This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

Short and Long Range Relation Based Spatio-Temporal Transformer for Micro-Expression Recognition

Liangfei Zhang, Xiaopeng Hong, *Member*, Ognjen Arandjelović, *Member* and Guoying Zhao, *Fellow*

Abstract—Being spontaneous, micro-expressions are useful in the inference of a person's true emotions even if an attempt is made to conceal them. Due to their short duration and low intensity, the recognition of micro-expressions is a difficult task in affective computing. The early work based on handcrafted spatio-temporal features which showed some promise, has recently been superseded by different deep learning approaches which now compete for the state of the art performance. Nevertheless, the problem of capturing both local and global spatio-temporal patterns remains challenging. To this end, herein we propose a novel spatio-temporal transformer architecture – to the best of our knowledge, the first purely transformer based approach (i.e. void of any convolutional network use) for micro-expression recognition. The architecture comprises a spatial encoder which learns spatial patterns, a temporal aggregator for temporal dimension analysis, and a classification head. A comprehensive evaluation on three widely used spontaneous micro-expression data sets, namely SMIC-HS, CASME II and SAMM, shows that the proposed approach consistently outperforms the state of the art, and is the first framework in the published literature on micro-expression recognition to achieve the unweighted F1-score greater than 0.9 on any of the aforementioned data sets.

Index Terms—Emotion recognition, long-term optical flow, temporal aggregator, self-attention mechanism.

1 Introduction

Facial expressions play an important role in interpersonal communication and their recognition is one of the most significant tasks in affective computing. Though there some disagreement on this remains, a notable number of psychologists believe that although due to different cultural environments individuals use different languages to communicate, the expression of their emotions is rather universal [1]. Correctly recognizing facial expressions is important in general communication and can help understanding people's mental state and emotions.

When colloquially used, the term 'facial expressions' refers to what are more precisely technically termed *facial macro-expressions* (MaEs). While crucial for human interaction, providing a universal and non-verbal means of articulating emotion [2], facial macro-expressions can be effected voluntarily which means that they can be used to deceive. In other words, a person's macro-expression may not accurately represent their truly felt emotion. However, whatever the conscious effort, felt emotions effect short-lasting contraction of facial muscles which are expressed *involuntarily* under psychological inhibition. The resulting minute, sudden, and transient expressions are referred to as *micro-expressions* (MEs). After being first observed and recognized as a phenomenon of interest by Haggard and Isaacs [3], and then further elaborated on by a case study reported by Ekman and Friesen [4], MEs began to be researched more widely by psychologists, and in the last decade attracting interest within the field of computer vision [5]. In contrast to MaEs, MEs are subtle. They are exhibited for 0.04s to 0.2s [1], and with lesser facial movement. These characteristics make MEs harder to be recognized than MaEs, whether manually (i.e. by humans) or automatically (i.e. by computers).

The seminal work by Pfister, *et al.* and the release of the database of micro-expression movie clips, namely SMIC-sub (Spontaneous Micro-expression Corpus) [6], effected a marked empowerment of computer scientists in the realm of micro-expression recognition (MER). The first generation of solutions built upon the well-established computer vision tradition and introduced a series of handcrafted features, such as Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) [7], 3 Dimensional Histograms of Oriented Gradients (3DHOG) [8], Histograms of Image Gradient Orientation (HIGO) [9] and Histograms of Oriented Optical Flow (HOOF) [10] and their variations. The next generation shifted focus towards Convolutional Neural Network (CNN) based deep learning methods [11], [12], [13], [14], [15]. Early work by and large uses convolutional kernels to extract spatial information from micro-expression video frames. This kind of pixel level operators can be considered as capturing "*short-range*", local spatial relationships. "*Long-range*", global relationships between different spatial regions have also been proposed and studied, notably by means of Graph Convolutional Network (GCN) based architectures [16], [17], [18], [19], [20]. The activations of Facial Action Units (AUs) are generally used as nodes

- *L. Zhang and O. Arandjelović are with the School of Computer Science, University of St Andrews, UK. E-mail: lz36@st-andrews.ac.uk, oa7@st-andrews.ac.uk*
- *X. Hong is with Harbin Institute of Technology, P.R.China. E-mail: hongxiaopeng@ieee.org*
- *G. Zhao is with University of Oulu, Finland. E-mail: guoying.zhao@oulu.fi*

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

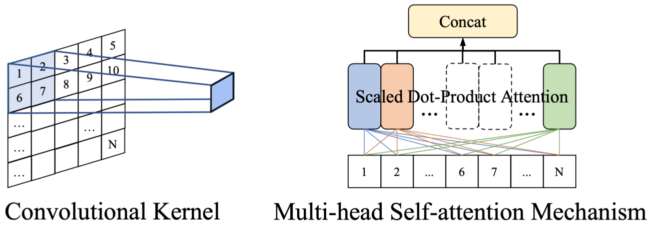


Fig. 1. Comparison of the different spatial feature extraction methods of CNN and transformer.

to build graphs. The relationships between different AU engagements are combined with image features to improve the discriminatory power in the context of MER. However, though these approaches consider global spatial relations so as to assist learning, they can only learn these after local features are extracted, i.e. they are unable to learn both kinds of relations jointly.

In order to capture automatically both short- and long-range relations at the same time, we apply Multi-head Self-attention Mechanism (MSM) instead of a Convolutional Kernel as the cornerstone of our deep learning MER architecture. As shown in Fig. 1, the relations between block 1 and N will hardly ever be learnt by CNN but has been considered at the beginning of MSM. MSM based networks are called *Transformer*. Short-range and long-range relationships between elements of a sequence can be learned in a parallelized manner because transformers utilize sequences in their entirety, as opposed to processing sequence elements sequentially like recurrent networks. Most recently, transformer networks came to the attention of the CV community. By dividing them into smaller constituent patches, two-dimensional images can be converted into one-dimensional sequences, translating the spatial relationships into the relationships between sequence elements (image patches). In this way, transformer networks can be simply applied to vision problems and on various tasks they have outperformed CNNs [21]. Examples include segmentation [22], image super-resolution [23], image recognition [24], [25], video understanding [26], [27] and object detection [28], [29].

Most MER research in the published literature is video based, as Ben et al. elaborated [30], though there is a small but notable body of work on single-frame analysis [31], [32], [33]. This statistic reflects the consensus that for best performance both spatial and temporal information need be considered. In particular, absolute and relative facial motions are extracted and analysed through spatial and temporal features respectively. Most handcrafted methods in existence use the same kind of operator to detect spatial and temporal information from different dimensions by considering the frames as 3D data. The resulting spatio-temporal features with uniform format are used together to implement video based MER. In deep learning based methods, spatial features are mainly extracted by means of a convolutional neural network. Some concatenate spatial features extracted from each frame and others use recurrent neural networks to derive temporal information. To integrate various spatio-temporal relations, our design makes use of long-term temporal information in spatial data (i.e. each frame of video sample) prior to the spatial encoder, and a temporal aggregation block to fuse both short- and long-term temporal relationships afterwards.

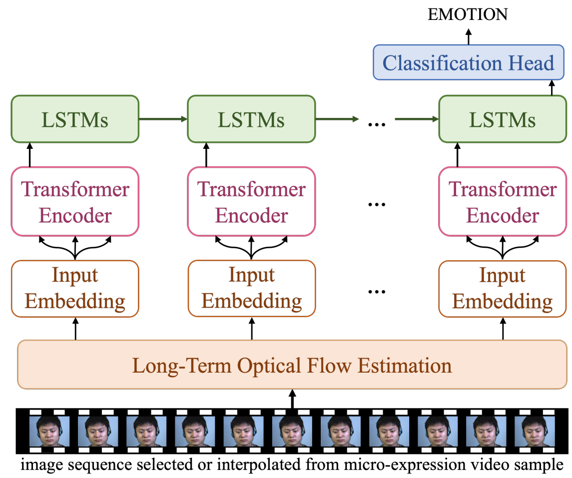


Fig. 2. The framework of proposed Short and Long range relation based Spatio-Temporal Transformer (SLSTT).

