Accepted Manuscript

To appear at the 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2021.

AULA-Caps: Lifecycle-Aware Capsule Networks for Spatio-Temporal Analysis of Facial Actions

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
Email: {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 on evaluating static frames, encoding a snapshot of heightened facial activity. In real-world interactions, however, facial expressions are more subtle and evolve over time requiring AU detection models to learn spatial as well as temporal information. In this work, we focus on both spatial and spatio-temporal features encoding the temporal evolution of facial AU activation. We propose the Action Unit Lifecycle-Aware Capsule Network (AULA-Caps) for 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 on spatial or spatio-temporal information depending upon the AU lifecycle. The proposed model is evaluated on popular benchmarks, namely BP4D and GFT datasets, obtaining state-of-the-art results for both.**

I. Introduction

Analysing facial expressions can be subjective and influenced by contextual and cultural variations [1]. To establish constants across varying cultural contexts and achieve objective evaluations for facial expressions, Ekman *et al.* [2] 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 different AUs reveals local relationships and dependencies 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* [2].

Furthermore, facial muscle activation follows a temporal evolution [3], 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 at the *apex* state to express peak intensity. This is followed by the relaxation of the muscles forming the *offset* state before returning 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 [4]. Understanding this evolution is essential for understanding how humans express affect, particularly for distinguishing *posed* from *spontaneous* expressions [5].

Computational models for AU detection, traditionally, have explored local spatial relationships between different face regions using shape-based representations or using spectral or histogram-based methods [3]. With deep learning gaining popularity, recent approaches [6], [7], [8], [9], [10], [11] have applied convolution or graph-based models to focus on learning such facial features directly from data, outperforming traditional approaches. More recently, capsule-based computations proposed by Sabour *et al.* [12] have further improved the learning of spatial dependencies in the form of facial feature *primitives*. These feature primitives are sensitive to local variations capturing dependencies between different facial regions and have been successfully applied for AU detection and expression recognition tasks [13], [14].

Most approaches, however, focus only on frame-based evaluation of peak-intensity facial frames [3], [15]. As a result, even though these approaches can detect strong AU activations in *posed* settings or highly *accentuated* expressions, they suffer when detecting more *subtle* expressions in *spontaneous* and naturalistic settings [5], [16], challenging their real-world applicability. A prevailing requirement for *automatic* AU detection is to be sensitive to the said *AU Lifecycle* by including temporal information, such as motion features or correlations amongst proximal frames, along with spatial features [9], [16], [17]. While spatial processing is important to determine relationships between different facial regions [17], 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 [6], [9], [16].

Leveraging the ability of capsule networks to learn local spatial and temporal features, we propose the Action Unit Lifecycle-Aware Capsule Network (AULA-Caps) for multi-label AU detection (see Fig. 1). AULA-Caps is a multi-stream capsule network, trained in an end-to-end manner, that not only learns spatial activation patterns within a frame

N. Churamani is funded by the EPSRC grant EP/R513180/1 (ref. 2107412). H. Gunes is supported by the EPSRC project ARoEQ under grant ref. EP/R030782/1, and partially by the European Union's Horizon 2020 research and innovation programme WorkingAge project under grant agreement No. 826232. S. Kalkan is supported by Scientific and Technological Research Council of Turkey (TÜBİTAK) through BIDEB 2219 International Postdoctoral Research Scholarship Program. The authors also thank Prof Lijun Yin from Binghamton University (USA) for providing access to the BP4D Dataset; Prof Jeff Cohn and Dr Jeffrey Girard from the University of Pittsburgh (USA) for providing access to the GFT dataset.

but also their dynamics across contiguous frames. Sensitive to these dynamics, it learns whether to focus more on spatial or spatio-temporal features during the progression of an AU Lifecycle. To the best of our knowledge, this is the first work combining multiple capsule-based processing streams to learn spatial and spatio-temporal features at frame and sequence-level, simultaneously. We perform benchmark evaluations on BP4D [18] and GFT [19] datasets where AULA-Caps achieves the best F1-scores for AUs 1, 6 and 17 and the best overall F1-score on the BP4D dataset and the best F1-scores for AUs $2, 7, 17$ and 23 and second-best overall F1-score on the GFT dataset.

