UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

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

***Nikhil Churamani*,** Sinan Kalkan and Hatice Gunes

Engineering and Physical Sciences Research Council

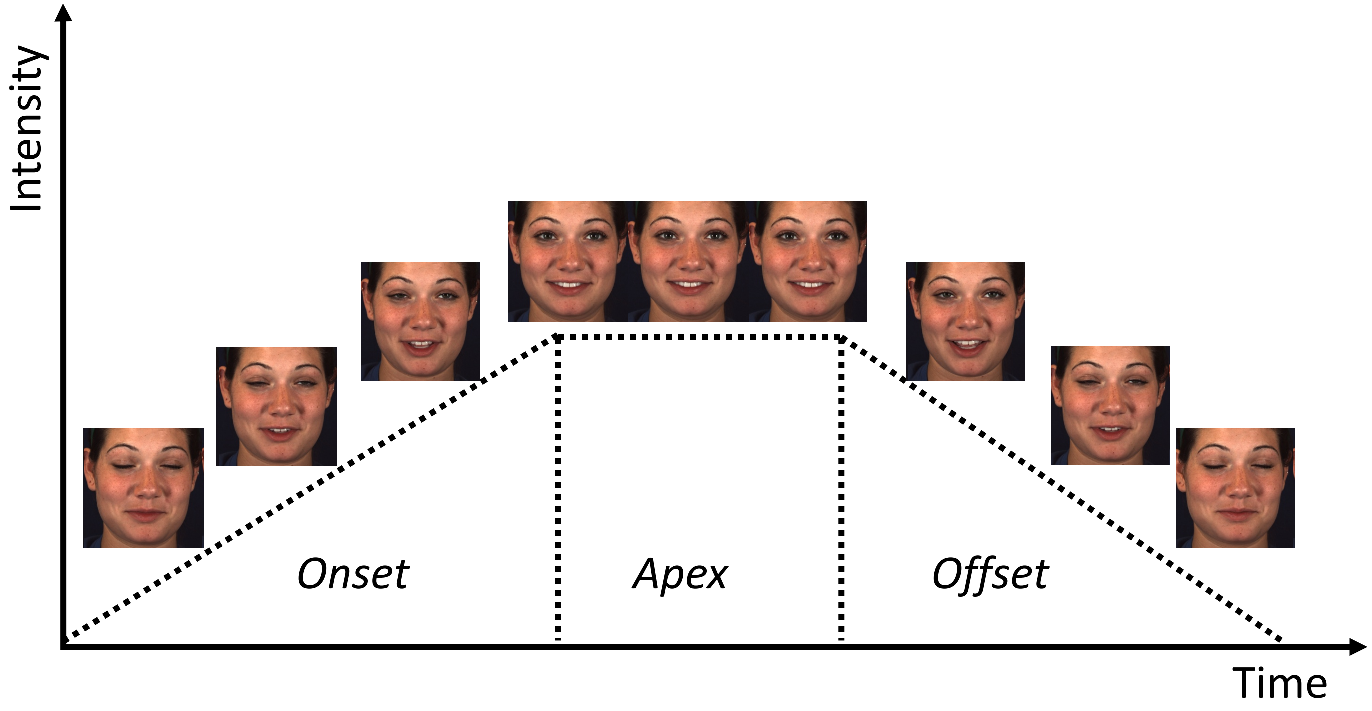
Motivation: Facial Action Coding System

- Facial Action Coding System (FACS) (Ekman *et al.* 1978) provides **objective** evaluations of Human Facial Expressions.
- Facial **AUs** encode muscle activity.
- **Precise representation** of facial activity.
- **No** subjective **interpretation** needed.

Valente, D, et al. “The role of visual experience in the production of emotional facial expressions by blind people: A review”. *Psychon Bull Rev* **25,** 483–497 (2018).

Motivation: The AU Lifecycle

- Facial Action Unit (AU) Activation follows a temporal evolution: the ***<u>AU Lifecycle</u>***.
- Facial muscles contract to form the ***onset*** phase.
- Complete contraction at the ***apex*** state.
- Muscles start to relax in the ***offset*** phase.



