Spatio-Temporal Analysis of Facial Actions using Lifecycle-Aware Capsule Networks

3 authors:

Nikhil Churamani
University of Cambridge
29 PUBLICATIONS 184 CITATIONS

SEE PROFILE

Sinan Kalkan
Middle East Technical University
109 PUBLICATIONS 1,244 CITATIONS

SEE PROFILE

Hatice Gunes
University of Cambridge
133 PUBLICATIONS 4,030 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project PhD Research View project

Project Intelligent subgrouping of multitrack audio View project

Spatio-Temporal Analysis of Facial Actions using Lifecycle-Aware Capsule Networks

Nikhil Churamani*, Sinan Kalkan† and Hatice Gunes*

*Department of Computer Science and Technology, University of Cambridge, United Kingdom
†Department of Computer Engineering, Middle East Technical University, Ankara, Turkey

{nikhil.churamani, hatice.gunes}@cl.cam.ac.uk, skalkan@metu.edu.tr

Abstract

Most state-of-the-art approaches for Facial Action Unit (AU) detection rely upon evaluating facial expressions from static frames, encoding a snapshot of heightened facial activity. In real-world interactions, however, facial expressions are usually more subtle and evolve in a temporal manner requiring AU detection models to learn spatial as well as temporal information. In this paper, we focus on both spatial and spatio-temporal features encoding the temporal evolution of facial AU activation. For this purpose, we propose the Action Unit Lifecycle-Aware Capsule Network (AULA-Caps) that performs AU detection using both frame and sequence-level features. While at the frame-level the capsule layers of AULA-Caps learn spatial feature primitives to determine AU activations, at the sequence-level, it learns temporal dependencies between contiguous frames by focusing on relevant spatio-temporal segments in the sequence. The learnt feature capsules are routed together such that the model learns to selectively focus more on spatial or spatio-temporal information depending upon the AU lifecycle. The proposed model is evaluated on the commonly used BP4D and GFT benchmark datasets obtaining state-of-the-art results on both the datasets.

1. Introduction

Analysing facial expressions can be highly subjective and influenced by different contextual and cultural variations [17]. To establish constants across varying cultural contexts and achieve objective evaluations for facial expressions, Ekman *et al.* [9, 10] developed the Facial Action Coding System (FACS). Facial actions, that is, the contraction and relaxation of facial muscles are encoded as 'activated' facial Action Units (AUs) that can be used to describe different facial expressions. As FACS only encodes the activation of facial muscles, no subjective or context-sensitive affective understanding is needed. Co-activation of differ-

ent AUs reveals local relationships and hierarchies where multiple facial muscles combine to form an expression, for example, raised eyebrows (involving AUs 1, 2) and jaw-drop (AU 26) together signify *surprise* [9] or communicate an intention (for example, AU 46 for winking).

Furthermore, facial muscle activation follows a temporal evolution [26, 32], referred to in this paper as the *AU Lifecycle*. Starting from a relaxed and *neutral* resting state, facial muscles start to contract, forming the *onset* of an expression with complete contraction achieved in the *apex* state to express peak intensity. This is followed by the relaxation of the muscles forming the *offset* state before returning back to *neutral*. This process may also be repeated several times for certain expressions, for example, spontaneous smiles typically have multiple *apices* with a much slower *onset* phase [6]. Understanding this evolution is essential to understand how humans express affect and is particularly crucial for distinguishing *posed* or *acted* expressions from *spontaneous* expressions [38].

Computational models for AU detection, traditionally, have explored local spatial relationships and hierarchies between different face regions using shape-based representations or using spectral or histogram-based methods [32, 44]. With deep learning gaining popularity, recent approaches [4, 14, 21, 35, 36, 46, 47] have applied convolution or graph-based models to focus on learning such facial features directly from data, outperforming most traditional approaches. More recently, capsule-based computations proposed by Sabour *et al.* [31] have further improved the learning of spatial dependencies in the form of facial feature *primitives*. These feature primitives are sensitive to local variations and capture local dependencies between different facial regions and have been successfully applied for AU detection and expression recognition tasks [12, 29, 30].

Most approaches, however, only focus on frame-based evaluation of facial actions, relying on analysing peak-intensity frames [23, 32]. As a result, even though these approaches are able to detect strong AU activations in *posed* settings or when an expression is highly *accentuated*, they

suffer when detecting more *subtle* expressions in *spontaneous* and naturalistic settings [38, 42], challenging their real-world applicability. A prevailing requirement for *automatic* AU detection is to be sensitive to the said *AU lifecycle* and include temporal information, such as motion features or correlations amongst proximal frames, along with spatial features [19, 36, 42]. While spatial processing is important to determine relationships between different facial regions [19], understanding temporal correlations between their activation patterns in contiguous frames provides essential information about the *AU lifecycle* and can be particularly useful in detecting subtle activations [4, 36, 42].

Leveraging the ability of capsule networks to learn relationships between local spatial and temporal features, in this work, we propose the Action Unit Lifecycle-Aware Capsule Network (AULA-Caps) for multi-label AU detection (see Figure 1). AULA-Caps is a multi-stream, lifecycle-aware capsule network trained in an end-to-end manner that not only learns spatial activation patterns within a frame but also their dynamics across several contiguous frames. To the best of our knowledge, this is the first work combining multiple capsule-based processing streams that learn spatial and spatio-temporal facial features at frame and sequence-level, simultaneously. We perform benchmark evaluations on the GFT [13] and BP4D [45] datasets comparing AULA-Caps performance with the state-of-the-art methodologies on AU detection. AULA-Caps achieves the best F1-score for AUs $2, 7, 17$ and 23 with the second-best overall (average) F1-score for the GFT dataset while achieving the best F1-scores for AUs $1, 6$ and 17 along with the best overall F1-score for the BP4D dataset.