II. Related Work

A. Spatial Analysis for AU Detection

AU detection approaches capture spatial relationships between different face regions [3], [20]. Popular methods include using geometric features that track facial landmarks [21], histogram-based methods to cluster local features into uniform regions [3] or using features that describe local neighbourhoods [22]. With the popularity of deep learning, CNN [7], [23] and graph-based [9], [17] methods have achieved state-of-the-art (SOTA) results for AU detection due to their ability to hierarchically learn spatial features. Capsule-based computations [12] offer an improvement as along with learning different facial features, they also learn how these are arranged with respect to each other. Recent works [13], [24] have explored capsule networks for AU detection by learning facial features that capture variations with respect to pose and orientation. Yet, relying only on spatial features ignores how AU activations evolve over time, impacting performance on automatic AU detection [16].

B. Spatio-Temporal Analysis for AU Detection

Learning spatio-temporal features provides information about the dynamics of AU activation. One way for computing these features is to extract spatial features from each frame separately and use recurrent models such as the LSTM [25] to learn how these evolve with time [6]. Alternatively, models may compute temporal features such as optical flow first and then process them using CNN-based networks [26]. Yet, most of these approaches focus on learning spatial and temporal information sequentially. Yang *et al.* [16] 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 recent methods learn semantic relationships between the face regions and represent these using structured knowledge-graphs to learn coupling patterns between regions using graph-based computations [9], [17].

C. Capsule Networks

Sabour *et al.* [12] 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.* [13] propose 'FACSCaps' that employs capsule networks to learn pose-independent spatial feature representations from multi-view facial images for AU detection. Rashid *et al.* [24] 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 employed by Quang *et al.* [14], 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 [27] 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 may contribute towards sequence-based predictions. Jayasekara *et al.* [28], on the other hand, apply capsule-based learning for time-series predictions learning to classify 1D ECG signals focusing on temporal dependencies.

In this work, we propose a multi-stream approach that applies capsule-based computations at frame and sequence-level concurrently, learning spatial and spatio-temporal dependencies from sequences of contiguous facial frames.

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

We propose AULA-Caps (see Fig. 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), here the middle frame from each input sequence, focuses on local spatial dependencies, spatio-temporal processing investigates contiguous frames to capture the 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 passed to a decoder that learns to reconstruct the FoI, further regularising learning.

A. Windowed Video Sequences as Input

AULA-Caps takes as input a video sequence of contiguous (96×96) grayscale frames of normalised face-centred images (each pixel $p \in [-1, 1]$) generated by taking each frame of the video, along with N frames immediately preceding and succeeding it. The middle FoI is passed to the spatial processing stream while the entire window of $2N+1$ frames is processed using the spatio-temporal stream. The overall task for the model is to predict the activated AUs in the FoI. We optimise AULA-Caps for the overall F1-Score comparing N=$\{1, 2, 3\ ,\ 4\}$. Setting N=2 performs the best, resulting in an input window of 5 frames (see Table III for a comparison).

B. Motivation for Lifecycle-Awareness

Following the AU lifecycle, different segments; *onset*, *apex* and *offset*, form the evolution of an AU. In *onset* and

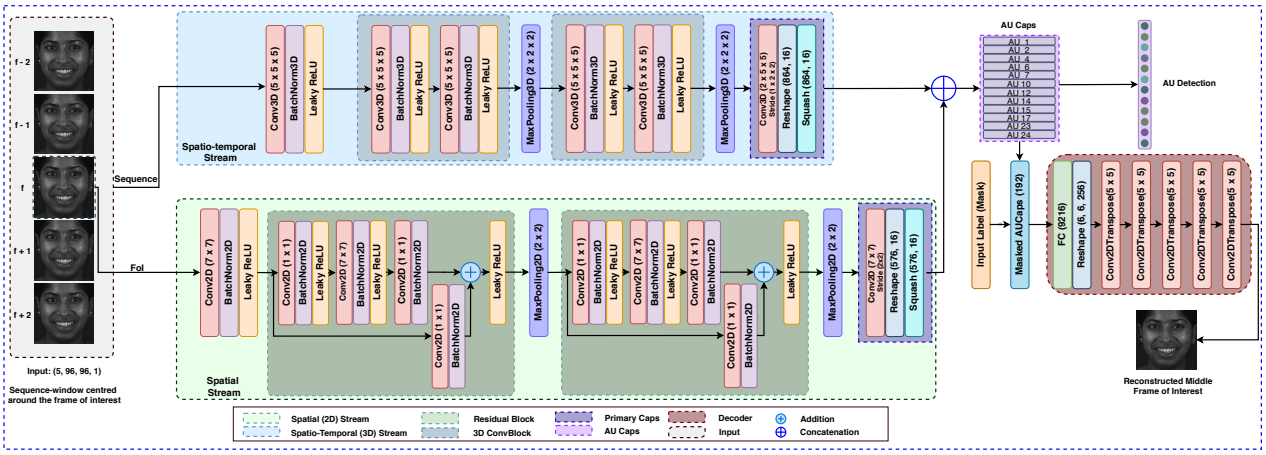
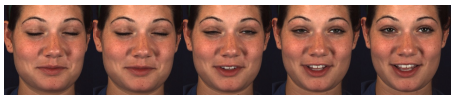
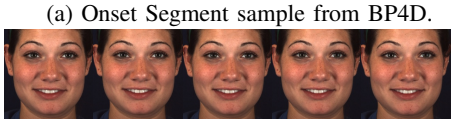