In this work we show how a transformer based deep learning architecture can be applied to MER in a manner which outperforms the current state of the art. The main contributions of the present work are as follows:

1) We propose a novel spatio-temporal deep learning transformer framework for video based micro-expression recognition, which we name *Short and Long range relation based Spatio-Temporal Transformer (SLSTT)*, the structure whereof is summarized in Fig. 2. To the best of our knowledge, ours is the first deep learning MER work of this kind, in that it does not employ a CNN at any stage, but is rather entirely centred on a transformer architecture.
2) We use matrices of long-term optical flow, computed in a novel way particularly suited for MER, instead of the original colour images as the input to our network. The feature ultimately arrived at combines long-term temporal information and short- and long-range spatial relations, and is derived by a transformer encoder block.
3) We design a temporal aggregation block to connect spatio-temporal features of spatial relations extracted from each frame by multiple transformer encoder layers and achieve video based MER. The empirical performance and analysis of mean and LSTM (long short-term memory) aggregators is presented too.

We evaluate our approach on the three well known and popular ME databases, Spontaneous Micro-Expression Corpus (SMIC) [34], Chinese Academy of Sciences Micro-Expression II (CASME II) [35] and Spontaneous Actions and Micro-Movements (SAMM) [36], in both Sole Database Evaluation (SDE) and Composite Database Evaluation (CDE) settings and achieve state of the art results.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

2 Related Work

2.1 Micro-expression Recognition

Since the publication of the SMIC data set in 2013, the volume of research on automatic micro-expression recognition has been increasing steadily over the years. From the handcrafted computer vision methods in the early years to the deep learning approaches more recently, the main ideas of micro-expression feature extraction could be categorized as primarily pursuing either a spatial strategy or a temporal one.

2.1.1 Spatial Features

The fundamental challenge of computer vision is that of extracting semantic information from images or videos. Whatever the approach, the extraction of some kind of spatial features is central to addressing this challenge. Micro-expression recognition is no exception. In a manner similar to many gradient based features applied previously on generic computer vision tasks, Polikovsky et al. [8] proposed the use of a gradient feature adapted to MER to describe local dynamics of the face. The magnitudes of local gradient projections in the XY plane is used to construct histograms across different regions, which are used as spatial features. LBP quickly became the most popular operator for micro-expression analysis after Pfister et al. [6] first applied it to MER. This operator describes local appearance in an image. The key idea behind it is that the relative brightness of neighbouring pixels can be used to describe local appearance in a geometrically and photometrically robust manner. Its widespread use and favourable performance often make it the default baseline method when new data sets are published, or a new ME related task proposed. As for deep learning approaches, CNN model can be thought as a combination of two components: a feature extraction part and a classification part. The convolution and pooling layers perform spatial feature extraction.

Further to local appearance based features, numerous other strategies have been described for spatial feature extraction in micro-expression analysis. One of the simplest and commonest of these employs facial Region Of Interest (ROI) segmentation. Polikovsky et al. [8] segmented each face sample into 12 regions according to the Facial Action Coding System (FACS) [37], each region corresponding to an independent facial muscle complex, and applied appearance normalization to individual regions. Others have modified or extended this strategy, e.g. employing different methods for segmentation or different salient regions – 11 [38], 16 [39], 36 [10] instead of 12 of Polikovsky et al. Spatial feature operators are applied with each ROI rather the whole image, thus providing a more nuanced description of the face. In recent years, a more principled equivalent of this strategy (in that it is learnt, rather than predetermined by a human), can be found in the form of attention blocks applied within neural networks to improve their ability to learn spatial features. These blocks can generate weight masks for feature maps, helping a network pay greater attention to significant regions. Most recently, GCNs have also been used within deep learning frameworks as a means of capturing spatial information, often using AUs as correponding to graph nodes. For example, Lei et al. [20] segment node patches based on facial landmarks and fuse them with an AU GCN. Xie et al. [18] infer AU node features from the backbone features by global average pooling and use them to build an AU relation graph for GCN layers. These optimization measures use a priori knowledge (AUs in FACS) to enhance the extracted spatial features. Long-range spatial relationships are not directly learnt by such networks.

2.1.2 Temporal Features

Since one of the most characteristic aspects of micro-expressions is their sudden occurrence, temporal features cannot be ignored. While some methods in the literature do use only the single, apex frame instead of all frames in each ME sample [31], [32], [33], [40], most employ all in the range between the onset frame and the offset, thus treating all temporal changes within this time period on the same footing. Some go further and employ temporal frame interpolation (as indeed we do herein) so as to increase the frame count [6], [9], [10], [12], [39].

A vast number of handcrafted feature based approaches treat raw video data as a 3D spatio-temporal volume, treating the temporal dimension as no different than the spatial ones. In other words, they apply the same kind of operator used to extract spatial features on pseudo-images formed by a cut through the 3D volume comprising one spatial dimension and the temporal dimension. For example, in LBP-TOP, LBP operators are applied on XT and YT planes to extract temporal features, and their histogram across the three dimensions forms the final representation. 3DHOG similarly treats videos as spatio-temporal cuboids with no distinction made between the three dimensions, but arguably with even greater uniformity than LBP-TOP in that the descriptor itself is inherently 3D based. Similar in this regard are optical flow based features, which too inherently combine local spatial and temporal elements – the use of optical strain [41], flow orientation [10] or its magnitude [31] are all variations on this theme.

As an alternative to the use of raw appearance imagery as input to a deep learning network, the use of preprocessed data in the form of optic flow matrices has been proposed by some authors [15], [19], [42]. In this manner, proximal temporal information is exploited directly. On the other hand, the learning of longer range temporal patterns has been approached in a variety of ways by different authors. Some extract temporal patterns simply by treating video sequences as 3-dimensional matrices [16], [41], [43], rather than 2-dimensional ones which naturally capture single images. Others employ structures such as the recurrent neural network (RNN) or the LSTM [12], [44]. In addition to the use of off-the-shelf recurrent deep learning strategies, recently there has been an emergence of methods which apply domain specific knowledge so as to make the learning particularly effective for micro-expression analysis [15].

2.2 Transformers in Computer Vision

For approximately a decade now, convolutional neural networks have established themselves as the backbone of most deep learning algorithms in computer vision. However, convolution always operates on fixed size windows and is thus unable to extract distal relations. The idea of a

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

transformer was first introduced in the context of NLP. It relies on a self-attention mechanism, learning the relationships between elements of a sequence. Transformers are able to capture 'long-term' dependence between sequence elements which is challenging for conventional recurrent models to encode. By dividing an image into sub-images and imposing a consistent ordering on them, a planar image can be converted into a sequence, so spatial dependencies can be learned in the same way as temporal features. For this reason, transformer based deep learning architectures have recently gained significant attention from the computer vision community and are starting to play an increasing role in a number of computer vision tasks.

A representative example in the context of object detection is the DEtection TRansformer (DETR) [28] framework which uses transformer blocks first, for regression and classification, but the visual features are still extracted by a CNN based backbone. The Image Generative Pre-Training (iGPT) approach of Chen et al. [45] attempts to exploit the strengths of transformers somewhat differently, pre-training BERT (Bidirectional Encoder Representations from Transformers) [46], originally proposed for language understanding, and thereafter fine tuning the network with a small classification head. iGPT uses pixels instead language tokens within BERT, but suffers from significant information loss effected by a necessary image resolution reduction. In the context of classification, the Vision Transformer (ViT) approach of Dosovitskiy et al. [24] applies transformer encoding of image patches as a means of extracting visual features directly. It is the first pure vision transformer, and in its spirit and design, follows the original transformer [47] architecture faithfully. As such, it facilitates the application of scalable transformer architectures used in NLP effortlessly.

Following these successes, transformers have been applied to a variety of computer vision tasks, including those in the realm of affective computing [48], [49]. Notable examples include facial action unit detection [50] and facial image-based macro-expression recognition [51]. However, none of the existing approaches to micro-expression recognition adequately make use of both the spatial and temporal information due to the design difficulties posed by the challenges we discussed in the previous sections.