2. Related Work

2.1. Spatial Analysis for AU Prediction

AU detection approaches rely on capturing spatial relationships between different face regions [27, 28, 32]. Popular methods include using geometric features that track facial landmarks [7], histogram-based approaches that cluster local features into uniform regions for processing [32] or using features that describe local neighbourhoods [2]. With the popularity of deep learning, CNN [11, 14, 21] and graph-based [19, 36] approaches have achieved state-of-the-art results for AU detection due to their ability to hierarchically learn spatial features. Capsule-based computations [31] offer an improvement as along with learning to detect different facial features, they also learn how these are arranged with respect to each other. Recent work [12, 30] has explored capsule-based computations for AU detection by learning facial features that capture variations with respect to pose and/or orientation. Yet, relying only on spatial features ignores how AU activations evolve over time, impacting performance on automatic AU detection [42].

2.2. Spatio-Temporal Analysis for AU Prediction

Learning dynamic spatio-temporal features provides information about the dynamics of AU activations. A straightforward way for computing spatio-temporal features is to extract spatial features from each frame separately and use recurrent models such as the LSTM [16] to learn how these evolve with time [4]. Alternatively, models may compute temporal features, such as optical flow, first and then process them using CNN-based networks [1]. Yet, most of these approaches focus on learning spatial and temporal information sequentially. Yang *et al*. [42] propose an alternative by concurrently learning spatial and temporal features, inspired by human AU coders. However, their approach focuses on extracting spatio-temporal features from complete video sequences at once, dropping certain adjacent frames to ensure all video sequences are of the same length. Other more recent approaches learn semantic relationships between different face regions and represent these using structured knowledge-graphs, learning coupling patterns between regions using graph-based computations [19, 36].

2.3. Capsule Networks

Sabour *et al*. [31] proposed the Capsule Networks that learn spatial dependencies in the form of feature *primitives* by extracting features corresponding to the different regions of an input image and learning how they combine together to contribute towards solving a particular task. This ability to learn local features and their inter-dependencies makes them a good fit for AU detection. Ertugrul *et al*. [12] propose the FACSCaps model that employs capsule networks to learn pose-independent spatial feature representations from multi-view facial images for AU detection. Rashid *et al*. [30] use capsule networks consisting of multiple convolutional operations to extract relevant spatial features from static frames before *routing* them together to obtain fully-connected class capsules. A similar approach is also employed by Quang *et al*. [29] applying capsule networks for micro expression recognition. These approaches, however, focus only on learning spatial features from static images.

Capsule networks have also been applied for video-based action recognition [8] that use 3D capsules for segmenting and tracking objects across frames. However, they explore temporal relations between frames only for segmentation and ignore how these might contribute towards sequence-based predictions. Jayasekara *et al*. [18], on the other hand, apply capsule-based learning for time-series predictions learning to classify 1D ECG signals by focusing on temporal dependencies.

In this paper, we propose a multi-stream approach that applies capsule-based computations both at frame and sequence-level, concurrently learning spatial as well as spatio-temporal dependencies from sequences of contiguous frames.
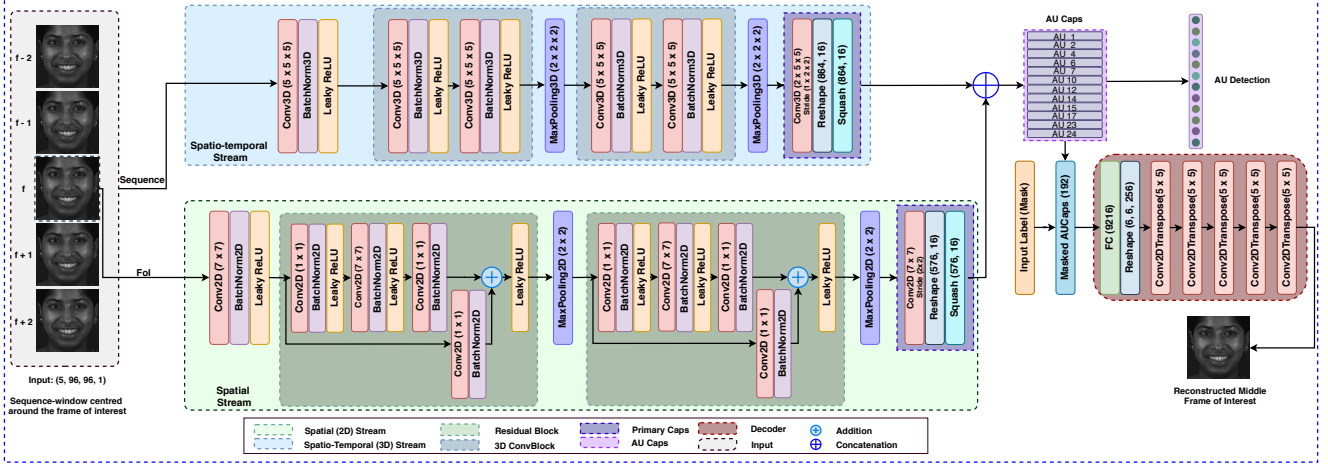
Figure 1: Action Unit Lifecycle-Aware Capsule Network (AULA-Caps) for Multi-label Facial Action Unit Detection.

3. Action Unit Lifecycle-Aware Capsule Network (AULA-Caps)

We propose the AULA-Caps network (see Figure. 1) that processes face-image sequences using two separate streams for computing spatial (2D) and spatio-temporal (3D) features. While spatial processing of a Frame-of-Interest (FoI), in this case, the middle frame from each input sequence, focuses on local spatial dependencies, spatio-temporal processing investigates contiguous frames to extract features that capture dynamics of AU activations. Both streams employ capsule-based computations with the extracted individual primary capsules combined and *routed* together to evaluate their influence on final class-capsules. The class-capsules are also passed to a decoder that learns to reconstruct the middle FoI, further regularising learning.

3.1. Windowed Video Sequences as Input

AULA-Caps takes as input a video sequence consisting of contiguous (96×96) frames of normalised face-centred images (each pixel $p \in [-1, 1]$). Input sequences are generated by taking each frame of the video, along with N frames immediately preceding and succeeding the frame. The middle FoI is passed to the spatial processing stream while the entire window of $2N + 1$ frames is processed using the spatio-temporal processing stream. The overall task for the model is to predict the activated AUs in the FoI. Here, we set N=2 forming an input window of 5 frames.