Fig. 1: Action Unit Lifecycle-Aware Capsule Network (AULA-Caps) for Multi-label Facial Action Unit Detection.



(a) Onset Segment sample from BP4D.



(b) Apex Segment sample from BP4D.

Fig. 2: Onset and Apex segment contiguous frames.

offset phases, the input images have high *variation*, in that, the contiguous frames are sufficiently different, as illustrated in Fig. 2a. Thus, focusing on this difference provides important temporal information about AU activations. In *apex* segment frames, however, the contiguous frames have low *variation* and are not sufficiently different, as illustrated in Fig. 2b. Here, spatial features extracted from a single FoI alone may provide sufficient information for AU detection.

The two streams in the AULA-Caps model are designed to exploit this difference by extracting relevant spatial and spatio-temporal features and combining them by weighting their individual contribution based on their relevance for AU prediction. Selectively tuning into these features based on where in the AU lifecycle the input sequence originates from, motivates the *lifecycle-awareness* of the model.

C. Computing Spatial Features

The spatial processing stream (see Fig. 1 bottom) processes the FoI ($\boldsymbol{x_f}$) from an input sequence and passes it through a convolutional (conv) layer with 128 filters of size (7×7) followed by *BatchNorm* and *LeakyReLU* ($\alpha = 0.2$) activation. The output is passed through 2 Residual blocks consisting of multi-resolution conv layers with shortcut connections [29], with 128 and 64 filters for each conv layer in the respective blocks using *LeakyReLU* ($\alpha = 0.2$) activation. Each block is followed by a (2×2) maxpooling layer and the final output is passed to the Primary Capsule layer consisting of a conv layer with reshaping and squashing of extracted spatial features into 576 capsules of 16 dimensions each.

D. Computing Spatio-Temporal Features

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

E. Combining Extracted Features

The extracted *primary* capsules representing spatial and spatio-temporal primitives from the two streams are concatenated together resulting in 1440 capsules of 16 dimensions each. The iterative *routing-by-agreement* algorithm [12] then couples these capsules with the AU-Caps layer, computing 12 capsules corresponding to the AU labels. Since the primary capsules are concatenated before routing, these are *competitively* weighted together based on whether spatial or spatio-temporal features contribute more towards detecting each of the activated AUs. The output of the AU-Caps layer is used to predict the AUs activated in the FoI, replacing the capsule with its length *squashed* between $[0, 1]$ depicting the activation probability for the AU label. The AU-Caps layer output is also used by the Decoder to reconstruct the FoI.

F. Decoder for Image Reconstruction

The Decoder regularises learning in the model 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. In AULA-Caps, we use transposed conv layers for the decoder, instead of dense layers proposed by Sabour *et al.* [12]. This significantly reduces the number of parameters in the decoder ($\approx$ 2.8M *vs.* 10M in [12]) while improving the photo-realistic quality of reconstructed

images. The decoder, adapted from the generator of [30], implements 4 stacked transposed conv layers, using *ReLU* activation, with $128, 64, 32, 16$ filters, respectively, of size (5×5) each with a stride of (2×2). Another transposed conv layer with *tanh* activation generates the resultant image ($\boldsymbol{x_{gen}}$) with the same dimensions as the FoI ($\boldsymbol{x_f}$).