3 Proposed Method

In the present work we propose a method that takes advantage both of the physiological understanding of micro-expressions and their characteristics, as well as of the transformer framework. The approach overcomes many of the weaknesses of the existing MER methods in the literature as discussed in the previous section. Importantly, our method is able to extract and thus benefit both from proximal (i.e. short-range) and distal (i.e. long-range) spatio-temporal features. Each element of the proposed framework is laid out in detail next, corresponds to each sub-section.

3.1 Long-term Optical Flow

Optical flow describes the apparent motion of brightness patterns between frames, caused by the relative movement of the content of a scene and the camera used to image it [52]. If the camera is static, optical flow can be used to infer both the direction and the magnitude of an imaged object's movement from the change in the appearance of pixels between frames [53].

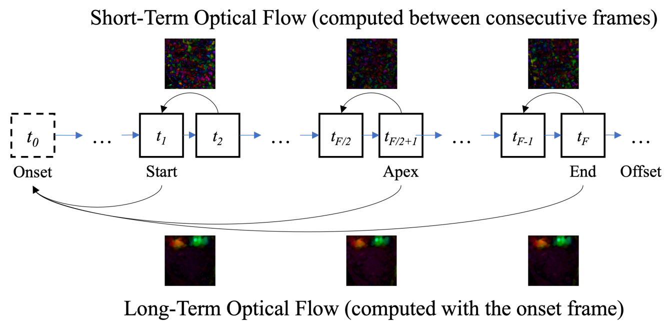
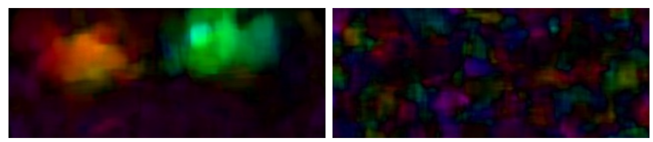


Fig. 3. Different computing mechanism between short- and long-term optical flow.

(a) Onset frame, magnified region of interest

(b) Apex frame with superimposed long-term optical flow samples



(c) Magnitude of long-term optical flow between onset & apex frames

(d) Magnitude of short-term optical flow between apex frame and its previous one

Fig. 4. Illustration of optic flow computed between the onset and the apex frame, corresponding to the motion effected by the activation unit Brow Lowerer (AU4).Compare with the one computed between consecutive frames.

Optical flow is inherently temporally local, i.e. save for practical considerations (numerical, efficiency, etc.) it is computed between consecutive frames of sequence. This introduces a problem when micro-expression videos are considered, created by the already noted limited motion exhibited during the expressions. Therefore, herein we propose to calculate optical flow between each sample frame and the onset frame instead of consecutive frames, see Fig. 3. To see the reasons behind this choice, consider Fig. 4 which shows optical flow fields of consecutive frames starting with the micro-expression onset frame. It can be readily observed that the fields are rather similar up to the apex frame, which can be attributed to the aforementioned brevity of the expression, with a similar trend thereafter but in the opposite direction. In contrast, our, temporally non-local modified optical flow – long-term optical flow in a manner of speaking – exhibits a much more structured pattern, always being in the same direction, increasing in magnitude up to the apex frame and declining in magnitude thereafter. This results in much more stable and discriminative features associated with each micro-expression.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

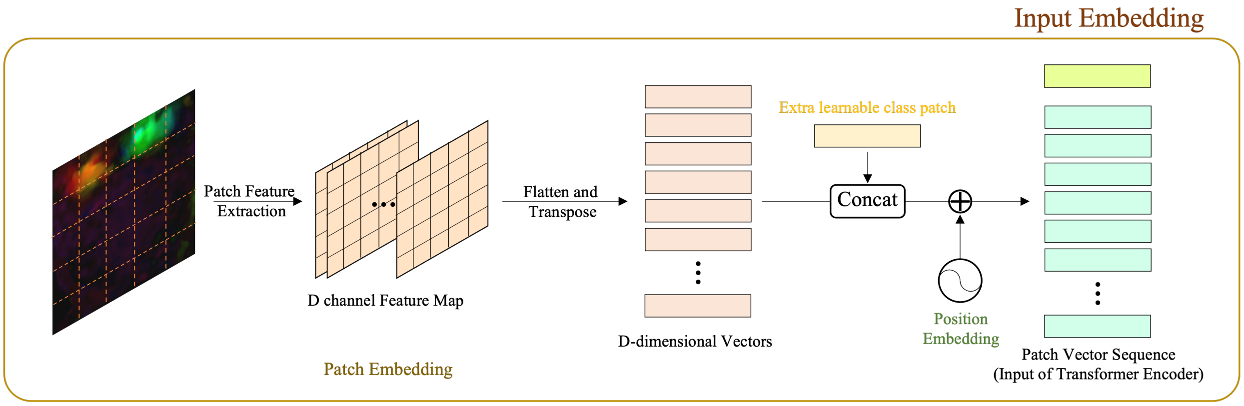


Fig. 5. Long-term optical flow fields are as inputs of the Input Embedding blocks. After short-range spatial feature extraction, patch and position embedding, the resulting sequence of vectors are fed to standard transformer encoder layers.

3.2 Spatial Feature Extraction

The key idea underlying the proposed method lies in the extraction of long-range spatial relations from each frame using a transformer encoder, with images as before being treated as sequences of constituent patches. More specifically, input frames are first represented as vector sequences with local spatial features of each image patch. The resulting sequences are then fed into the transformer encoder for long-term spatial feature extraction.

3.2.1 Input Embedding and Short-Range Spatial Relation Learning

The standard transformer receives a 1D sequence as input. To handle 2D images, we represent each image as a sequence of rasterized 2D patches. Herein we do not use appearance images, that is the original video sequence frames, as input but rather the corresponding optical flow fields. An input embedding block is proposed as a means of representing input images as vector sequences for input to the transformer encoder.

The general input embedding mechanism considers the image $X \in \mathbb{R}^{H\times W\times C}$ as a sequence of non-overlapping $P \times P$ pixel patches, where H, W, and C are respectively the height, the width, and the channel count of the input. Different from the "separate and flat" linear patch embedding proposed by Dosovitskiy et al. [24], we first extract local spatial features in patch regions with a patch-wise fully connected layer. Patches of image X are represented as $X_p \in \mathbb{R}^{N\times(P^2,C)}$. As shown in Fig. 5, we extract the short-range spatial features from image X to feature map $X \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times D}$, flatten and transpose them to N D-dimensional vectors, where $N = \frac{HW}{P^2}$ the resulting number of patches in each image. D-dimensional vectors are passed through all transformer encoder layers. The specific values of parameters used in our experiments are stated in Section 4.

After that, a learnable D-dimensional vector is concatenated with the sequence, as the class token ($Z_0[0] = x_{class}$), whose state as the output of the transformer encoder ($Z_{L_T}[0]$). The effective input sequence length for the transformer encoder is thus $N + 1$. Then a position embedding is added to each vector in the sequence. The whole input embedding procedure can be described as follows:

$$Z_0 = [X_{class}; X_p^1E; X_p^2E; \dots; X_p^NE] + E_{pos},$$
$$E \in \mathbb{R}^{(P^2,C)\times D}, E_{pos} \in \mathbb{R}^{(N+1)\times D}, \qquad (1)$$

where $Z_0 \in \mathbb{R}^{(N\times D)}$ is the input of the transformer encoder.

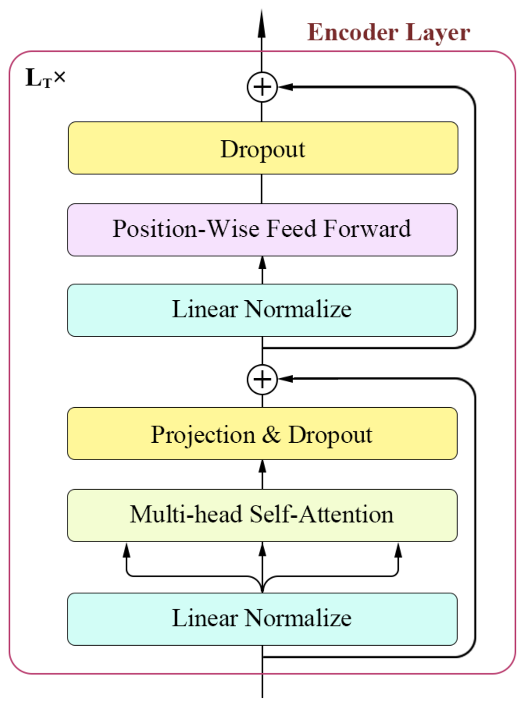


Fig. 6. Detailed structure of a Transformer Encoder layer. The output of frame t processed by spatial encoder is $Z_{L_T}^t$.