3.2. Motivation for Lifecycle-Awareness

Following the AU lifecycle, different segments of activation, that is, *onset*, *apex* and *offset* form the evolution of an AU. In the *onset* and *offset* phases, the input images, or the extracted facial features, have high *variation*, in that, the contiguous frames are sufficiently different, as illustrated in Figure 2a. Thus, focusing on this difference provides im-

(a) Onset Segment sample from BP4D.

(b) Apex Segment sample from BP4D.

Figure 2: Onset and Apex segment contiguous frames.

portant temporal information about AU activation patterns. In *apex* segment frames, however, the contiguous frames have low *variation* and are not sufficiently different, as illustrated in Figure 2b. As a result, instead of focusing on temporal changes, spatial features extracted from just a single FoI provide sufficient information for AU prediction.

The two streams of processing in the AULA-Caps model are designed to exploit this difference by extracting relevant spatial and spatio-temporal features and combining them in a manner where their individual contribution is weighted based on their relevance for AU prediction. Realising such an ability for the model to selectively and automatically tune into separate features based on where in the AU lifecycle the input sequence originates from, motivates the *lifecycle-awareness* of the AULA-Caps model.

3.3. Computing Spatial Features

The spatial processing stream (see Spatial Stream in Figure 1) takes as input the middle FoI (f) from an input sequence and passes it through a convolutional (conv) layer with 128 filters of size (7×7) followed by *BatchNormalisation* and a *LeakyReLU* ($\alpha = 0.2$) activation. The output of this conv layer is passed through two Residual blocks consisting of multi-resolution conv layers with shortcut con-

Table 1: Action Units examined in this work.

AU	Description	AU	Description	AU	Description
1	Inner Brow Raiser	7	Eyelid Tightener	15	Lip Corner Depressor
2	Outer Brow Raiser	10	Upper Lip Raiser	17	Chin Raiser
4	Brow Lowerer	12	Lip Corner Puller	23	Lip Tightener
6	Cheek Raiser	14	Dimpler	24	Lip Pressor

nections [15], with 128 and 64 filters for each conv layer in the respective blocks using a *LeakyReLU* ($\alpha = 0.2$) activation. Each residual block is followed by a (2×2) maxpooling layer for dimensionality reduction. The output of the final maxpooling layer is passed to the Primary Capsule layer consisting of a conv layer followed by reshaping and squashing of the extracted spatial features into 576 capsules of 16 dimensions each.

3.4. Computing Spatio-Temporal Features

The spatio-temporal processing stream (see Spatio-temporal Stream in Figure 1) processes the entire input window sequence. The sequence is first passed through a 3DConv layer consisting of 128 filters of size ($5 \times 5 \times 5$) followed by *BatchNormalisation* and a *LeakyReLU* ($\alpha = 0.2$) activation. The output of the 3DConv layer is passed through two 3DConv blocks consisting of two conv layers each followed by *BatchNormalisation* and *LeakyReLU* ($\alpha = 0.2$) activation. Conv layers in the first block consist of 128 filters of size ($5 \times 5 \times 5$) while the second block layers consist of 64 filters. Each block is followed by a ($2 \times 2 \times 2$) 3D maxpooling layer for dimensionality reduction. The final maxpooling layer output is passed to the 3D Primary Capsule layer consisting of a 3DConv layer followed by reshaping and squashing of the extracted spatio-temporal features into 864 capsules of 16 dimensions each.

3.5. Combining Extracted Features

The extracted *primary* capsules representing spatial and spatio-temporal primitives computed by the two streams are concatenated together resulting in 1440 capsules of 16 dimensions each. The iterative *routing-by-agreement* algorithm [31] then couples the concatenated capsules with the AU Capsule Layer (AU-Caps), computing 12 capsules each corresponding to an AU label (see Table 1). The output of the AU-Caps layer is used to predict the AUs activated in the FoI, replacing the computed capsule with its length *squashed* between $[0, 1]$ depicting the activation probability for each AU label. The AU-Caps layer output is also used by the Decoder model to reconstruct the FoI.

3.6. Decoder for Image Reconstruction

The Decoder is used to regularise learning in the model [31] making sure it learns task-relevant features, as well as to enable visualisation of learnt features through

the reconstructed images. The AU-capsules are masked using the label *y* for reconstructing the FoI (x_{gen}). In AULA-Caps, we use transposed conv layers for the decoder, instead of dense layers proposed by Sabour *et al.* [31]. This significantly reduces the number of parameters in the decoder (≈ 2.8M as opposed to 10M in [31]) while also improving the photo-realistic quality of reconstructed images. The decoder, adapted from the generator of [5], implements 4 stacked conv layers, using *ReLU* activation, performing transposed convolutions with $128, 64, 32, 16$ filters, respectively, of size (5×5) each with a stride of (2×2). The output of the last conv layer is passed through another transposed conv layer using *tanh* activation to generate the resultant image (x_{gen}) with the same dimensions as the FoI (x_r).

3.7. Learning Objectives

The two streams of AULA-Caps, along with the decoder, are trained together in an end-to-end manner. The AULA-Caps model predicts 2 outputs in each run, that is, the activation probabilities for 12 AUs (see Table 1) as well as the reconstructed FoI. The learning objectives for the model are as follows:

AU Prediction: The AULA-Caps model predicts the activation probabilities for the 12 AUs in the FoI as the length of the AU-class capsules. Learning to detect the activated AUs focuses on minimising a *weighted* margin loss. The loss for each of the AUs ($\mathcal{L}_{au}$) is defined as:

$$\mathcal{L}_{au} = w_{au}(T_{au} \max(0, m^+ - ||p_{au}||)^2 \\ + \lambda_{au}(1 - T_{au}) \max(0, ||p_{au}|| - m^-)^2), \quad (1)$$

where $T_{au} = 1$ if an AU is present and 0 otherwise, $||p_{au}||$ is the prediction (output probability) for an AU computed as the magnitude (length) of the respective class capsule, m^+ and m^- are the positive and negative sample margins, λ_{au} is a constant weighting the effect of positive and negative samples, and w_{au} is a class balancing weight. We set $m^+ = 0.9$, $m^- = 0.1$ and $\lambda_{au} = 0.5$, following [31]. w_{au} is computed using the occurrence-rate for the respective AUs in the training data. This is done to reduce the effect of the class imbalance under multi-label classification settings. Following [33], w_{au} is computed as follows:

$$w_{au} = \frac{(1/r_i)N}{\sum_i^N (1/r_i)}, \quad (2)$$

where N is the number of AUs (in this case $N = 12$) and r_i is the occurrence rate of AU$_i$. The resultant loss ($\mathcal{L}_{margin}$) is computed as the sum of the losses for each AU ($\mathcal{L}_{au}$).