G. 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 generates 2 outputs in each run: the activation probabilities for the 12 AUs and the reconstructed FoI. The learning objectives for the model are as follows:

1) 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 ($\boldsymbol{\mathcal{L}_{au}}$) is defined as:

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

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 [12]. $\boldsymbol{w_{au}}$ 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 [31], $\boldsymbol{w_{au}}$ is computed as follows:

$$w_{au} = \frac{(1/r_i)N}{\sum_{i}^{N}(1/r_i)}, \tag{2}$$

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 ($\boldsymbol{\mathcal{L}_{margin}}$) is computed as the sum of the losses for each AU ($\boldsymbol{\mathcal{L}_{au}}$).

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

$$\min_{X_f, X_{gen}} \mathcal{L}_{rec} = L_2(x_f, x_{gen}), \tag{3}$$

where $\boldsymbol{x_f}$ is the FoI and $\boldsymbol{x_{gen}}$ is the reconstructed image.

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

$$\mathcal{L}_{AULA} = \mathcal{L}_{margin} + \lambda_d \mathcal{L}_{rec}, \tag{4}$$

where λ_d is set to 0.05 to balance the loss terms.

IV. Experiments

A. Datasets

We evaluate AULA-Caps on two popular AU benchmarks; BP4D and GFT. For both datasets, samples representing the 12 most frequently occurring AUs; namely AUs $1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 24$, are used.

1) BP4D: The BP4D dataset [18] consists of videos from 41 subjects performing 8 different affective tasks to elicit emotional reactions. Approximately 500 frames for each video are annotated for AUs occurrence and intensity. In our experiments, we only use occurrence labels for AU detection.

2) GFT: The Sayett-GFT Dataset [19] 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.

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

B. Experiment Settings

1) Evaluation Metric: Similar to other approaches [9], [23], we follow $3-$fold cross-validation for our evaluations, splitting the data into 3 folds where each subject occurs in the 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 [32].

2) Implementation Details: The AULA-Caps is implemented using Keras-Tensorflow. The model is trained individually on each dataset in an end-to-end manner using the *Adam* optimiser with an initial learning-rate of $2.0e^{-4}$, decayed each epoch by a factor of 0.9. The model is trained for 12 epochs with early stopping with a batch-size of 24. No data-augmentation is performed during training on either of the datasets. Model hyper-parameters: filter number and size for each layer, capsule dimensions, batch-size and learning-rate are optimised using the Hyperopt Python Library.

C. Results

1) BP4D: Table I presents AULA-Caps results for BP4D and compares them to the SOTA approaches (scores reported from respective papers) such as the [CNN-LSTM] [6] learning temporal variation in facial features, the [EAC] method [7] that employs enhancing and cropping mechanism to focus on selective regions in an image, the [ROI] network [33] that focuses on learning regional features using separate local CNN, a 2D Capule-Net based model [CapsNet] [24], the [JÂA] [34] approach that uses multi-scale high-level facial features, the semantic learning-based [SRERL] [17] model and the [STRAL] [9] approach that employs a spatio-temporal graph CNN to capture both spatial and temporal relations for AU prediction. AULA-Caps uses a multi-stream approach that simultaneously learns and combines spatial and spatio-temporal features making it sensitive to the temporal evolution of AU activations. AULA-Caps achieves the best results for 3 AUs and second-best results for

TABLE I: Performance Evaluation (F1-Scores) on BP4D. **Bold** values denote best while [*bracketed*] denote second-best values for each row.

AU	CNN-LSTM [6]	EAC [7]	ROI [33]	CapsNet [24]	JÂA [34]	SRERL [17]	STRAL [9]	AULA-Caps [*Ours*]
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**

another 3. Overall, the model outperforms other models, with closest *Avg.* F1-score difference to the [STRAL] approach [9] being 0.013 with both [STRAL] and AULA-Caps combining spatial and temporal analysis of facial features. Yet, while [STRAL] employs a multi-stage training strategy where different components of the model are trained sequentially one after the other, all of the components of the AULA-Caps are trained together in an end-to-end manner.

2) GFT: Table II presents AULA-Caps performance on the GFT dataset in comparison to the SOTA (scores reported from respective papers) consisting of different spatial and spatio-temporal approaches such as the CNN-based cross-domain learning [CRD] [23], an Alex-Net-based model [ANet] for frame-based AU detection [6], 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 [6]. 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, AULA-Caps 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 AUs and second-best results for another 3. [CNN-LSTM] [6] reports the best F1-scores, however the model is evaluated with data from only 50 out of the 96 participants. Despite achieving the second-best overall results, AULA-Caps performs rather poorly for under-represented AU $1, 4$ and 14 impacting the overall F1-score.