3.2.2 Long-Range Spatial Relation Learning by Transformer Encoder

After short-range spatial relation are extracted from the input long-term optical flow fields of each frame and embedded as vectors, they are passed to a transformer encoder for further long-range spatial feature extraction.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

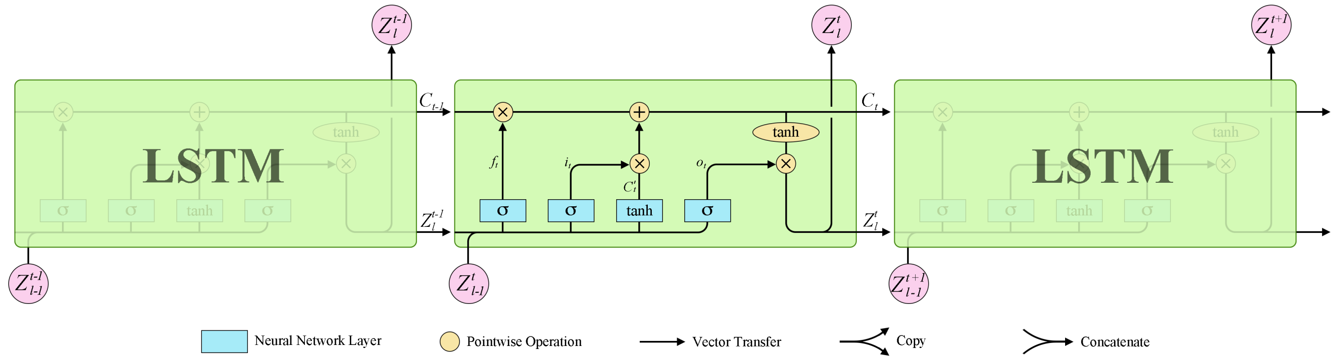


Fig. 7. The repeating module in an LSTM aggregator layer.

Our encoder contains L_T transformer layers; herein we use $L_T = 12$, adopting this value from the ViT-Base model of Dosovitskiy et al. [24] (the pre-trained encoder we use in experiments).Each layer involves two blocks, a Multi-head Self-attention Mechanism (MSM) and a Position-Wise fully connected Feed-Forward network (PWFF), as shown in Fig. 6. Layer Normalisation (LN) is applied before each block and residual connections after each block [54], [55]. The output of the transformer layer can be written as follows:

$$Z'_l = MSM(LN(Z_{l-1})) + Z_{l-1}, l = 1 \dots L_T, \quad (2)$$
$$Z_l = PWFF(LN(Z'_l)) + Z'_l, l = 1 \dots L_T, \quad (3)$$

where Z_l is the output of layer l. The PWFF block contains two layers with the Gaussian Error Linear Unit (GELU) non-linear activation function. The feature embedding dimension thereby first increases from D to $4D$ and then reduces back to D, which equals 768 in our experiments.

Multi-head attention allows the model to focus simultaneously on information content from different parts of the sequences, so both long-range and short-range spatial relations can be learnt. An attention function is mapping a query and a set of key-value pairs to the output, a weighted sum of the values. The weights are computed using a compatibility function of the queries with the corresponding keys, and they are all vectors. The self-attention function is computed on a set of queries simultaneously. The queries, keys and values can be grouped together and represented as matrices Q, K and V, so the computation of the matrix of outputs can be written as:

$$Q = Z_{l-1}W_Q, \quad (4)$$
$$K = Z_{l-1}W_K, \quad (5)$$
$$V = Z_{l-1}W_V, \quad (6)$$

$$SA(Z_l) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V, \quad (7)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D_m}$ are learnable matrices and SA is the self-attention module. MSM can be seen as a type of self-attention with M heads in parallel operation and a projection of their concatenated outputs:

$$MSM(Z_l) = \text{Concat}(\{SA_h(Z_l), \forall h \in [1..M]\})W_O, \quad (8)$$

where $W_O \in \mathbb{R}^{M \cdot D_m \times D}$ is a re-projection matrix. D_m is typically set to $\frac{D}{M}$, so as to keep the number of parameters constant with changing M.

3.3 Temporal Aggregation

After extracting both local and global spatial features associated with each frame using a transformer encoder, we introduce an aggregation block to extract temporal features before performing the ultimate classification. The aggregation function ensures that our transformer model can be trained and applied to the spatial feature sets of each frame, subsequently processing the temporal relations between frames in each sample. Since facial movement during micro-expressions is almost imperceptible, all frames from a single video sample are rather similar one to another. Nevertheless, it is still possible to identify reliably a number of salient frames, such as the apex frame, that play a particularly important role in the analysis of a micro-expression. Therefore, we propose an LSTM architecture for temporal aggregation.

Long Short-Term Memory (LSTM) [56] is a type of recurrent neural network with feedback connections, which overcomes two well-known problems associated with RNNs: the vanishing gradient problem, and the sensitivity to the variation of the temporal gap length between salient events in a processed sequence. The elements of the input are the sets of outputs from the transformer encoder for each frame. The inputs are not concatenated, and the input sequence length is thus dependent on the number of frames in each ME video sample.

We used three LSTM layers in the aggregation block. The computation details of each layer are:

$$t = 1 \dots F, l = L_T + 1 \dots L_A,$$
$$f_t = \sigma(W_f \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_f), \quad (9)$$
$$i_t = \sigma(W_i \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_i), \quad (10)$$
$$o_t = \sigma(W_o \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_o), \quad (11)$$

$$C'_t = tanh(W_C \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_C), \quad (12)$$
$$C_t = f_t \times C_{t-1} + i_t \times C'_t, \quad (13)$$

$$Z_l^t = o_t \times tanh(C_t), \quad (14)$$

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

where F is the number of chosen frames in each video sample, L_A is the total number of layers in both the transformer encoder and the LSTM aggregator. Z_l^t denotes the outputs of the layer l after t frames have been processed. After all frames are processed in this manner, the result is a single feature set describing the entire micro-expression video sample. Finally, these features are fed into an MLP which is used for the ultimate MER classification. The details of how previous output join the latter training are presented in Fig. 7. We also design a comparative experiment to demonstrate the effectiveness of the LSTM aggregator, the details of which are described in the Section 4.3.2.

3.4 Network Optimization

Following the aggregation block, our network contains two fully connected layers which facilitate the final classification achieved using the SoftMax activation function. Cross Entropy loss is used as the objective function for training:

$$L = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i \sum_{c=1}^{C} y_{ic}\log(p_{ic}), \tag{15}$$

where N is the number of the ME video samples and C the number of emotion classes. The value of y_{ic} is 1 when the true class of sample i is equal to c and 0 otherwise. Similarly, p_{ic} is the predicted probability that sample i belongs to class c.

When using gradient descent to optimize the objective function during network training, as the parameter set gets closer to its optimum, the learning rate should be reduced. Herein we achieve this using cosine annealing [57], i.e. using the the cosine function to modulate the learning rate which initially decreases slowly, and then rather rapidly before stabilizing again. This learning rate adjustment is particularly important in the context of the problem at hand, considering that the number of available micro-expression video samples is not large even in the largest corpora, readily learning to overfitting if due care is not taken.

4 Experiments and Evaluation

In this section we describe the empirical experiments used to evaluate the proposed method. We begin with a description of the data sets used, follow up with details on the data pre-processing performed, relevant implementation details, and evaluation metrics, and conclude with a report of the results and a discussion of the findings.