Image Reconstruction: The Decoder reconstructs the FoI using the extracted AU capsules imposing a mean squared error *reconstruction loss* ($\mathcal{L}_{rec}$):

$$\min_{X_r, X_{gen}} \mathcal{L}_{rec} = L_2(x_r, x_{gen}), \quad (3)$$

where x_r is the FoI and x_{gen} is the reconstructed image.

Overall Objective: The overall objective for AULA-Caps is a weighted sum of the overall AU prediction ($\mathcal{L}_{margin}$) and image reconstruction ($\mathcal{L}_{rec}$) objectives:

$$\mathcal{L}_{AULA} = \mathcal{L}_{margin} + \lambda_d \mathcal{L}_{rec}, \qquad (4)$$

where λ_d is set to 0.05 balance the loss terms.

4. Experiments

4.1. Datasets

We train and test AULA-Caps on two popular AU benchmark datasets, namely GFT and BP4D. For both datasets, samples representing the 12 most frequently occurring AUs are used (see Table 1).

GFT: The Sayett Group Formation Task (GFT) Dataset [13] consists of 1-minute video recordings from 96 subjects, spontaneously interacting with each other in group settings ($2 - 3$ persons per group). The interactions are unstructured, allowing for natural and spontaneous reactions by the participants, annotated for each group-member at frame-level. Annotations are made available at frame-level.

BP4D: The BP4D dataset [45] consists of video sequences from 41 subjects performing 8 different affective tasks to elicit emotional reactions. Approximately 500 frames for each video are annotated for occurrence and intensity of the activated AUs. In our experiments, we only use occurrence labels for AU detection.

Both BP4D and GFT represent different data settings, enabling a comprehensive evaluation of the proposed model. While GFT represents complex, naturalistic recording settings, BP4D, on the other hand, consists of cleaner, face-centred images and provides much more data per subject.

4.2. Experiment Settings

Evaluation Metric: Similar to other approaches [11, 23, 36], we follow $3-$fold cross-validation for our evaluations, splitting the data into 3 folds where each subject occurs in test-set once. For each run, the model is trained on 2 folds and tested on the third. Results are collated across the 3 folds. We report model performance using *F1-Scores* computed as the harmonic mean (F1= $\frac{2RP}{R+P}$) of the precision (P) and recall (R) scores, providing for a robust evaluation of the model. F1-score is the most commonly employed metric for reporting AU detection performance [39].

Implementation Details: The AULA-Caps model is implemented using Keras-Tensorflow[1]. We train the model individually on each dataset in an end-to-end manner using

the *Adam* optimiser with an initial learning rate of $2.0e^{-4}$ and decayed each epoch by a factor of 0.9. For each fold, the model is trained for 12 epochs with early stopping with a batch-size of 24. The model hyper-parameters namely, the filter sizes, number of filters for each layer, capsule dimension size, the batch-size and the initial learning-rate are optimised using the Hyperopt[2] Python Library. The models are trained on a NVIDIA GeForce GTX 1080Ti GPU.

4.3. Results

GFT: Table 2 presents the model performance results on the GFT dataset in comparison to the state-of-the-art. We compare AULA-Caps to different spatial and spatio-temporal approaches such as the CNN-based cross-domain learning [CRD] [11], an Alex-Net-based model [ANet] for frame-based AU detection [4], the [JÂA] [34] approach that uses multi-scale high-level facial features extracted from face alignment tasks to aid AU prediction, and learning temporal variation in facial features using a CNN-LSTM model [4]. The [CNN-LSTM] model applies frame-based spatial computations and extends this learning to the temporal domain by evaluating how spatial features evolve over time. In contrast, the proposed AULA-Caps model simultaneously extracts spatial and spatio-temporal features from input sequences and learns to combine them to selectively focus on relevant features for respective AU predictions.

AULA-Caps achieves the best results for 4 AU labels and second-best results for another 3. Despite achieving the second-best overall results (-0.002 difference in *Avg.* F1-score from the best approach LSTM), the model performs rather poorly for AU $1, 4$ and 14 impacting the overall score of the model. We discuss these results in Section 5.2.

BP4D: Table 4 presents the AULA-Caps results for BP4D and compares them to the state-of-the-art approaches such as the [CNN-LSTM] [4] learning temporal variation in facial features, the [EAC] method [21] that employs enhancing and cropping mechanism to focus on selective regions in an image, the [ROI] network [20] that focuses on learning regional features using separate local CNN, a 2D Capule-Net based model [CapsNet] proposed by [30], the [JÂA] [34] approach that uses multi-scale high-level facial features, the semantic learning-based [SRERL] [19] model and the [STRAL] [36] approach that employs a spatio-temporal graph convolutional network to capture both spatial and temporal relations for AU prediction. The AULA-Caps model, on the other hand, uses a multi-stream approach that simultaneously learns and combines spatial as well as spatio-temporal features making it sensitive to the temporal evolution of AU activations.