D. 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 contributes to the overall performance of the model. To evaluate the contribution of the learnt spatial features, we use the trained 2D stream to predict

TABLE II: Performance Evaluation (F1-Scores) on GFT. **Bold** values denote best while [*bracketed*] denote second-best values for each row. *`Averaged for 10 AUs.`

AU	CRD [23]	ANet [6]	JÂA [34]	CNN-LSTM [6]	AULA-Caps [*Ours*]
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 III: Ablations using BP4D dataset. Decoder parameters (≈ 2.8M) excluded for comparison with CNN baselines.

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
Dual-Stream CNN Baseline	0.596	25.6M	0.64s
2D Stream AULA-Caps	0.580	3.06M	0.35s
3D Stream AULA-Caps	0.550	8.46M	0.66s
AULA-Caps (N=1)	0.599	11.67M	0.71s
AULA-Caps (N=2)	**0.645**	11.51M	1.22s
AULA-Caps (N=3)	0.603	14.24M	1.66s
AULA-Caps (N=4)	0.619	14.32M	1.78s

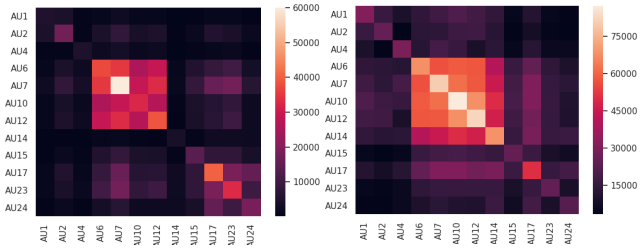
AUs by appending a separate AU-Caps layer to the primary capsule layer. The weights of the 2D stream are frozen and only the routing algorithm is run for the added AU capsule layer. Similarly, for assessing the effect of learning spatio-temporal features, we use the trained spatio-temporal (3D) stream to predict the AU labels by appending a separate AU-Caps layer to the primary capsules. Additionally, we also evaluate different windows sizes of $2N + 1$ frames with $N \in \{1, 2, 3, 4\}$.

Furthermore, for highlighting the contribution of capsule-based computation, we compare the results with 2D, 3D and Dual-Stream CNN-based models. The CNN streams are unchanged with the capsule-block replaced by fully-connected layers. The results for the different ablations conducted are presented in Table III. Analysing ablations with BP4D provides a fair comparison as it consists of more samples per subject with cleaner, face-centred images.

V. Analysis and Discussion

A. Lifecycle-Awareness

The capsule-based computations of the multi-stream AULA-Caps allow it to weigh the contribution of spatial



(a) GFT (b) BP4D

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

and spatio-temporal feature capsules towards predicting AU activations. If spatial features are more relevant, for example, for *apex* frames where an AU is activated with highest intensity, the model may choose to give precedence to spatial capsules. For off-peak intensity frames, for example, the *onset* or *offset* 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 III) 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 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 [9], [16]. Interestingly, the windowed computation of spatio-temporal features (3D Stream) performs worse than spatial features (2D Stream), unlike other approaches [16] where 3D features perform better. This may be due to the choice of a smaller input window (5 frames in AULA-Caps) unlike [16] where an entire video is considered for computing spatio-temporal features (see Section VI-A.1 for a discussion).

B. AU Prediction

The AULA-Caps model achieves SOTA results for both BP4D (see Table I) and GFT (see Table II) datasets. 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 Fig. 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 samples while AUs $1, 4$ and 14, the lowest. In such an imbalanced data distribution, where AUs $1, 4$ and 14 correspond to less than 2% of the total samples, the model is unable to learn relevant features to detect these AUs. The imbalance in data correlates with the model performance on individual AUs.

A similar imbalance is also witnessed for the BP4D dataset (see Fig. 3b) yet, an overall larger number of samples per AU helps mitigate some of these effects for BP4D. Furthermore, for the GFT dataset, subjects are recorded interacting 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 SOTA compared in Table II. 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 I. AULA-Caps is able to achieve competitive scores on the GFT dataset despite its more complex and challenging settings while outperforming SOTA evaluations on the BP4D dataset.

C. Temporal Evaluation

AU detection evaluations commonly use only frame-wise performance metrics. 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 Fig. 4, we plot, across time, the true labels as well as model predictions 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 ground-truth 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 within the video. AULA-Caps is able to model this switch effectively, predicting AU activations efficiently.

