641.

[12] M. Verburg and V. Menkovski, "Micro-expression detection in long videos using optical flow and recurrent neural networks," in *14th IEEE FG*, 2019, pp. 1–6.

[13] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, 2017.

[14] T.-K. Tran, Q.-N. Vo, X. Hong, X. Li, and G. Zhao, "Micro-expression spotting: A new benchmark," *arXiv preprint arXiv:2007.12421*, 2020.

[15] S.-T. Liong, Y. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *14th IEEE FG*, 2019, pp. 1–5.

[16] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 36–40.

[17] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[18] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo, "Tv-l1 optical flow estimation," *Image Processing On Line*, vol. 2013, pp. 137–150, 2013.

[19] L.-W. Zhang, J. Li, S.-J. Wang, X.-H. Duan, W.-J. Yan, H.-Y. Xie, and S.-C. Huang, "Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences," in *15th IEEE FG*, 2020, pp. 245–252.

[20] Y. He, S.-J. Wang, J. Li, and M. H. Yap, "Spotting macro-and micro-expression intervals in long video sequences," *arXiv preprint arXiv:1912.11985*, 2019.

[21] C. H. Yap, C. Kendrick, and M. H. Yap, "Samm long videos: A spontaneous facial micro-and macro-expressions dataset," *arXiv preprint arXiv:1911.01519*, 2019.

[22] F. Qu, S.-J. Wang, W.-J. Yan, and X. Fu, "Cas (me) 2: A database of spontaneous macro-expressions and micro-expressions," in *International Conference on Human-Computer Interaction*. Springer, 2016, pp. 48–59.

[23] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2016.

[24] R. Girshick, "Fast r-cnn," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2015, pp. 1440–1448.

[25] 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 *IEEE ICCV*, 2017, pp. 618–626.