AULA-Caps achieves the best results for 3 AU labels and second-best results for another 3. Overall, the model

---

Table 2: Performance Evaluation (F1-Scores) on GFT Dataset: CRD [11], ANet [4], JÂA [34], CNN-LSTM [4]. **Bold** values denote best while [*bracketed*] denote second-best values for each row. All scores reported from the respective papers. *Averaged for 10 AUs.*

AU	CRD	ANet	JÂA	CNN-LSTM	AULA-Caps
1	[*0.437*]	0.312	**0.465**	0.299	0.313
2	0.449	0.292	[*0.493*]	0.257	**0.498**
4	0.198	**0.719**	0.192	[*0.689*]	0.297
6	0.746	0.645	**0.790**	0.673	[*0.775*]
7	0.721	0.671	–	[*0.725*]	**0.772**
10	**0.765**	0.426	[*0.75*]	0.670	0.749
12	[*0.798*]	0.731	**0.848**	0.751	0.785
14	0.500	[*0.691*]	0.441	**0.807**	0.236
15	0.339	0.279	0.335	**0.435**	[*0.371*]
17	0.170	[*0.504*]	–	0.491	**0.592**
23	0.168	0.348	**0.549**	0.350	[*0.522*]
24	0.129	0.390	[*0.507*]	0.319	**0.530**
Avg.	0.452	0.500	0.537*	**0.539**	[*0.537*]

Table 4: Performance Evaluation (F1-Scores) on BP4D Dataset: CNN-LSTM [4], EAC [21], ROI [20], CapsNet [30], JÂA [34], SRERL [19], STRAL [36]. **Bold** values denote best while [*bracketed*] denote second-best values for each row. All scores reported from the respective papers.

AU	CNN-LSTM	EAC	ROI	CapsNet	JÂA	SRERL	STRAL	AULA-Caps
1	0.314	0.390	0.362	0.468	[*0.538*]	0.469	0.482	**0.562**
2	0.311	0.352	0.316	0.291	**0.478**	0.453	[*0.477*]	0.465
4	**0.714**	0.486	0.434	0.529	[*0.582*]	0.556	0.581	0.573
6	0.633	0.761	0.771	0.753	[*0.785*]	0.771	0.758	**0.796**
7	0.771	0.729	0.737	0.776	0.758	**0.784**	[*0.781*]	0.765
10	0.450	0.819	**0.850**	0.824	0.827	0.835	0.816	[*0.843*]
12	0.826	0.862	0.870	0.850	**0.882**	[*0.876*]	[*0.876*]	0.874
14	**0.729**	0.588	0.626	0.657	0.637	0.639	0.605	[*0.718*]
15	0.340	0.375	0.457	0.337	0.433	**0.522**	[*0.502*]	0.457
17	0.539	0.591	0.580	0.606	0.618	0.639	[*0.640*]	**0.694**
23	0.386	0.359	0.383	0.369	0.456	0.471	**0.512**	[*0.495*]
24	0.370	0.358	0.374	0.431	0.499	[*0.533*]	**0.552**	0.502
Avg.	0.532	0.559	0.564	0.574	0.624	0.629	[*0.632*]	**0.645**

Table 3: Ablations on AULA-Caps for BP4D. Decoder parameters ($\approx$ 2M) excluded for comparison with CNN baselines. Batch size is set to 24.

Model	Avg. F1-Score	#Params	RunTime / Batch
2D CNN Baseline	0.573	3.44M	0.31s
3D CNN Baseline	0.540	15.09M	0.63s
2D Stream AULA-Caps	0.580	3.06M	0.35s
3D Stream AULA-Caps	0.550	8.46M	0.66s
AULA-Caps	**0.645**	11.51M	1.22s

outperforms other models, with closest *Avg.* F1-score difference to the STRAL approach [36] being 0.013 with both STRAL and the AULA-Caps model combining spatial and spatio-temporal analysis of facial features.

4.4. Ablation: Spatial *vs.* Spatio-Temporal Features

Since AULA-Caps focuses on learning spatial and spatio-temporal features simultaneously, it is important to understand how each of these feature-sets contribute towards the overall performance of the model. To evaluate the contribution of the learnt spatial features, we use the trained 2D stream layers to predict AUs by appending a separate AU-Caps layer to the primary capsule layer. The weights of the 2D stream are frozen and only the routing algorithm is run for the added AU capsule layer. Similarly, for assessing the effect of learning spatio-temporal dependencies across contiguous frames, we use the trained spatio-temporal (3D) stream layers to predict the AU labels by appending a sepa-

rate AU-Caps layer to the primary capsules.

Furthermore, for highlighting the contribution of capsule-based computation, we compare these results with 2D and 3D Convolutional Neural Network (CNN)-based models. The two streams are unchanged with only the capsule-block replaced by fully-connected dense layers. The results on BP4D for the different ablations conducted are presented in Table 3. Analysing the ablations on the BP4D provides a fairer comparison as it consists of more samples per subject with cleaner and face-centred images.

5. Analysis and Discussion

5.1. Lifecycle-Awareness

The capsule-based computations of the multi-stream Action Unit Lifecycle-Aware Capsule Network (AULA-Caps) allow it to weight the contribution of spatial and spatio-temporal feature capsules towards predicting AU activations. If spatial features are more important, for example, in the case of *apex* frames where an AU is activated with highest intensity, the model may choose to give precedence to the spatial capsules. For the off-peak intensity frames, for example, originating from the *onset* or *offset* temporal segments where the activation is low, the model may focus more on temporal differences in contiguous frames, captured using the spatio-temporal feature capsules. The ablation study results (see Table 3) highlight the individual contribution of spatial (2D) and spatio-temporal (3D) streams where a combination of both, that is, when the model learns to balance these two feature-sets, results in the best model
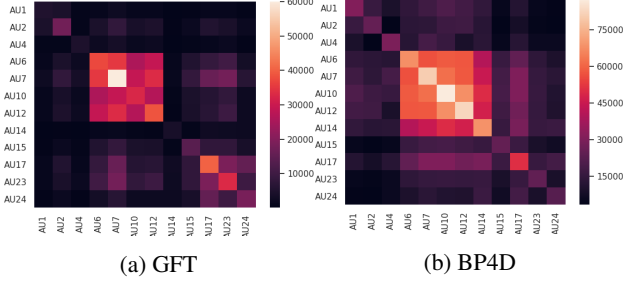
(a) GFT   (b) BP4D

Figure 3: AU co-activation heatmaps based on True Labels.

(a) Input FoI Images.   (b) Reconstructed FoI Images.

Figure 4: FoI Image reconstruction by the Decoder.

performance. This is consistent with other findings in literature where a combination of spatial and spatio-temporal features results in high performance for AU detection [36, 42]. Interestingly, the windowed computation of spatio-temporal features in AULA-Caps performs worse than spatial features, unlike other approaches [42] where 3D features perform better. This may be due to the choice of a smaller input window (5 frames in AULA-Caps) unlike [42] where an entire video is considered for computing spatio-temporal features (see Section 6.1 for a discussion).