Furthermore, we see that the 3D stream, on average, models the changing dynamics of AU activations better than the 2D stream, especially in regions where ground truth switches from *absent* to *activated* or vice-versa. Yet, the 2D stream has a better average performance across all videos. As frame-based evaluation only reports *average* F1-scores, they ignore temporal correspondences commonly examined for continuous affect prediction [35]. Yet, these can be beneficial for understanding real-time model performance, underlining its applicability for real-world *automatic* AU prediction.

D. Visualisations

1) Image Reconstruction: The decoder regularises learning by ensuring the model learns task-relevant features. Additionally, the reconstructed images enable a visual interpretation of the learnt features. The convolution-based AULA-Caps decoder is able to reconstruct images using a much 'lighter' network ($\approx$ 2.8M parameters *vs.* $\approx$ 10M [12]) without compromising on quality, as can be seen in Fig. 5. The data imbalance problem is witnessed in the reconstructed images as well where FoIs for certain under-represented subjects and AUs are reconstructed incorrectly. For example, faces at (row 1, col 2) and (row 2, col 1) are reconstructed as *generic mean* faces representing the corresponding AUs, with a visible bias for ethnicity and gender.

2) Visualising Saliency Maps: Visualising learnt features helps understand what the model pays attention to while making its predictions. In Fig. 6, we see Saliency Maps [36] generated by visualising the pixels in the FoIs that contribute