Motivation: Spatial *vs.* Spatio-temporal Features

Spatial Features

- Capture **local relationships** between facial regions.
- **Hierarchical features** sensitive to **local variations**.
- Contiguous frames in the ***apex*** phase experience **low** variations.
- Spatial features provide more **descriptive** information during the ***apex*** phase.

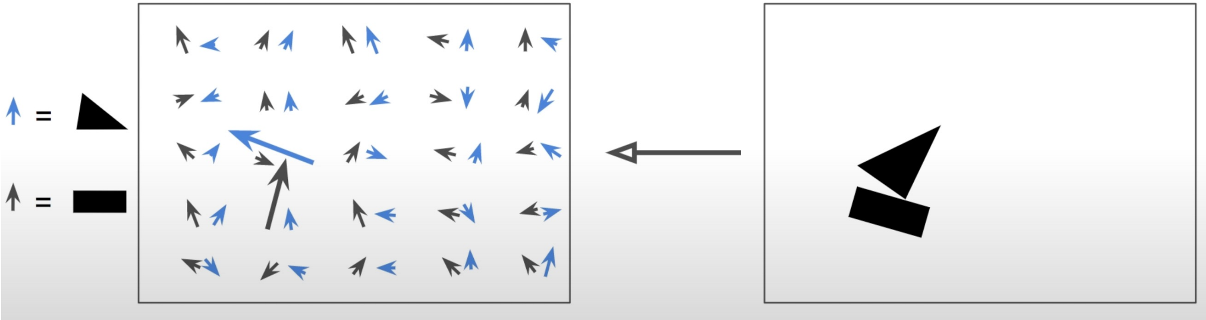
Spatio-temporal Features

- Capture how facial features **vary across frames**.
- **Temporal features** sensitive to variations over time.
- Contiguous frames in the ***onset*** and ***offset*** phases experience **high** variations.
- Spatio-temporal fetures provide more **descriptive** information during ***onset*** and ***offset*** phases.

Can we dynamically learn to selectively focus on spatial or spatio-temporal features?

Motivation: Capsule Networks

- Capsules help encode **spatial primitives** or features constituting the object of interest.
- **Length** encodes **probability** of presence.
- **Orientation** encodes parameters such as **pose** variations.
- Local **spatial relationships learnt** between the object of interest and its surroundings.

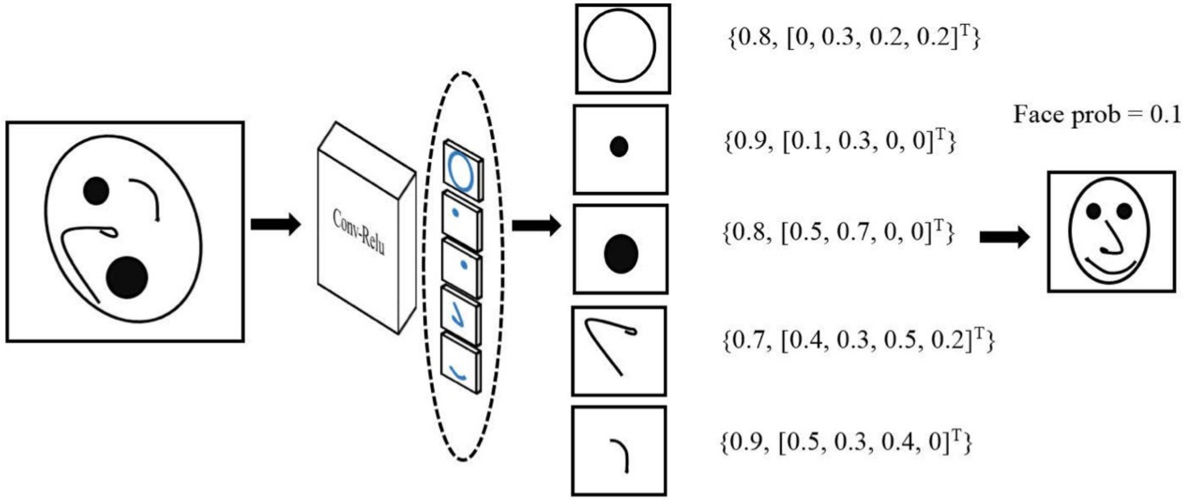


(a)

a) https://www.slideshare.net/aureliengeron/introduction-to-capsule-networks-capsnets

Motivation: Capsule Networks

- Each capsule may learn **features** relevant for **different parts** of the face.
- Capsules may **encode position, rotation, pose features** for each individual part.
- **Local relationships** between these features **guide** model **predictions**.
- Observing **contiguous frames** may help provide insights into how these relationships **vary with time**.