5.2. AU Prediction

The AULA-Caps model achieves state-of-the-art results for both BP4D and GFT datasets achieving the best overall results on BP4D (see Table 4) while second-best for GFT (see Table 2). Despite the good overall performance, individual F1-scores for AUs $1, 4$ and 14 are quite poor for GFT evaluations. Investigating the data distribution for GFT by plotting the AU co-activation heatmap (see Figure 3a), we find that certain AUs dominate the data distribution. In particular, we see that AUs $6, 7, 10$ and 12 have the highest number of data samples while AUs $1, 4$ and 14, the lowest. In such an imbalanced data distribution, where AUs $1, 4$ and 14 correspond to $< 2\%$ of the total samples, the model is unable to learn relevant features that can be used for detecting these AUs. The imbalance in data correlates with the performance of the model on individual AU (see Table 2). A similar imbalance in the data distribution is also witnessed for the BP4D dataset (see Figure 3b) yet, owing to the overall larger number of samples per AU label helps mitigate some of these effects for BP4D (see Table 2). Furthermore, for the GFT dataset, subjects are recorded interacting with each other in group settings while performing a drink-tasting task which results in a lot of the recorded frames ($\approx 23\%$ of the entire dataset) being dropped and not annotated due to occlusions and varying perspectives, impacting the overall data quality as well as distribution. This also negatively impacts the overall results on the GFT database, across the state-of-the-art compared in Table 2. BP4D, on the other hand, provides cleaner and occlusion-free frames where the subjects are
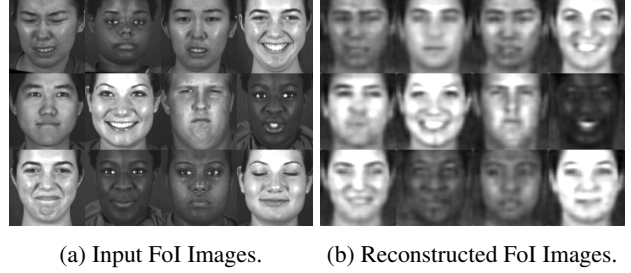
recorded mostly in face-centred videos resulting in higher performance scores across all the models compared in Table 4. AULA-Caps is able to achieve competitive scores on the GFT dataset despite its more complex and challenging settings while outperforming state-of-the-art evaluations on the BP4D dataset.

5.3. Temporal Evaluation

AU detection evaluations commonly use frame-wise performance metrics for evaluating model performance. However, for automatic AU detection, it is also important to evaluate model's performance across time. Considering the data settings in our set-up where video recordings of subjects are examined, predicting AU labels in contiguous frames can provide for a *continuous* evaluation of the model. In Figure 5, we plot, across time, the true labels as well as model predictions, that is, the length of AU class capsules for the corresponding FoIs depicting the activation probabilities for respective AUs for the 2D stream, 3D stream and the AULA-Caps model. We see that AULA-Caps predictions are able to model how the the ground-truth (AU activation) varies across time for an entire video. For example, for AU 4, we see the ground truth AU activation switching from *absent* to *activated* and then back to *absent* representing its entire lifecycle, while for AUs $6, 10, 14, 15, 17$ and 24 we see this switch occurring multiple times (multiple cycles) within the video. AULA-Caps is able to model this switch effectively, predicting AU activations efficiently.

Furthermore, Figure 5 also shows the predictions from the 2D and 3D streams where it can be seen that the 3D stream, on average, is able to better model the changing dynamics of AU activations as compared to the 2D streams, especially in regions where the ground truth switches from absent to activated or vice-versa. Yet, the 2D stream outperforms the average frame-based performance of the 3D stream across all the videos and all AUs. As frame-based performance evaluations for AU detection algorithms only report the *average* frame-based F1-scores, they ignore temporal correspondences, with respect to model performance, which are commonly examined for continuous affect perception such as arousal-valence prediction [24, 43]. Yet,
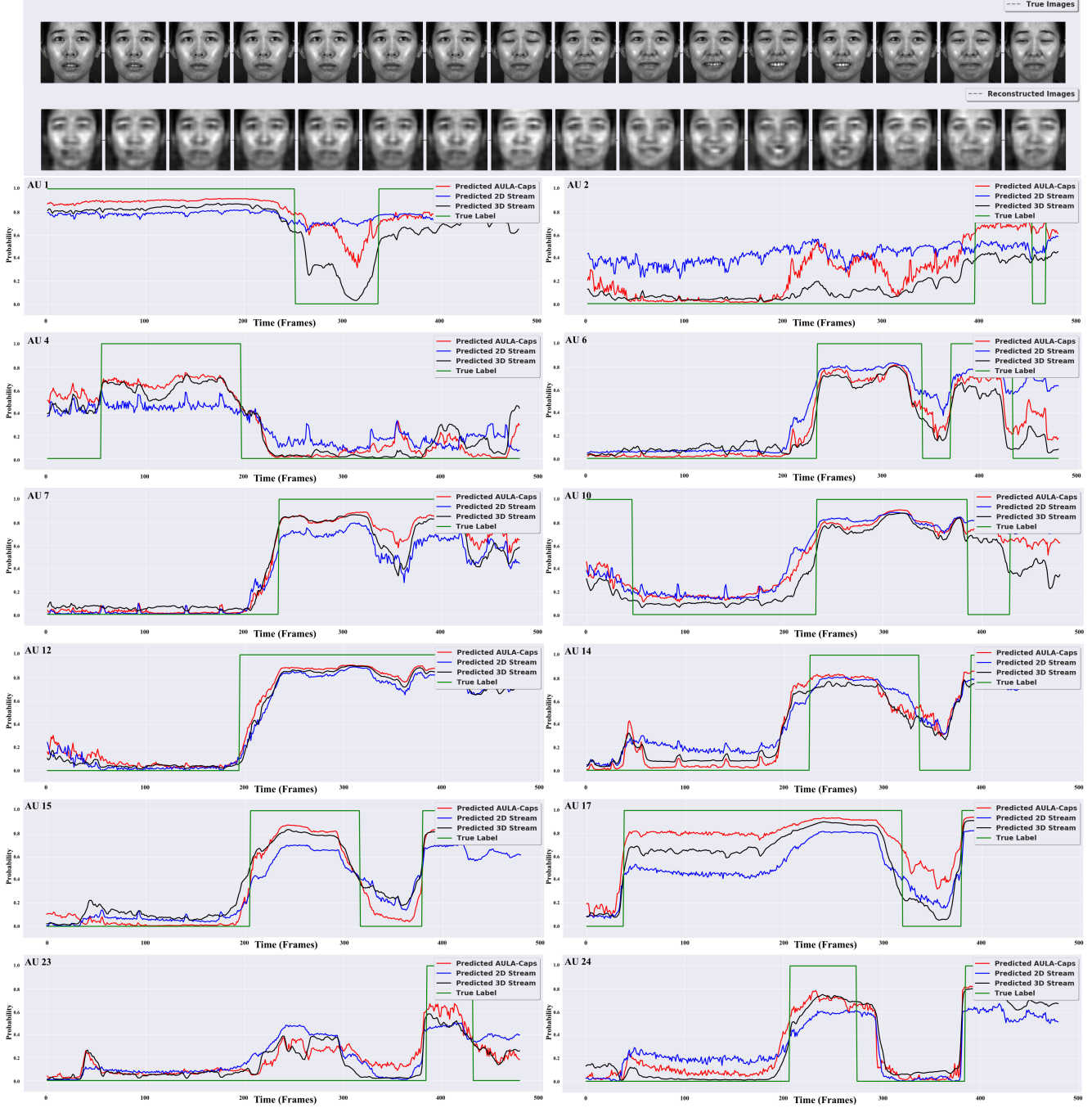
Figure 5: Comparing predictions for the 2D stream, 3D stream and AULA-Caps for the 12 AUs for a sample BP4D video.