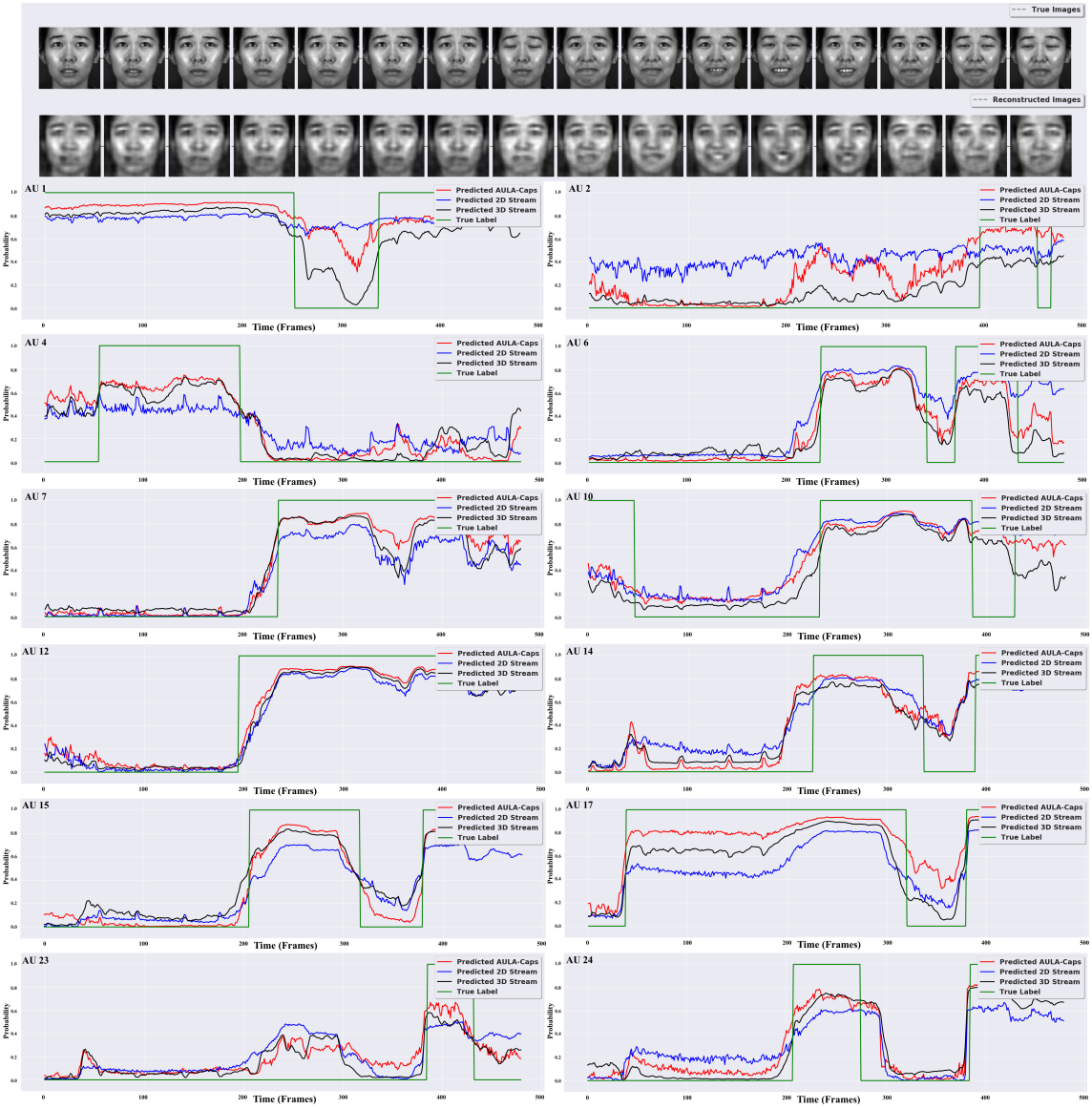


Fig. 4: Comparing predictions for the 2D stream, 3D stream and AULA-Caps for the 12 AUs for a sample BP4D video.

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

Fig. 5: FoI Image reconstruction by the Decoder.

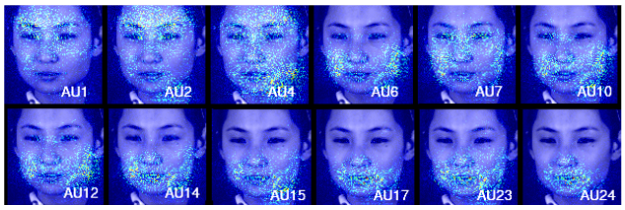


Fig. 6: Saliency Maps generated using *guided backpropagation* of gradients corresponding to each AU label.

most to model predictions. As desired, for different AUs the model learns to focus on different regions of the face. For example, for AUs 1 and 2 it focuses more on the *forehead* and *eyebrows* while for AUs 23 and 24, it focuses on the *nose* and *mouth*. For certain AUs however, we see additional activity in other 'irrelevant' face regions. For example, for AU 4, we see activity in the lower face region near the mouth and cheeks. This is due to the co-occurrence pattern (see Fig. 3) observed in the data distribution where samples containing AU 4 also encode activity for AU 7 and AU 17. Understanding such co-occurrence patterns can be important to improve model predictions for AU activations [17].

VI. Conclusion

Our experiments with the AULA-Caps demonstrate that 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 in line with other findings [9], [37]. Furthermore, capsule-based

computations in the spatial stream enable learning local spatial relationships corresponding to the different face regions while the spatio-temporal stream is 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.

A. Limitations and Future Work

1) 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 AU activity is annotated for those frames. This impacts model performance resulting in poor performance for AUs $1, 4$ 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 a wider window is expected to improve performance. In our experiments, however, we optimised the window-size for the highest overall F1-score, only comparing sizes $3, 5, 7$ and 9. Further experimentation is needed to investigate which window-sizes work best for different AUs. Also, learning to dynamically adapt the windows based on AU activity may offer improvements. Lu *et al.* [38] provide an insightful approach to address this 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 sequence compared to this anchor frame. This robustly captures temporal dependencies in facial activites, improving AU detection performance.

2) Imbalanced Data Distributions: Another problem faced by most approaches is the imbalanced label distribution of the datasets. In Fig. 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. Understanding AU co-activations can provide additional contextual information to improve performance on scarce AU samples [17]. Furthermore, 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 [39], [40], [41].

REFERENCES

[1] R. E. Jack *et al.*, "Facial expressions of emotion are not culturally universal," *PNAS*, vol. 109, no. 19, pp. 7241–7244, 2012.

[2] P. Ekman *et al.*, *Facial action coding systems*. Consulting Psychologists Press, 1978.

[3] E. Sariyanidi *et al.*, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE PAMI*, vol. 37, no. 6, pp. 1113–1133, Jun 2015.

[4] J. F. Cohn *et al.*, "The Timing of Facial Motion in Posed and Spontaneous Smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 02, no. 02, pp. 121–132, Jun. 2004.

[5] M. F. Valstar *et al.*, "How to distinguish posed from spontaneous smiles using geometric features," in *ICMI*, 2007.

[6] W.-S. Chu *et al.*, "Learning spatial and temporal cues for multi-label facial action unit detection," in *FG*, 2017.