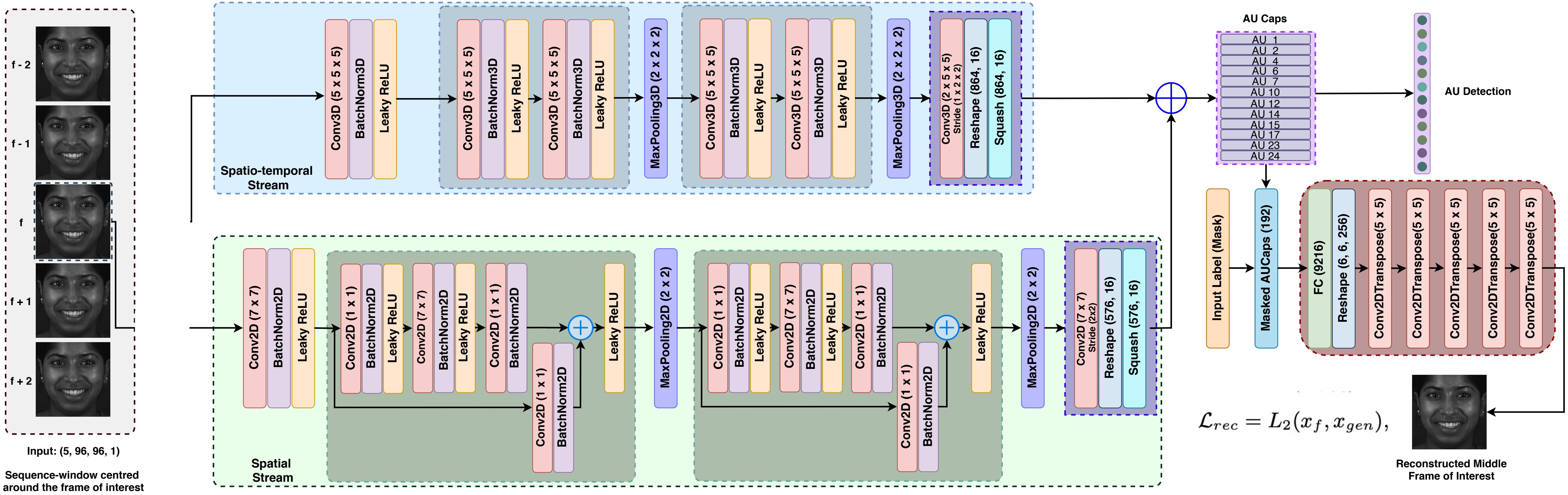


(a)

a) A. Shahroudnejad, et al., "Improved Explainability Of Capsule Networks: Relevance Path By Agreement," IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2018, pp. 549-553.

Action Unit Lifecycle-Aware Capsule Networks

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



$$\mathcal{L}_{rec} = L_2(x_f, x_{gen}),$$

Evaluations

- Multi-label AU Prediction:
  - Evaluate model performance on **two datasets** for *12 Action Units*:

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

- Model Ablations:
  - Spatial vs. Spatio-temporal Features.
  - Convolutional vs. Capsule-based computations.
  - Window sizes.
- Model Visualisations:
  - Image Reconstructions.
  - Visualising Saliency Maps.