4.1 Databases

Following the best practices in the field, for our evaluation we adopt the use of three large data sets, namely the Spontaneous Micro-Expression Corpus (SMIC) [34], the Chinese Academy of Sciences Micro-Expression II data set (CASME II) [35], and the Spontaneous Actions and Micro-Movement database (SAMM) [36], thus ensuring sufficient diversity of data, evaluation scale, and ready and fair comparison with other methods in the literature. All video samples in these databases capture spontaneously exhibited, rather than acted micro-expressions (see Zhang and Arandjelović [5] for discussion), which is important for establishing the real-world applicability of findings.

4.1.1 SMIC

The Spontaneous Micro-Expression Corpus (SMIC) is the earliest published spontaneous micro-expression database [34]. It comprises three distinct parts captured by cameras of different types, namely a conventional visual camera (VIS), a near-infrared camera (NIR) and a high-speed camera (HS). These subsets are designed to study micro-expression analysis tasks in various application scenarios. To achieve uniformity with the other two corpora, namely CASME-II and SAMM which are described next, which only contain high-speed camera videos, it is the HS subset from SMIC that we make use of herein. The SMIC-HS contains 164 video sequences (samples) from 16 subjects of 3 ethnicities. Using two human labellers, these videos are categorized as corresponding to either negative (70), positive (51), or surprised (43) expression, and both raw and cropped frames are provided.

4.1.2 CASME II

The Chinese Academy of Sciences Micro-Expression II (CASME II) data set contains 247 micro-expression video samples from 26 Chinese participants. The full videos have the resolution of 640×480 pixels. Cropped facial frames in 280×340 pixel resolution (higher than both CASME and SMIC-HS), extracted using the same face registration and alignment method as for SMIC, are also provided. The micro-expression samples in CASME II are labelled by 2 coders to 5 classes, namely Happiness (33), Disgust (60), Surprise (25), Repression (27), and Others (102).

4.1.3 SAMM

The Spontaneous Actions and Micro-Movement (SAMM) database is the newest MER corpus. The 159 micro-expression short videos in the corpus were collected using 32 participants of 13 ethnicities, with an even gender distribution (16 male and 16 female), at 200 fps and the resolution of 2040×1088 pixels, with the face region size being approximately 400×400 pixels. The samples are assigned to one of 8 emotion classes, namely Anger (57), Happiness (26), Other (26), Surprise (15), Contempt (12), Disgust (9), Fear (8) and Sadness (6).

4.2 Data Pre-Processing

4.2.1 Face Cropping

As noted in the previous section, cropped face images are explicitly provided in both SMIC-HS and CASME II data sets, with the same registration method used in both; no cropped faces are provided as part of SAMM. In order to maintain data consistency across different databases, in our experiments we employ a different face extraction approach. In particular, we utilize the Ensemble of Regression Trees (ERT) [58] algorithm implemented in DLib [59] to localize salient facial loci (68 of them) in a uniform manner regardless of which data set a specific video sample came from.

In the case of SMIC-HS and CASME II videos, the original authors' face extraction process consists of facial landmarks detection in the first frame of a micro-expression clip and then the detected face being registered to the model face using a LWM transformation. Motivated by the

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

Fig. 8. The 68 facial landmarks used by our method, shown for the onset (green) and the apex frame (red).

short duration of MEs, the faces in all remaining frames of the video sample are registered using the same matrix. However, in this paper we employ an alternative strategy. The primary reason lies in the need for sufficient and representative data diversity, which is particularly important in deep learning. In particular, the original face extraction method just described, often results in the close resemblance of samples which increases the risk of model overfitting. Therefore, herein we instead simply use a non-reflective 2D Euclidean transformation, i.e. one comprising only rotation and translation. By doing so, at the same time we ensure the correct alignment of salient facial points and maintain information containing facial contour variability.

Furthermore, unlike the authors of SMIC-HS and CASME II, we do not perform facial landmark detection in the first frame of a micro-expression sample, but rather in the apex, thereby increasing the registration accuracy of the most informative parts of the video. As shown in Fig. 8, points 27–30 can be used to determine the centre line of the nose that can be considered as the vertical symmetry line of the entire face area. Point 30 is set as the centre point, and the square size s (in pixels) is computed by adding the vertical distance from the centre point of the eyebrows (19) to the lowest point of the chin (8), $y_{apex[8]} - y_{apex[19]}$, to the height of chin, $y_{apex[8]} - y_{apex[57]}$, so that nearly the entire face is included in the cropped image:

$$s = (y_{apex[8]} - y_{apex[19]}) + (y_{apex[8]} - y_{apex[57]}). \qquad (16)$$

4.2.2 Temporal Interpolation

Considering the short duration of micro-expressions, even when samples are acquired using high-speed cameras, in some instances only a small number (cc. 10) of frames is available. In an attempt to extract accurate temporal information, we also apply frame interpolation from raw videos, effectively synthetically augmenting data. In previous work, the Temporal Interpolation Model (TIM) relies on a path graph to characterize the structure of a sequence of frames, popularly used in several handcrafted feature based methods [9], [13], [60], whereas Liu et al. [10] use simple linear interpolation. Herein we propose a novel approach to interpolation so that its result is smoother in terms of optical flow, it being the nexus of our entire MER methodology. Most existing optical flow based methods produce artifacts on motion boundaries by estimating bidirectional optical flows, scaling and reversing them to approximate intermediate flows. We adopt the Real-time Intermediate Flow Estimation (RIFE) method [61], which uses an end-to-end trainable neural network, IFNet, which speedily and directly estimates the intermediate flows.

Original RIFE interpolates one frame between two given consecutive frames, so we apply it recursively to interpolate multiple intermediate frames. Specifically, given any two consecutive input frames I_0, I_1, we apply RIFE once to get intermediate frame $\hat{I}_{0.5}$ at $t = 0.5$. We then apply RIFE to interpolate between I_0 and $\hat{I}_{0.5}$ to get $\hat{I}_{0.25}$, and so on. In our experiment, we prioritize interpolation in the temporal vicinity of the apex frame. The interpolated queue can be expressed as $\left\{\hat{I}_{a-0.5}, \hat{I}_{a+0.5}, \hat{I}_{a-1.5}, \hat{I}_{a+1.5}, \ldots, \hat{I}_{o+0.5} \; or \; \hat{I}_{f-0.5}\right\}$, where a, o and f are frame indices of the apex, onset, and offset frames respectively. Recall that the apex frames are specified explicitly in CASME II and SAMM, and for SMIC-HS we choose the middle frame of each sample video as the apex. If the number of interpolation frames is lower than the reference count (the average number of frames in this period across the database), we use the same method on the updated frame sequence iteratively to generate further intermediate frames.

4.3 Experimental Settings

4.3.1 Implementation Details

In the spatial feature extraction procedure, we employed base ViT blocks, with 12 Encoder layers, hidden size of 768, MLP size of 3072, and 12 heads. For initialization, we use the official ViT-B/16 model [24] pre-trained on ImageNet [62]. We resize our input images to 384×384 pixels and split each image into patches with 16×16 pixels, so that the number of patches is 24×24. 768-dimensional vectors are passed though all transformer encoder layers. For temporal aggregation, we select 11 frames (apex, and five preceding and succeeding it) per sample as inputs for the mean aggregator and LSTM aggregator. We have tried other options with different number of frame, but it didn't work any better. We only use long-term optical flow in experiments, as motivated by the arguments discussed in Section 3.1. For learning parameters, the initial learning rate and weight decay are set to be 1e-3 and 1e-4, respectively. The momentum for Stochastic Gradient Decent (SGD) is set to 0.9, with the batch size 4 for all experiments. All the experiments were conducted with PyTorch.

