Self-Supervised Apparent Emotional Reaction Recognition from Video

Meta AI | Imperial College London

Marija Jegorova, Stavros Petridis, Maja Pantic

07/01/2023

Why apparent emotional reaction recognition?

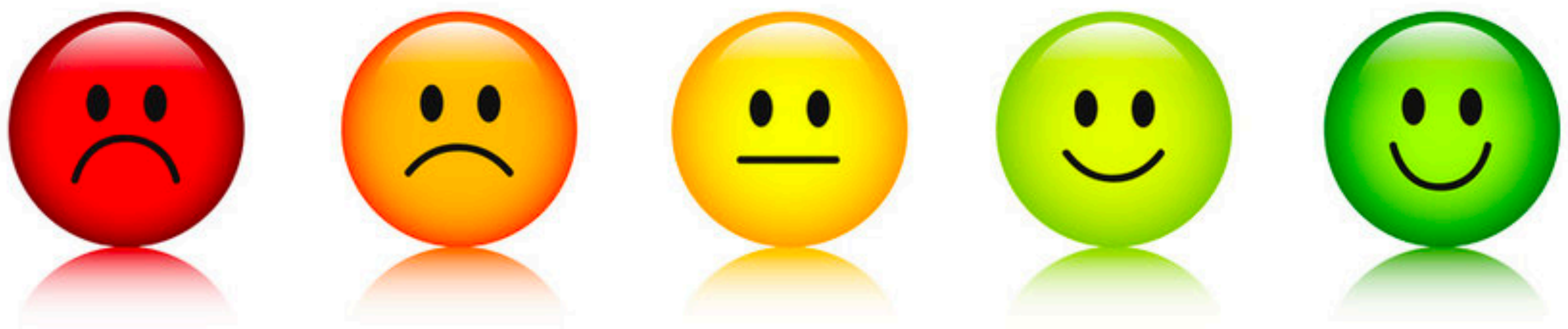
Potential applications include:

- More precise customer feedback
- Better recommendation systems
- Empathic autonomous personal assistant
- Identifying happy occasions
- Identifying accidents

Why video-only?

- Sometimes sound is not available with the video
- Noisy environments can be a challenge when relying on audio signal
- Even when the sound is available in multiple speaker setting speaker detection is a problem in its own right

How humans perceive emotions of others?



- Most humans are relatively good at classifying apparent emotional reactions

- Typically we think **categorical**:

  Happy

  Sad

  Angry

  Surprised

  Disgusted

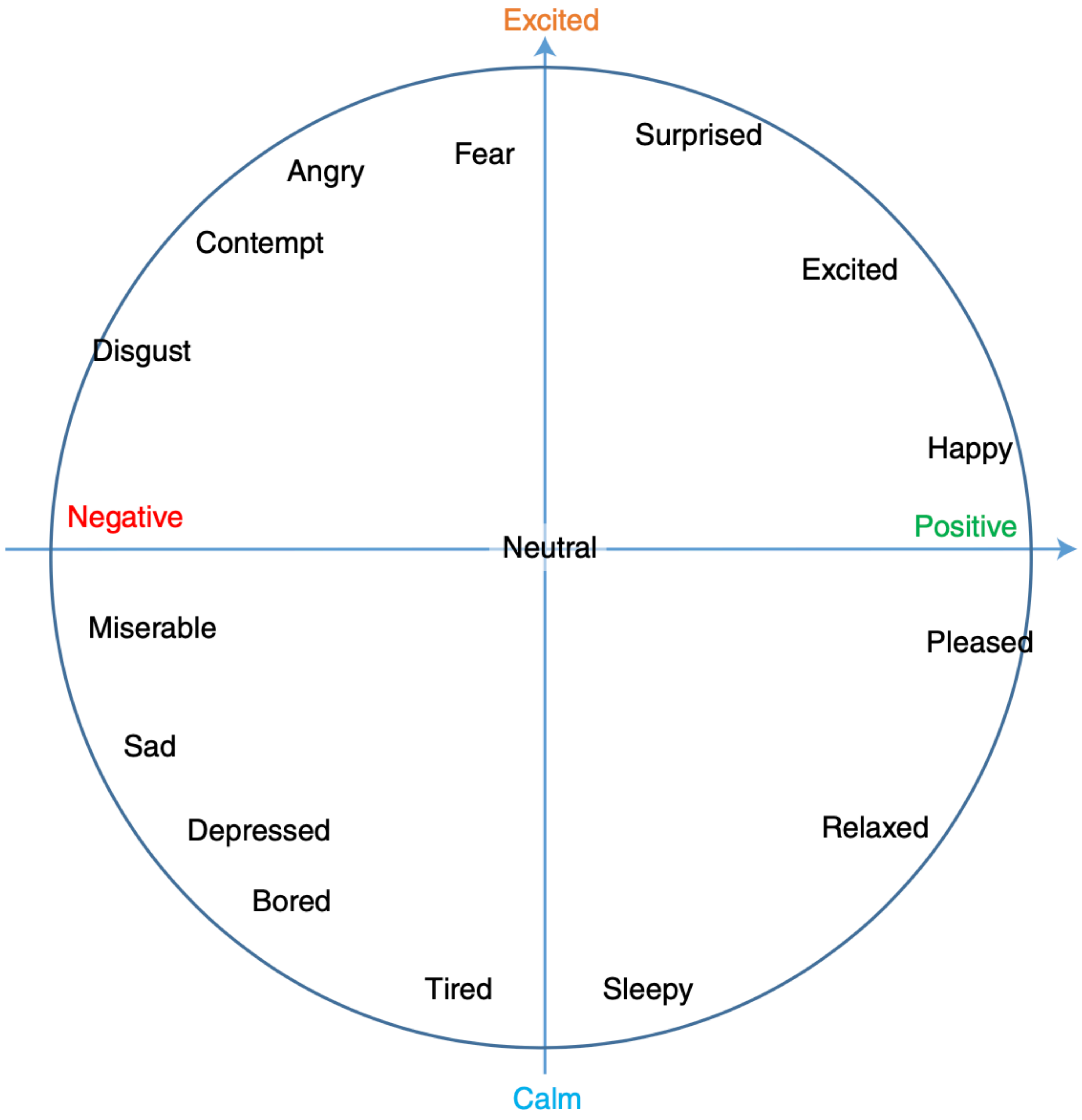
  Neutral

  etc

BUT…

- How many are there, exactly?
- Categories fail to account for
  - more subtle cues
  - mixtures of emotional reactions
- So we use continuous metrics for apparent emotional reactions —