BP4D

(a)

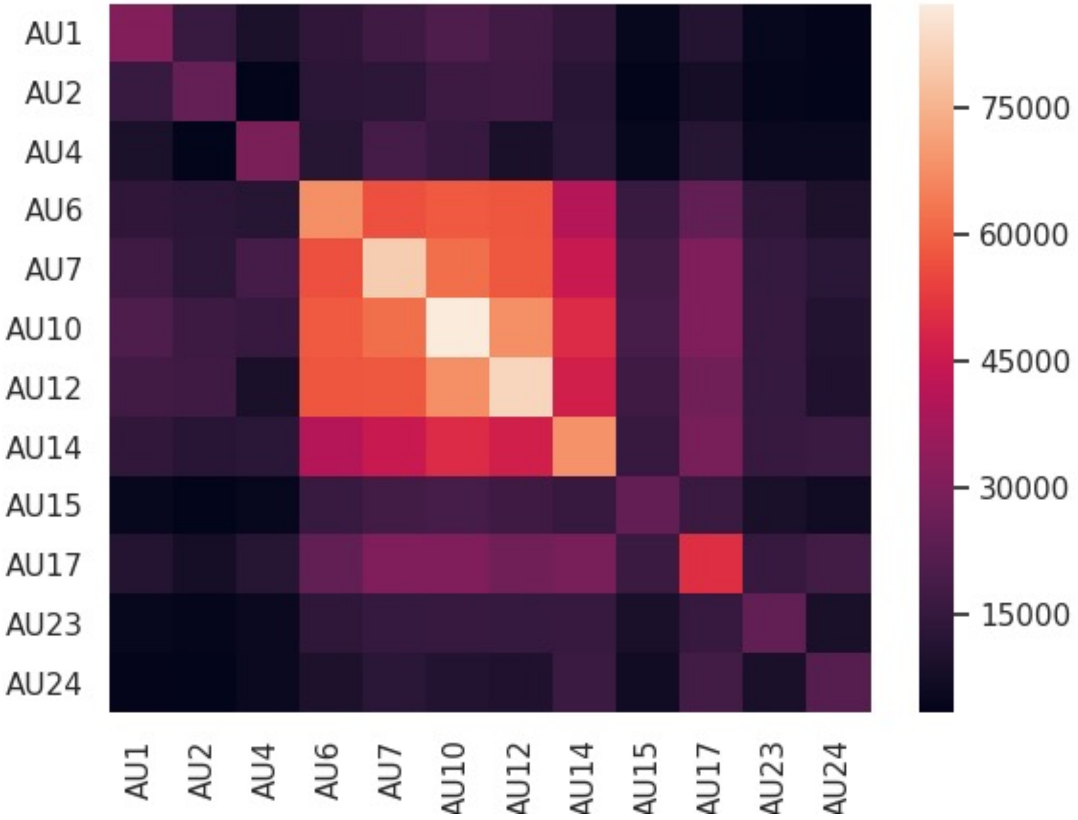
GFT

(b)

a) Xing Zhang, et al. "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database", Image and Vision Computing, Volume 32, Issue 10, 2014,Pages 692-706.

a) J. M. Girard, et al. "Sayette Group Formation Task (GFT) Spontaneous Facial Expression Database," IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 581-588.