4.3.2 Mean Versus LSTM Aggregator

We compare our LSTM aggregator with an alternative which uses the simple mean operator for temporal aggregation. After each frame is processed by spatial encoder, the corresponding output is used in the computation by the mean aggregation layer (layer $L_T + 1$):

$$Z^t_{L_T+1} = \frac{t-1}{t} Z^{t-1}_{L_T+1} + \frac{1}{t} Z^t_{L_T}, t = 1 \ldots F, \qquad (17)$$

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

In a manner similar to that described previously in the context of the LSTM Aggregator, outputs of each frame from our transformer encoder are taken as inputs to the temporal feature extraction module. Compared to the mean operator, LSTM has the advantage of larger expressive capability, resulting in different extracted relationships between different frames. Within the specific context of our work, this means that its ability to distinguish between emotions is also different, with LSTM expected to perform better.

4.3.3 Evaluation Metrics

Following previous work and the Micro-Expressions Grand Challenges (MEGCs), we conducted experiments on SMIC-HS, CASME II, and SAMM, evaluating the classification performance using the corresponding original emotion classes, as well as the composite corpus formed using all three data sets and relabelled using three classes as proposed in MEGC 2019 [63]. All results are reported using LOSO cross-validation. Evaluation is repeated multiple times by holding out test samples of each subject group while the remaining samples are used for training. In this way we best mimic real-world situations and in particular assess the robustness to variability in ethnicity, gender, emotional sensitivity, etc.

4.3.3.1 Sole Database Evaluation (SDE): In the first part of our empirical evaluation, experiments are conducted on three databases individually, using the corresponding original emotion labels, excepting the very rare (and thus underrepresented) classes in CASME II and SAMM. SMIC-HS uses 3 class labels whereas the other two sets both use 5. We use *accuracy* and *macro F1-score* to assess the recognition performance.

4.3.3.2 Composite Database Evaluation (CDE): In the second part of our empirical evaluation, experiments are conducted on the composite database with 3 emotion classes (negative, positive, and surprise). The composite database, that is the database obtained by merging SMIC, CASME II, and SAMM contains the total of 68 subjects, 16 from SMIC, 24 from CASME II and 28 from SAMM. LOSO cross-validation is applied on each database separately and together on the composite database. *Unweighted F1-score (UF1)*, also known as the *macro F1-score* and *Unweighted Average Recall (UAR)* are used to assess performance:

$$UF1 = \textit{macro F1-score}, \tag{18}$$

$$UAR = \frac{\sum_{c=1}^{C} \frac{\sum_{i=1}^{S} TP_{i,c}}{N_c}}{C}, \tag{19}$$

where N_c is the total number of samples of class c across all subjects.

4.4 Results and Discussion

We compare the performance of the proposed approach with baseline handcrafted feature extraction methods and the most prominent recent deep learning based methods on the widely used micro-expression databases, SMIC-HS, CASME II, and SAMM, described in the previous section, both in the SDE and the CDE settings. To ensure uniformity and fairness of the comparison, the SDE results for all methods were obtained in identical conditions, i.e. for the identical number of samples, the number of labels (classes), and using the same cross-validation approach. The details of the performance of our *SLSTT* on different emotion categories are shown in Fig. 9.

As can be readily seen in Table 1 which presents a comprehensive overview of our experimental results in the SDE setting, the method proposed in the present paper performs best (n.b. shown in bold) in all but one testing scenario, in which it is second best (n.b. second best performance is denoted by square brackets), trailing marginally behind the method introduced by Sun et al. [69]. What is more, in most cases our method outperforms rivals by a significant margin.

Moving next to the results of our experiments in the CDE setting, these are summarized in Table 2. It can be readily seen that our method's performance is again shown to be excellent. In particular, in most cases our method again comes out either at the top or second best (as before the former being shown in bold and the latter denoted by square brackets enclosure). The only existing method in the literature which remains competitive against ours is that of Lei et al. [20]. To elaborate in further detail, our approach achieved the best results both in terms of UF1 and UAR on CASME II, and on UF1 on the full composite database, and second best on UAR on the composite database and on UF1 on SMIC-HS. The performance of all methods on CASME II is consistently higher than when applied on other data sets, which suggests that the challenge of MER is increased with ethnic diversity of participants – this should be born in mind in future research and any comparative analysis. It is insightful to observe that in contrast with the results in the SDE setting already discussed (see Table 1), our method does not come out as dominant in the context of CDE. This suggests an important conclusion, namely that our method is particularly capable of nuanced learning over finer grained classes and that its superiority is less able to come through in a simpler setting when only 3 emotional classes as used.

Taking into account the results from both the sole and the composite database experiments, it is useful to observe that when only short-range patterns are utilized, convolutional neural network approaches do not outperform methods based on handcrafted feature. It is the inclusion of long-range spatial learning that is key, as shown by the marked improvement in performance of the corresponding methods. Yet, the proposed method's exceeds even their performance, owing to its use of a multi-head self-attention mechanism, thus demonstrating its importance in MER. The superiority of our short- and long-range relation based spatiotemporal transformer is further corroborated by the results shown in the latest two rows in both Table 1 and Table 2 which summarize our comparison of the proposed LSTM aggregator with the simpler mean operator aggregator.

In CASME II, distinguishing whether a micro-expression is Disgust or Others is inherently difficult because the database contains multiple inconsistently labelled samples with only AU4 activated – some of them are labelled as Others, some as Disgust. It is also worth noting that in SAMM, some AU labels ('AU12 or 14') for the Contempt class were not manually verified, which also causes confusion with the Happiness class (mostly with AU12 labelled). In part,

TABLE 1
SDE results comparison with LOSO on SMIC-HS (3 classes), CASME II (5 classes) and SAMM (5 classes). Best performances are shown in bold, second best by square brackets enclosure. (* Reported by Huang et al. [64], ** Reported by Khor et al. [65])

	SMIC-HS		CASME II		SAMM	
	Acc(%)	**F1**	**Acc(%)**	**F1**	**Acc(%)**	**F1**
Handcrafted						
LBP-TOP*	53.66	0.538	46.46	0.424	–	–
LBP-SIP*	44.51	0.449	46.56	0.448	–	–
STLBP-IP [66] (2015)	57.93	–	59.51	–	–	–
STCLQP [64] (2015)	64.02	0.638	58.39	0.584	–	–
Hierarchical STLBP-IP [67] (2018)	60.37	0.613	–	–	–	–
HIGO+Mag [9] (2018)	68.29	–	67.21	–	–	–
Deep Learning						
AlexNet**	59.76	0.601	62.96	0.668	52.94	0.426
DSSN [65] (2019)	63.41	0.646	70.78	0.730	57.35	0.464
AU-GACN [18] (2020)	–	–	49.20	0.273	48.90	0.310
MER-GCN [16] (2020)	–	–	42.71	–	–	–
Micro-attention [68] (2020)	49.40	0.496	65.90	0.539	48.50	0.402
Dynamic [69] (2020)	**76.06**	0.710	72.61	0.670	–	–
GEME [70] (2021)	64.63	0.616	[75.20]	[0.735]	55.88	0.454
SLSTT-Mean (Ours)	73.17	[0.719]	73.79	0.723	[66.42]	[0.547]
SLSTT-LSTM (Ours)	[75.00]	**0.740**	**75.81**	**0.753**	**72.39**	**0.640**

TABLE 2
CDE results comparison with LOSO on SMIC-HS, CASME II, SAMM and composite database (3 classes). Best performances are shown in bold, second best by square brackets enclosure. (*Reported by See et al. [63], **Reported by Xia et al. [71])