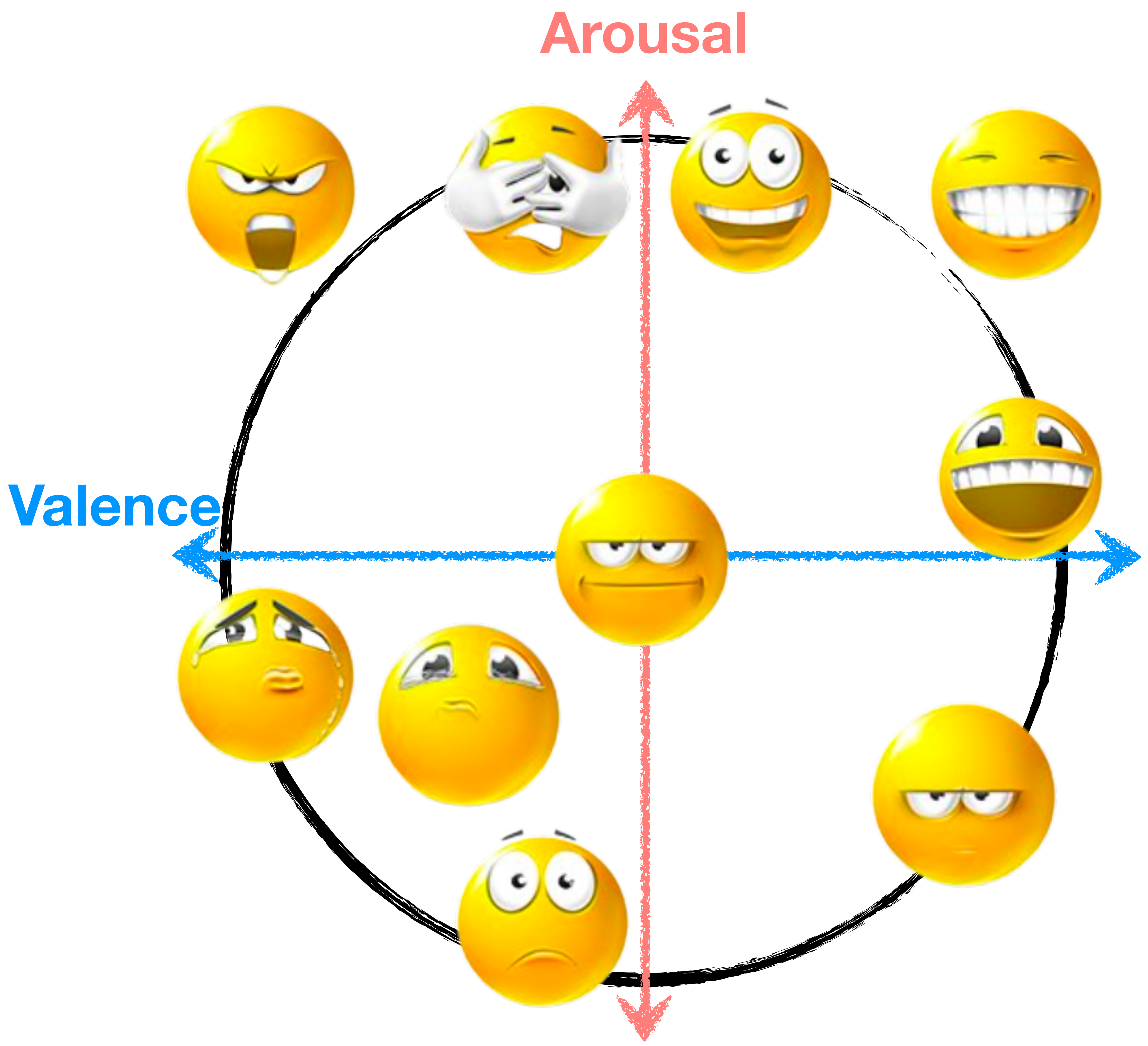
  arousal and valence



Valence ~ positivity / negativity

Arousal ~ level of excitation

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nat. Mach. Intell. 3, 1 (2021), 42–50



Valence ~ positivity / negativity

Arousal ~ level of excitation

Data & Pre-processing

Pre-text data: LRS3

Downstream data: SEWA and RECOLA

Pre-processing:

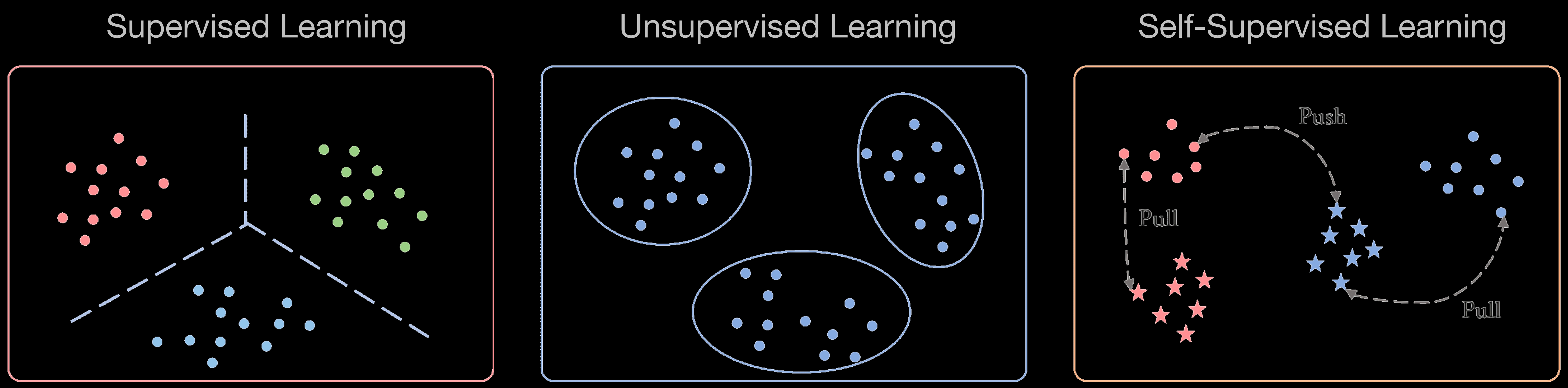
- Gray-scale
- Align & Crop based on landmarks (RetinaNet and FAN)

* SEWA dataset has been collected across the residents of 6 countries: the UK, Germany, Hungary, Serbia, Greece, and China. RECOLA is a database of multi-domain data recordings of native French-speaking participants completing a collaborative task in pairs during a video conference call, collected in France.