AU Prediction: BP4D Dataset



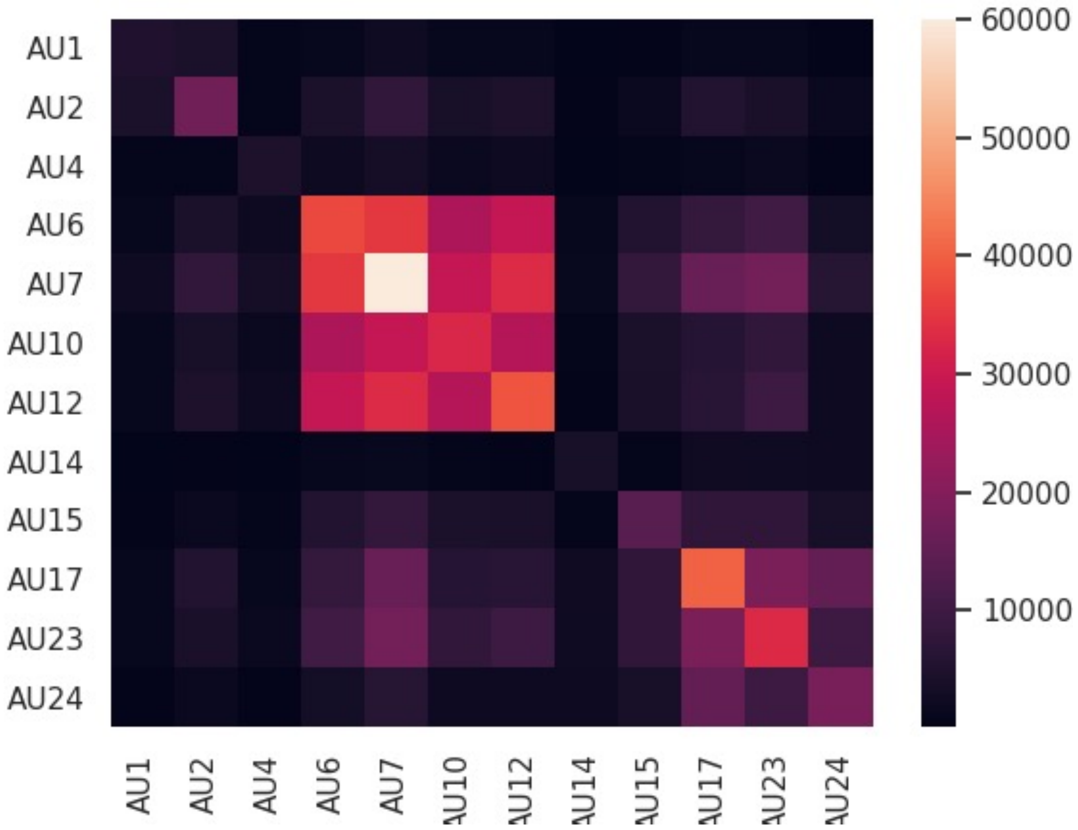
Co-activation Heatmap

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**

UNIVERSITY OF CAMBRIDGE

AU Prediction: GFT Dataset



Co-activation Heatmap

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*]

**** Results on 50 out of 96 subjects.**

Model Ablations

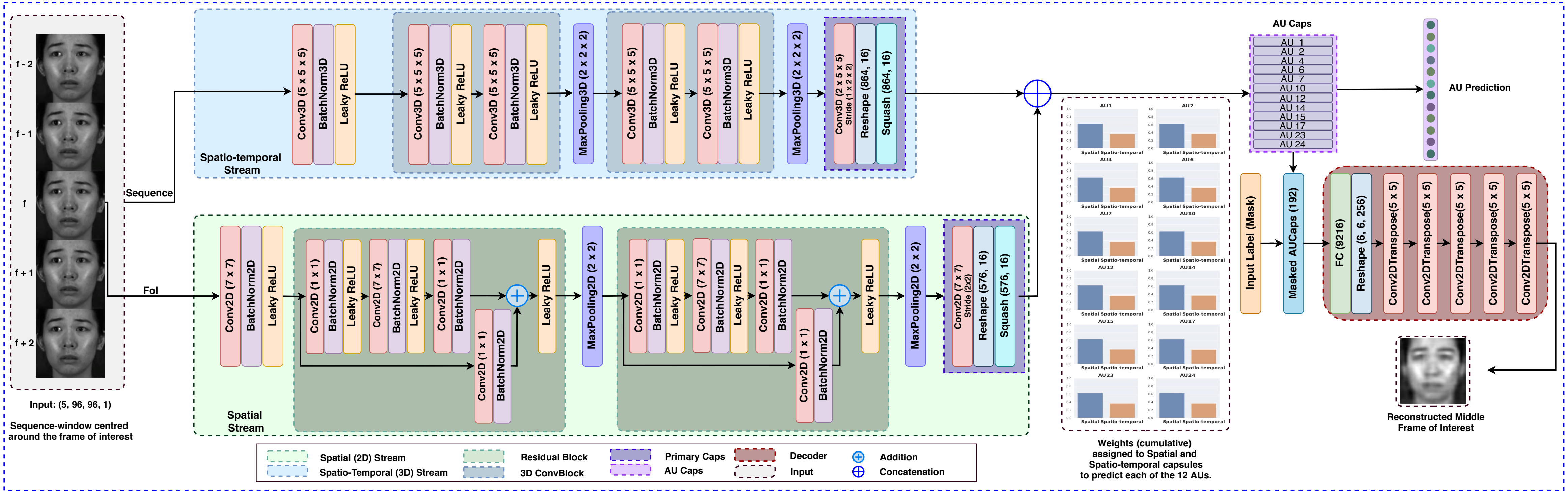
- Spatial vs. Spatio-Temporal Features:
  - 2D performs better than 3D on frame—based analyses.
  - Combining 2D and 3D features results in improved performance overall.
- Convolution vs. Capsule-based Computation:
  - Capsule-based computations provide improvements across evaluations.
  - # Parameters to be trained are decreased.
- Ablating Window Sizes:
  - Increasing Window size, on average improves performance.
  - Window size 5 (N=2) performs the best.