	Composite		SMIC-HS		CASME II		SAMM	
	UF1	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**
Handcrafted								
LBP-TOP*	0.588	0.579	0.200	0.528	0.703	0.743	0.395	0.410
Bi-WOOF*	0.630	0.623	0.573	0.583	0.781	0.803	0.521	0.514
Deep learning								
ResNet18**	0.589	0.563	0.461	0.433	0.625	0.614	0.476	0.436
DenseNet121**	0.425	0.341	0.460	0.333	0.291	0.352	0.565	0.337
Inception V3**	0.516	0.504	0.411	0.401	0.589	0.562	0.414	0.404
WideResNet28-2**	0.505	0.513	0.410	0.401	0.559	0.569	0.410	0.404
OFF-ApexNet* [32] (2019)	0.720	0.710	0.682	0.670	0.876	0.868	0.541	0.539
CapsuleNet [72] (2019)	0.652	0.651	0.582	0.588	0.707	0.701	0.621	0.599
Dual-Inception [73] (2019)	0.732	0.728	0.665	0.673	0.862	0.856	0.587	0.566
STSTNet [41] (2019)	0.735	0.761	0.680	0.701	0.838	0.869	0.659	0.681
EMR [42] (2019)	0.789	0.782	**0.746**	**0.753**	0.829	0.821	**0.775**	[0.715]
ATNet [40] (2019)	0.631	0.613	0.553	0.543	0.798	0.775	0.496	0.482
RCN [71] (2020)	0.705	0.716	0.598	0.599	0.809	0.856	0.677	0.698
AUGCN+AUFsuion [20] (2021)	[0.791]	**0.793**	0.719	[0.722]	[0.880]	[0.871]	[0.775]	**0.789**
SLSTT-Mean (Ours)	0.788	0.767	0.719	0.699	0.844	0.830	0.625	0.566
SLSTT-LSTM (Ours)	**0.816**	[0.790]	[0.740]	0.720	**0.901**	**0.885**	0.715	0.643

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

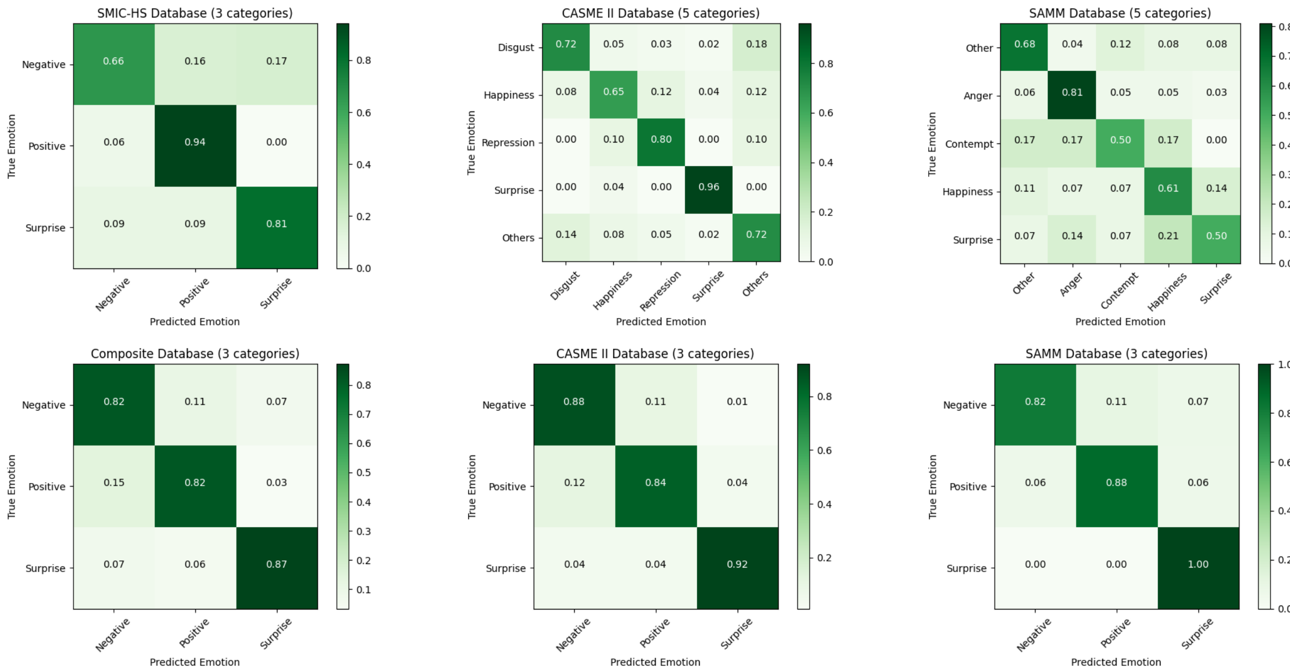


Fig. 9. Confusion matrices corresponding to each of our experiments. Only one is shown for SMIC-HS because the SDE and the CDE are identical when this database is used alone.

these labelling issues emerge from the fact that the mapping between facial action unit activations and emotions (as understood by psychologists) is not a bijection. It is also the case that imperfect information is made use of because only visual data is used. Hence, it should be understood that the theoretical highest accuracy of automated micro-expression recognition on the MER corpora currently used for research purposes is not 100%. The micro-expression databases containing multi-modal signals [74], [75], which have begun emerging recently, seem promising in overcoming some of the limitations of the existing corpora, and we intend to make use of them in our future work.

5 Conclusion

In this paper, we proposed a novel transformer based spatio-temporal deep learning framework for micro-expression recognition, which is the first deep learning work in the field entirely void of convolutional neural network use. In our framework both short- and long-term relations between pixels in spatial and temporal directions of the sample videos can be learned. We use transformer encoder layers with multi-head self-attention mechanism to learn spatial relations from visualized long-term optical flow frames and design a temporal aggregation block for temporal relations. Extensive experimental results using three large MER databases, both in the context of sole database evaluation and composite database evaluation settings and the Leave One Subject Out cross validation protocol, consistently demonstrate that our approach is effective and outperforms the current state of the art. These findings strongly motivate further research on the use of transformer based architectures rather than convolutional neural networks in micro-expression analysis, and we hope that our theoretical contributions will help direct such future efforts.

Acknowledgments

The authors would like to thank the China Scholarship Council – University of St Andrews Scholarships (No.201908060250) funds L. Zhang for her PhD. This work is funded by the National Key Research and Development Project of China under Grant No. 2019YFB1312000, the National Natural Science Foundation of China under Grant No. 62076195, and the Fundamental Research Funds for the Central Universities under Grant No. AUGA5710011522.

References

[1] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” *Journal of Personality and Social Psychology*, 1971.

[2] L. Zhang, O. Arandjelović, S. Dewar, A. Astell, G. Doherty, and M. Ellis, “Quantification of advanced dementia patients’ engagement in therapeutic sessions: An automatic video based approach using computer vision and machine learning,” in *Proceedings International Conference of the IEEE Engineering in Medicine & Biology Society*. IEEE, 2020, pp. 5785–5788.

[3] L. A. Gottschalk, A. H. Auerbach, E. A. Haggard, and K. S. Isaacs, “Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy,” in *Methods of Research in Psychotherapy*, 1966.

[4] P. Ekman and W. V. Friesen, “ Nonverbal Leakage and Clues to Deception,” *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[5] L. Zhang and O. Arandjelović, “Review of Automatic Micro-expression Recognition in the Past Decade,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 414–434, 2021.

[6] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, “Recognising spontaneous facial micro-expressions,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[7] G. Zhao and M. Pietikäinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[8] S. Polikovsky, Y. Kameda, and Y. Ohta, “Facial micro-expressions recognition using high speed camera and 3D-Gradient descriptor,” in *IET Seminar Digest*, 2009.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3213509

[9] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, 2018.

[10] 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*, 2016.

[11] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *Proceedings of 23rd international conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 2258–2263.

[12] H. Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *Proceedings of 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, 2018.

[13] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Analysis and Applications*, vol. 22, no. 4, pp. 1331–1339, 2019.

[14] Z. Xia, X. Feng, X. Hong, and G. Zhao, "Spontaneous facial micro-expression recognition via deep convolutional network," in *Proceedings of 8th International Conference on Image Processing Theory, Tools and Applications, IPTA*. Institute of Electrical and Electronics Engineers Inc., 2019.

[15] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-Expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, 2020.

[16] L. Lo, H. X. Xie, H. H. Shuai, and W. H. Cheng, "MER-GCN: Micro-Expression Recognition Based on Relation Modeling with Graph Convolutional Networks," in *Proceedings of 3rd International Conference on Multimedia Information Processing and Retrieval*, 2020.

[17] A. M. Buhari, C.-P. Ooi, V. M. Baskaran, R. C. W. Phan, K. Wong, and W.-H. Tan, "FACS-Based Graph Features for Real-Time Micro-Expression Recognition," *Journal of Imaging*, vol. 6, no. 12, 2020.