these can be really beneficial for understanding real-time model performance, underlining its applicability for *automatic* AU prediction in real-world applications.

5.4. Visualisations

Image Reconstruction: The decoder is used to regularise learning by ensuring the model learns task-relevant features. Additionally, the reconstructed images enable a visual in-

terpretation of the learnt features. In AULA-Caps, we propose a convolution-based decoder model as opposed to the densely-connected decoder originally proposed by Sabour *et al.* [31]. The decoder is able to generate reconstructed images using a much 'lighter' network ($\approx$ 2M parameters *vs.* $\approx$ 10M [31]) without compromising on the quality of the images generated, as can be seen in Figure 4.

The data imbalance problem is witnessed in the re-

constructed images as well where FoIs for certain under-represented subjects and labels are reconstructed incorrectly. For example, for the images at (row 1, column 2) and (row 2, column 1) in both Figures 4a and 4b, the model reconstructs *generic mean* faces representing the corresponding AU labels, with a visible bias in terms of ethnicity and/or gender.

Visualising Saliency Maps: Visualising learnt features enables us to understand what the model pays attention to while making its predictions. In Figure 6, we see the Saliency Maps [37] generated by visualising the pixels in the FoIs that contribute most to model predictions. As desired, for different AUs the model learns to focus on different regions of the face, in line with Table 1. For example, for AUs 1 and 2 it focuses more on the *forehead* and *eyebrows*, for AUs 12 and 14 it focuses on *cheeks* while for AUs 23 and 24, it focuses on the *nose* and *mouth*. For certain AUs however, we see additional activity in other 'non-relevant' regions of the face as well. For example, for AU 4, we see activity in the lower region of the face near the mouth and cheeks. This can be due to the co-occurrence pattern observed in the label distribution where samples containing AU 4 also encode activity for AU 7 and AU 17, as illustrated in Figure 3. Understanding and evaluating such co-occurrence patterns can be important to improve model predictions for AU activations [19, 41] as knowing if a certain AU is activated or not can help improve model predictions for other correlated AUs.

6. Conclusion

Our experiments with the AULA-Caps highlight the importance of combining spatial and spatio-temporal feature learning for automatic AU prediction. Evaluating the temporal evolution of AU activation positively impacts model performance and allows for the dynamic evaluation of AU activity in a continuous manner. This is inline with other findings [26, 36]. Furthermore, capsule-based computations in the spatial stream enable learning local spatial relationships within an image frame corresponding to the different face regions while in the spatio-temporal stream they are able to learn temporal dependencies based on how these spatial relationships evolve across time. Combining such features allows the model to learn where to focus in an image while also being sensitive to the AU activation lifecycle.

6.1. Limitations and Future Work

Choosing the Right Window for Context: As the model evaluates AU activity across a window of input frames, it is highly sensitive to how these windows are processed. For GFT, we see that due to occlusions and complex recording conditions, several frames are dropped randomly as no
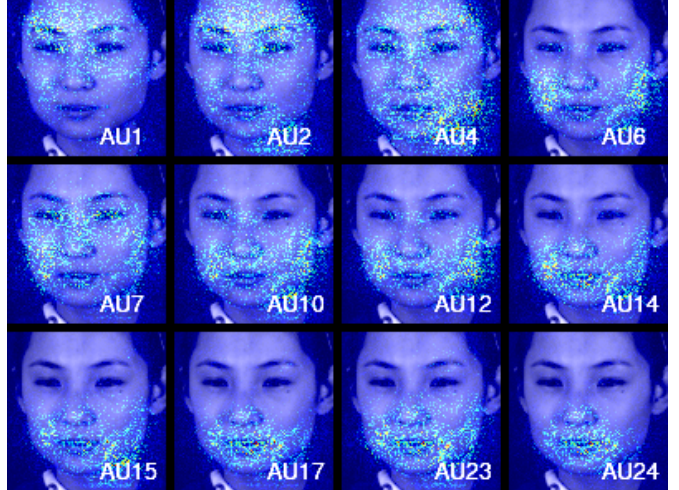

Figure 6: Saliency Maps generated using *guided backpropagation* of gradients corresponding to each AU label.