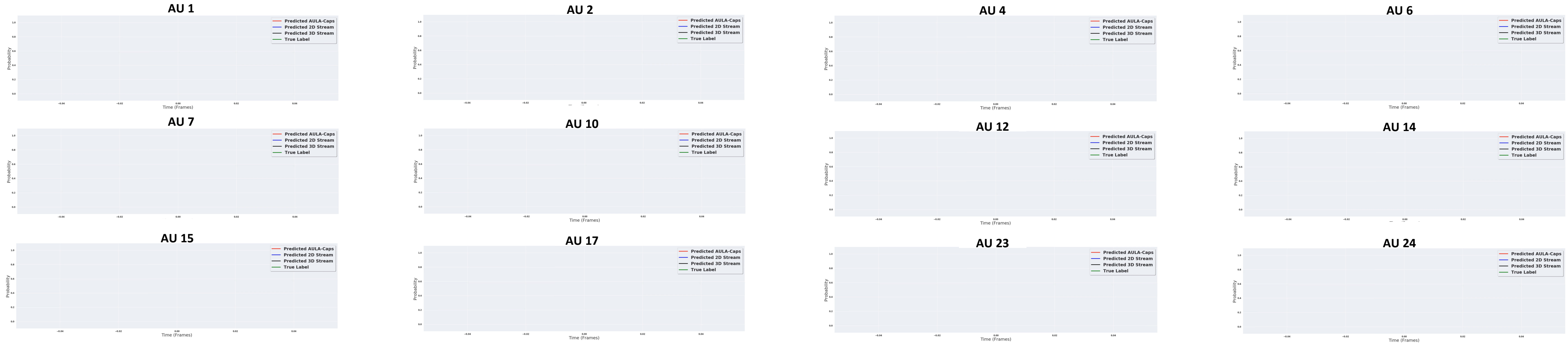
TABLE III: Ablations using BP4D dataset. Decoder parameters ($\approx$ 2.8M) excluded for comparison with CNN baselines.