[18] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition," 2020.

[19] A. J. R. Kumar and B. Bhanu, "Micro-Expression Classification Based on Landmark Relations With Graph Attention Convolutional Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[20] L. Lei, T. Chen, S. Li, and J. Li, "Micro-Expression Recognition Based on Facial Graph Representation Learning and Facial Action Unit Fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *arXiv preprint arXiv:2101.01169*, 2021.

[22] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 494–10 503.

[23] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and Others, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.

[26] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[27] R. Girdhar, J. Joao Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[29] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," *arXiv preprint arXiv:2010.04159*, 2020.

[30] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[31] S. T. Liong, J. See, K. S. Wong, and R. C. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, 2018.

[32] Y. S. Gan, S. T. Liong, W. C. Yau, Y. C. Huang, and L. K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Processing: Image Communication*, 2019.

[33] Y. Li, X. Huang, and G. Zhao, "Joint Local and Global Information Learning With Single Apex Frame Detection for Micro-Expression Recognition," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 30, 2021.

[34] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A Spontaneous Micro-expression Database: Inducement, collection and baseline," in *Proceedings of 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG*, 2013.

[35] W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, 2014.

[36] 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*, 2018.

[37] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System - Investigator's Guide*, 2002.

[38] L. Zhang, O. Arandjelović, and X. Hong, "Facial Action Unit Detection with Local Key Facial Sub-region Based Multi-label Classification for Micro-expression Analysis," in *ACM international conference on Multimedia Workshops*, 2021.

[39] S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, and C.-G. Zhou, "Micro-Expression Recognition Using Robust Principal Component Analysis and Local Spatiotemporal Directional Features," in *European Conference on Computer Vision*, vol. 8925, 2014, pp. 325–338.

[40] M. Peng, C. Wang, T. Bi, Y. Shi, X. Zhou, and T. Chen, "A novel apex-time network for cross-dataset micro-expression recognition," *Proceedings of 8th International Conference on Affective Computing and Intelligent Interaction, ACII*, 2019.

[41] S. T. Liong, Y. S. Gan, J. See, H. Q. Khor, and Y. C. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.

[42] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *Proceedings of 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, 2019.

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

[44] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," *Proceedings of the 2016 ACM Multimedia Conference*, pp. 382–386, 2016.

[45] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-Trained Image Processing Transformer," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. T. Google, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[48] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2020.

[49] Y. Wang, W.-B. Jiang, R. Li, and B.-L. Lu, "Emotion transformer fusion: Complementary representation properties of eeg and eye movements on recognizing anger and surprise," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1575–1578.

[50] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 7680–7689.

[51] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, 2021.

[52] D.-S. Pham, O. Arandjelović, and S. Venkatesh, "Detection of dynamic background due to swaying movements from motion features," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 332–344, 2014.
[53] O. Arandjelović, D.-S. Pham, and S. Venkatesh, "Cctv scene perspective distortion estimation from low-level motion features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 5, pp. 939–949, 2015.
[54] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, "Learning deep transformer models for machine translation." Association for Computational Linguistics, 2019, pp. 1810–1822.
[55] A. Baevski and M. Auli, "Adaptive input representations for neural language modeling," 2019.
[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
[57] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2017.
[58] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.
[59] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, 2009.
[60] S. J. Wang, B. J. Li, Y. J. Liu, W. J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, 2018.
[61] H. Zhewei, Z. Tianyuan, H. Wen, S. Boxin, and Z. Shuchang, "RIFE: Real-time intermediate flow estimation for video frame interpolation," *arXiv preprint arXiv:2011.06294*, 2020.
[62] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009.
[63] J. See, M. H. Yap, J. Li, X. Hong, and S. J. Wang, "Megc 2019 - the second facial micro-expressions grand challenge." IEEE, 5 2019.
[64] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using Spatiotemporal Completed Local Quantized Patterns," *Neurocomputing*, 2015.
[65] H. Q. Khor, J. See, S. T. Liong, R. C. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," 2019.
[66] X. Huang, S. J. Wang, G. Zhao, and M. Piteikainen, "Facial Micro-Expression Recognition Using Spatiotemporal Local Binary Pattern with Integral Projection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
[67] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Transactions on Multimedia*, 2018.
[68] C. Wang, M. Peng, T. Bi, and T. Chen, "Micro-attention for micro-expression recognition," *Neurocomputing*, 2020.
[69] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic Micro-Expression Recognition Using Knowledge Distillation," *IEEE Transactions on Affective Computing*, 2020.
[70] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "Geme: Dual-stream multi-task gender-based micro-expression recognition," *Neurocomputing*, vol. 427, 2021.
[71] Z. Xia, W. Peng, H. Q. Khor, X. Feng, and G. Zhao, "Revealing the Invisible with Model and Data Shrinking for Composite-Database Micro-Expression Recognition," *IEEE Transactions on Image Processing*, vol. 29, 2020.
[72] N. Van Quang, J. Chun, and T. Tokuyama, "CapsuleNet for micro-expression recognition," in *Proceedings of 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, 2019.
[73] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Proceedings of 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG*. Institute of Electrical and Electronics Engineers Inc., 2019.
[74] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, "Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[75] X. Li, S. Cheng, Y. Li, M. Behzad, J. Shen, S. Zafeiriou, M. Pantic, and G. Zhao, "4dme: A spontaneous 4d micro-expression dataset with multimodalities," *IEEE Transactions on Affective Computing*, 2022.

Liangfei Zhang received the B.Eng. degree in 2018, from the School of Computer Science and Technology, Northwestern Polytechnical University, P.R.China. She is currently a Ph.D. student in the School of Computer Science, University of St Andrews, where she received her MSc Artificial Intelligence degree in 2019. Her research interests are artificial intelligence, computer vision and pattern recognition, and their application such as affective computing, facial behaviour analysis and expression recognition.

Xiaopeng Hong is a professor at Harbin Institute of Technology (HIT), PRC. He had been a professor at Xi'an Jiaotong University, P. R. China, and an adjunct professor with the University of Oulu, Finland. Xiaopeng received his Ph.D. degree in computer application and technology at HIT, in 2010. He has authored over 50 articles in top-tier publications and conferences such as IEEE T-PAMI, CVPR, ICCV, and AAAI. He has served as an area chair/senior program committee member for ACM MM, AAAI, IJCAI, and ICME, a guest editor for peer-reviewed journals like Patter Recognition Letter and Signal, Image and Video Processing, a co-organizer for six international workshops in conjunction with IEEE CVPR, ACM MM, IEEE FG, and a co-lecturer for two tutorials in conjunction with ACM MM21 and IJCB21. His studies about subtle facial movement analysis have been reported by International media like MIT Technology Review and been awarded the 2020 IEEE Finland Section best student conference paper.

Ognjen Arandjelović graduated top of his class from the Department of Engineering Science at the University of Oxford (M.Eng.). In 2007 he was awarded his Ph.D. by the University of Cambridge where he stayed thereafter as Fellow of Trinity College Cambridge. Currently he is a Reader in the School of Computer Science at the University of St. Andrews in Scotland. Ognjen's main research interests are computer vision and pattern recognition, and their application in various fields of science such as bioinformatics, medicine, physiology, etc. He is a Fellow of the Cambridge Overseas Trust, winner of numerous awards, and Area Editor in Chief of Pattern Recognition and Associate Editor of Information, Cancers, and Frontiers in AI.

Guoying Zhao (IEEE Fellow) is currently an Academy Professor with Academy of Finland and (tenured) full professor (from 2017) with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences and then joined University of Oulu as a senior researcher. She was Academy Fellow in 2011-2017 and an Associate Professor from 2014 to 2017 with University of Oulu. She has authored or co-authored more than 280 papers in journals and conferences. Her papers have currently over 18390 citations in Google Scholar (h-index 65). She has served as associate editor for Pattern Recognition, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Trans. on Multimedia and Image and Vision Computing Journals. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics. Her research has been reported by Finnish national TV, newspapers and MIT Technology Review.