AU activity is annotated for those frames. This impacts model performance resulting in poor performance for certain AUs such as AU 1, 4 and 14. Additionally, the size of the input window may impact model performance differently for the different AUs. For some AUs the lifecycle is much longer than the others, for example, AU 12 (smile) *vs.* AU 45 (blink), and thus wider windows are expected to improve performance. In our experiments, however, we optimised the window-size based on the highest overall F1-score. Further experimentation is needed to investigate which window-sizes work best for different AUs. Also, learning to dynamically adapt such windows based on AU activity may offer improvements. Lu *et al.* [22] provide an insightful approach to address such problems by focusing on the temporal consistency in video sequences rather than relying on pre-defined window-sizes. They randomly assign *anchor frames* in input sequences and apply self-supervised learning to encode the temporal consistency of an input video sequence compared to this anchor frame. This robustly captures temporal dependencies in facial activites, improving AU detection performance.

Imbalanced Data Distributions: Another problem faced by most approaches is the imbalanced label distribution of the datasets. In Figure 3, we see that AUs 6, 7, 10 and 12 dominate the data distributions, resulting in the models performing worse on scarce labels such as AUs 1, 4 and 14. It is important to address this imbalance either at the data-level by recording evenly distributed datasets that offer a fairer comparison of models or by including mitigation strategies that handle biases arising from such imbalances [3, 25, 40]. Furthermore, understanding AU co-activations can provide additional contextual information to improve performance on scarce AU samples [19, 41, 46].

Acknowledgement

References

[1] Benjamin Allaert, Isaac Ronald Ward, Ioan Marius Bilasco, Chaabane Djeraba, and Mohammed Bennamoun. Optical flow techniques for facial expression analysis: Performance evaluation and improvements. *CoRR*, 2019.

[2] Sarah Adel Bargal, Amr Goneid, Rana el Kaliouby, and Anas Nayfeh. Classification of mouth action units using local binary patterns. In *International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, 2012.

[3] Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385 – 397, 2015.

[4] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 25–32, 2017.

[5] Nikhil Churamani and Hatice Gunes. CLIFER: Continual Learning with Imagination for Facial Expression Recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 322–328, 2020.

[6] J. F. Cohn and K. L. Schmidt. The Timing of Facial Motion in Posed and Spontaneous Smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 02(02):121–132, June 2004.

[7] Fernando De la Torre, Wen-Sheng Chu, Xuehan Xiong, Francisco Vicente, Xiaoyu Ding, and Jeffrey F. Cohn. Intraface. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8, 2015.

[8] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. VideoCapsuleNet: A Simplified Network for Action Detection. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7610–7619. Curran Associates, Inc., 2018.

[9] Paul Ekman and Wallace V. Friesen. *Facial action coding systems*. Consulting Psychologists Press, 1978.

[10] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. The Facial Action Coding System. Weidenfeld and Nicolson, 2002. 2nd ed. London, U.K.

[11] I. O. Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji. Cross-domain au detection: Domains, learning approaches, and measures. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8, 2019.

[12] I. O. Ertugrul, L. A. Jeni, and J. F. Cohn. Facscaps: Pose-independent facial action coding with capsules. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2211–2220, 2018.

[13] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, Jeffrey Cohn, Fernando De La Torre, and Michael A Sayette. Sayette group formation task (GFT) spontaneous facial expression database. In *IEEE International Conference on Automated Face & Gesture Recognition*, 2017.

[14] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–5, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997.

[17] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[18] Hirunima Jayasekara, Vinoj Jayasundara, Jathushan Rajasegaran, Sandaru Jayasekara, Suranga Seneviratne, and Ranga Rodrigo. TimeCaps: Capturing Time Series Data with Capsule Networks. *ArXiv*, abs/1911.11800, 2019.

[19] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI Conference on Artificial Intelligence*, pages 8594–8601, 2019.

[20] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6766–6775, 2017.

[21] W. Li, F. Abtahi, Z. Zhu, and L. Yin. EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2583–2596, 2018.

[22] Liupei Lu, Leili Tavabi, and Mohammad Soleymani. Self-Supervised Learning for Facial Action Unit Recognition through Temporal Consistency. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2020.

[23] Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: a survey. *IEEE Transactions on Affective Computing*, June 2017.

[24] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, April 2011.

[25] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[26] M. Pantic and I. Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, pages 3358–3363 Vol. 4, 2005.

[27] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, Dec 2000.

[28] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3):1449–1461, June 2004.

[29] N. V. Quang, J. Chun, and T. Tokuyama. CapsuleNet for Micro-Expression Recognition. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–7, May 2019.

[30] Maheen Rashid and Yong Jae Lee. Facial action unit detection with capsules. In *PREPRINT*, 2018. Available Online: https://menorashid.github.io/pubs/334.pdf.

[31] S Sabour, N Frosst, and G. E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.

[32] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, June 2015.

[33] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Computer Vision – ECCV 2018*, pages 725–740, Cham, 2018. Springer International Publishing.

[34] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. JÂA-Net: Joint Facial Action Unit Detection and Face Alignment Via Adaptive Attention. *International Journal of Computer Vision*, Sept. 2020.

[35] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing*, pages 1–14, 2019.

[36] Zhiwen Shao, Lixin Zou, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Spatio-temporal relation and attention learning for facial action unit detection. *arXiv preprint arXiv:2001.01168*, 2020.

[37] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[38] Michel F. Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, pages 38–45, New York, NY, USA, 2007. ACM.

[39] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The First Facial Expression Recognition and Analysis Challenge. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 921–926, 2011.

[40] T. Xu, J. White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. *ArXiv*, abs/2007.10075, 2020.

[41] Yan Tong and Qiang Ji. Learning bayesian networks with qualitative constraints. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[42] L. Yang, I. O. Ertugrul, J. F. Cohn, Z. Hammal, D. Jiang, and H. Sahli. FACS3D-Net: 3D Convolution based Spatiotemporal Representation for Action Unit Detection. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 538–544, 2019.

[43] G. N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255, 2017.

[44] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, January 2009.

[45] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. Best of Automatic Face and Gesture Recognition 2013.

[46] K. Zhao, W. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3391–3399, 2016.

[47] K. Zhao, Wen-Sheng Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2207–2216, 2015.