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

UNIVERSITY OF CAMBRIDGE

Dynamically Weighting Features

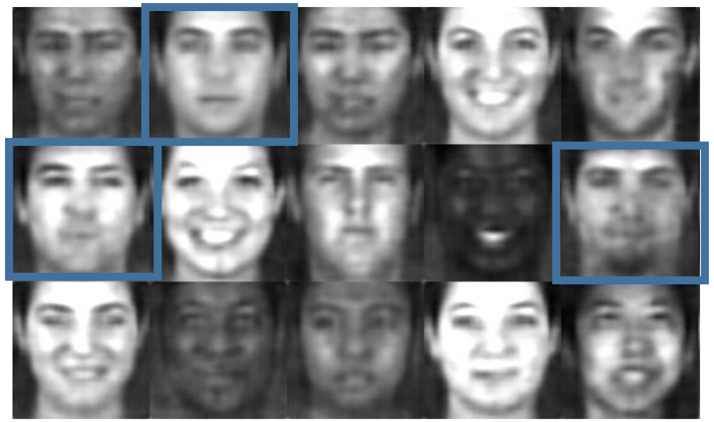




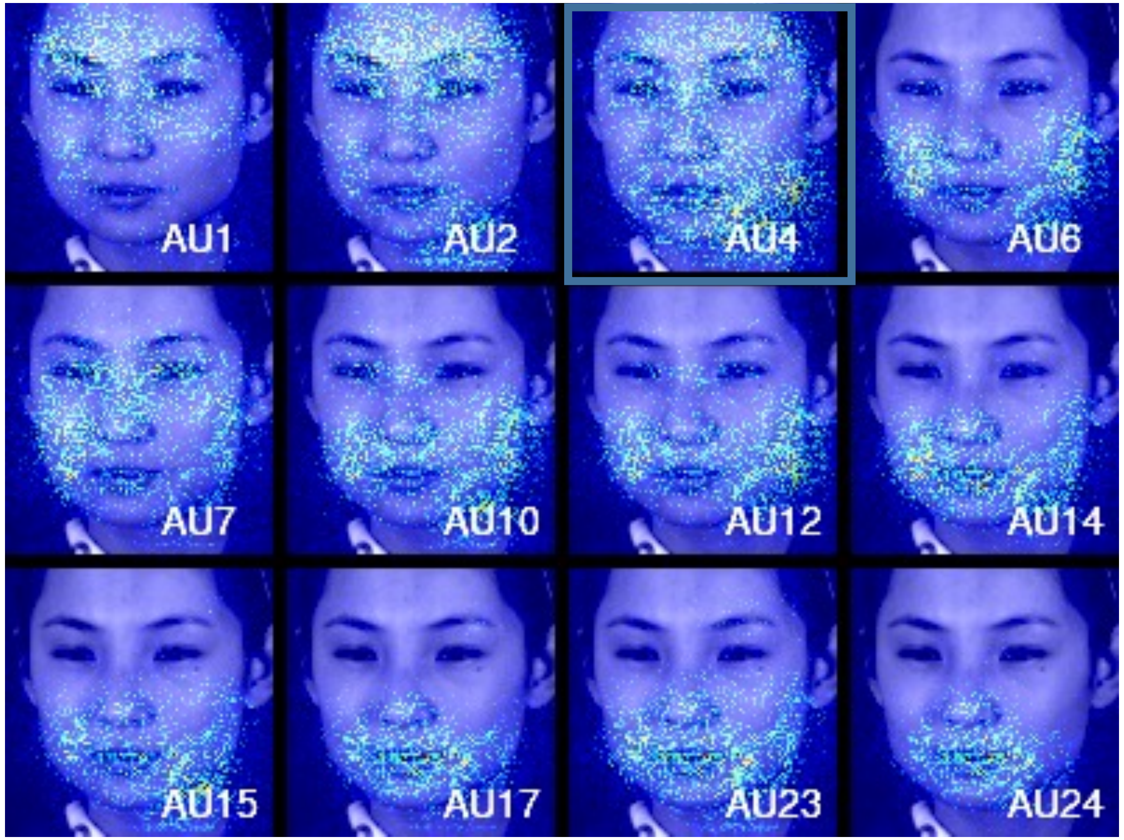
Visualisations

Input FoI Images

Reconstructed FoI Images



Saliency Maps



Take Away Message

Conclusions

- First implementation combining **spatial** and **spatio-temporal** capsule-based computations.
- Spatio-temporal information provides **context** for continuous AU prediction.
- Combining **spatial** and **spatio-temporal** feature primitives improves model performance.
- **Selectively focusing** on spatial and spatio-temporal features through capsule routing enables robustness.

Next Steps

- Model performance **sensitive** to sequence window length.
- Dynamically **adapting** window-size based on specific AU lifecycles using **anchor frames** (Lu *et al.* 2020).
- Data Imbalance major hurdle for multi-label classification problems.
    - Using co-activation patterns as context to improve model performance (Li *et al.* 2019).
    - Advanced methodologies such as Synthetic Instance Generation (Charte *et al. 2015)* or Continual Learning (Churamani *et al. 2021).*

Acknowledgement

Nikhil Churamani

Sinan Kalkan

Hatice Gunes

Nikhil Churamani is supported by the EPSRC (grant EP/R513180/1 ref. 2107412).

Hatice Gunes is supported by European Union's Horizon 2020 research and innovation programme, under grant agreement No. 826232.

S. Kalkan is supported by Scientific and Technological Research Council of Turkey (TUBITAK) through BIDEB 2219 International Postdoctoral Research Scholarship Program.

UKRI Engineering and Physical Sciences Research Council

UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

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

Nikhil Churamani, Sinan Kalkan and Hatice Gunes

UKRI Engineering and Physical Sciences Research Council