[7] W. Li *et al.*, "EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection," *IEEE PAMI*, vol. 40, no. 11, pp. 2583–2596, 2018.

[8] Z. Shao *et al.*, "Facial action unit detection using attention and relation learning," *IEEE Transactions on Affective Computing*, pp. 1–14, 2019.

[9] Z. Shao *et al.*, "Spatio-temporal relation and attention learning for facial action unit detection," *arXiv preprint arXiv:2001.01168*, 2020.

[10] K. Zhao *et al.*, "Deep region and multi-label learning for facial action unit detection," in *CVPR*, 2016, pp. 3391–3399.

[11] K. Zhao *et al.*, "Joint patch and multi-label learning for facial action unit detection," in *CVPR*, 2015, pp. 2207–2216.

[12] S. Sabour *et al.*, "Dynamic routing between capsules," in *NIPS*, 2017.

[13] I. O. Ertugrul *et al.*, "Facscaps: Pose-independent facial action coding with capsules," in *IEEE CVPR Workshops*, 2018, pp. 2211–2220.

[14] N. V. Quang *et al.*, "CapsuleNet for Micro-Expression Recognition," in *FG*, 2019.

[15] B. Martinez *et al.*, "Automatic analysis of facial actions: a survey," *IEEE Transactions on Affective Computing*, Jun 2017.

[16] L. Yang *et al.*, "FACS3D-Net: 3D Convolution based Spatiotemporal Representation for Action Unit Detection," in *ACII*, 2019.

[17] G. Li *et al.*, "Semantic relationships guided representation learning for facial action unit recognition," in *AAAI*, 2019.

[18] X. Zhang *et al.*, "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[19] J. M. Girard *et al.*, "Sayette group formation task (GFT) spontaneous facial expression database," in *FG*, 2017.

[20] M. Pantic *et al.*, "Facial action recognition for facial expression analysis from static face images," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 34, no. 3, pp. 1449–1461, 2004.

[21] F. De la Torre *et al.*, "Intraface," in *FG*, 2015, pp. 1–8.

[22] S. A. Bargal *et al.*, "Classification of mouth action units using local binary patterns," in *IPCV*, 2012.

[23] I. O. Ertugrul *et al.*, "Cross-domain au detection: Domains, learning approaches, and measures," in *FG*, 2019, pp. 1–8.

[24] M. Rashid *et al.*, "Facial action unit detection with capsules," in *PREPRINT*, 2018.

[25] S. Hochreiter *et al.*, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[26] B. Allaert *et al.*, "Optical flow techniques for facial expression analysis: Performance evaluation and improvements," *CoRR*, 2019.

[27] K. Duarte *et al.*, "VideoCapsuleNet: A Simplified Network for Action Detection," in *NIPS*, 2018.

[28] H. Jayasekara *et al.*, "TimeCaps: Capturing Time Series Data with Capsule Networks," *ArXiv*, vol. abs/1911.11800, 2019.

[29] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[30] N. Churamani *et al.*, "CLIFER: Continual Learning with Imagination for Facial Expression Recognition," in *FG*, 2020, pp. 322–328.

[31] Z. Shao *et al.*, "Deep adaptive attention for joint facial action unit detection and face alignment," in *ECCV*, 2018.

[32] M. F. Valstar *et al.*, "The First Facial Expression Recognition and Analysis Challenge," in *FG*, 2011, pp. 921–926.

[33] W. Li *et al.*, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *CVPR*, 2017.

[34] Z. Shao *et al.*, "JÂA-Net: Joint Facial Action Unit Detection and Face Alignment Via Adaptive Attention," *IJCV*, 2020.

[35] G. N. Yannakakis *et al.*, "The ordinal nature of emotions," in *ACII*, 2017.

[36] K. Simonyan *et al.*, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *ICLR Workshop*, 2014.

[37] M. Pantic *et al.*, "Detecting facial actions and their temporal segments in nearly frontal-view face image sequences," in *Int. Conf. Systems, Man and Cybernetics*, 2005.

[38] L. Lu *et al.*, "Self-Supervised Learning for Facial Action Unit Recognition through Temporal Consistency," in *BMVC*. BMVA Press, 2020.

[39] F. Charte *et al.*, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, vol. 89, pp. 385 – 397, 2015.

[40] K. Oksuz *et al.*, "Imbalance problems in object detection: A review," *IEEE PAMI*, 2020.

[41] N. Churamani *et al.*, "Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition," 2021, arXiv:2103.08637.