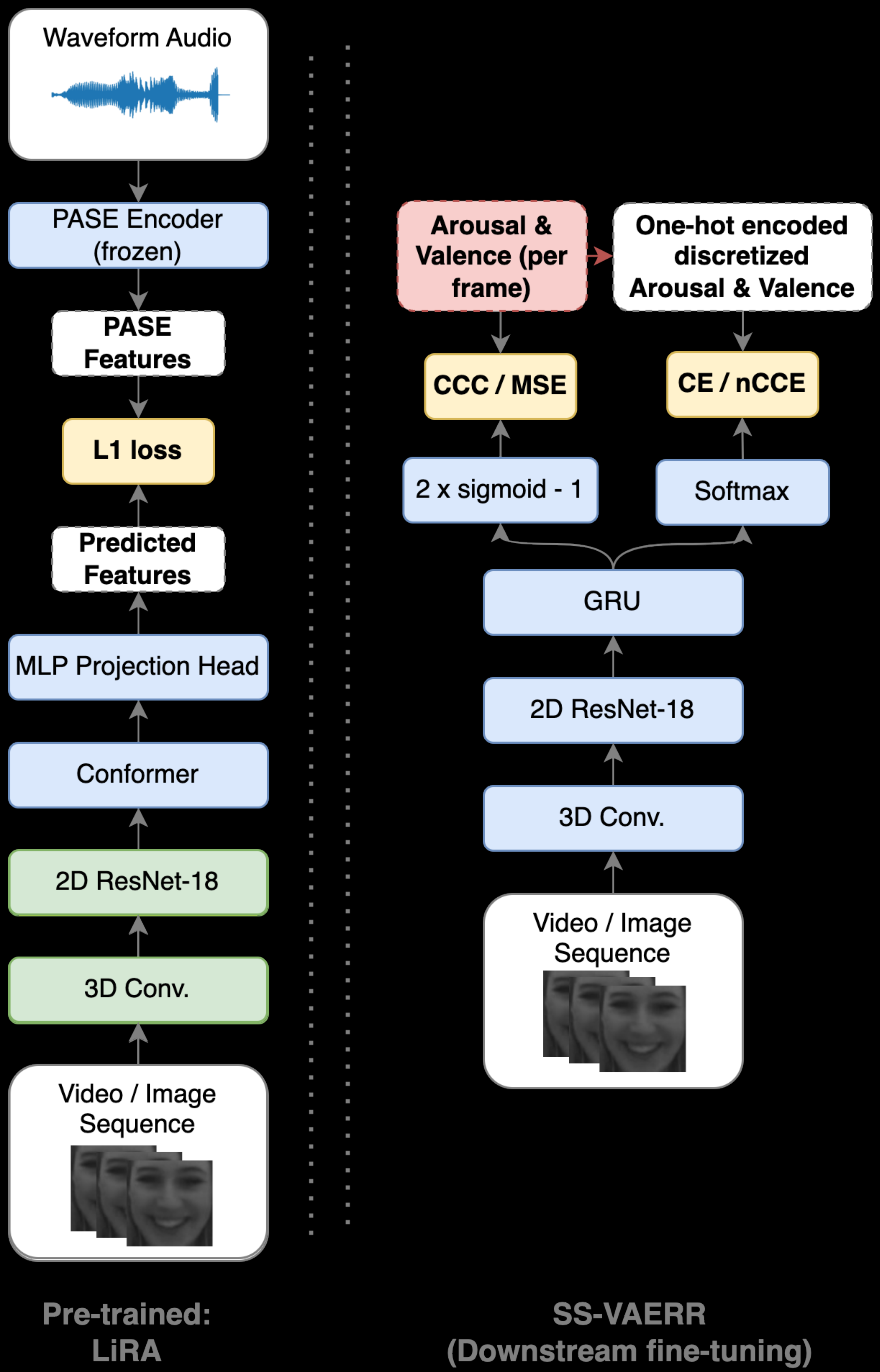
Why Self-Supervised Learning?

Self-supervised learning

- Doesn't require explicit annotations
- Can utilise databases that are not labelled for the task at hand
- Utilises potentially useful features learnt by otherwise purposed models



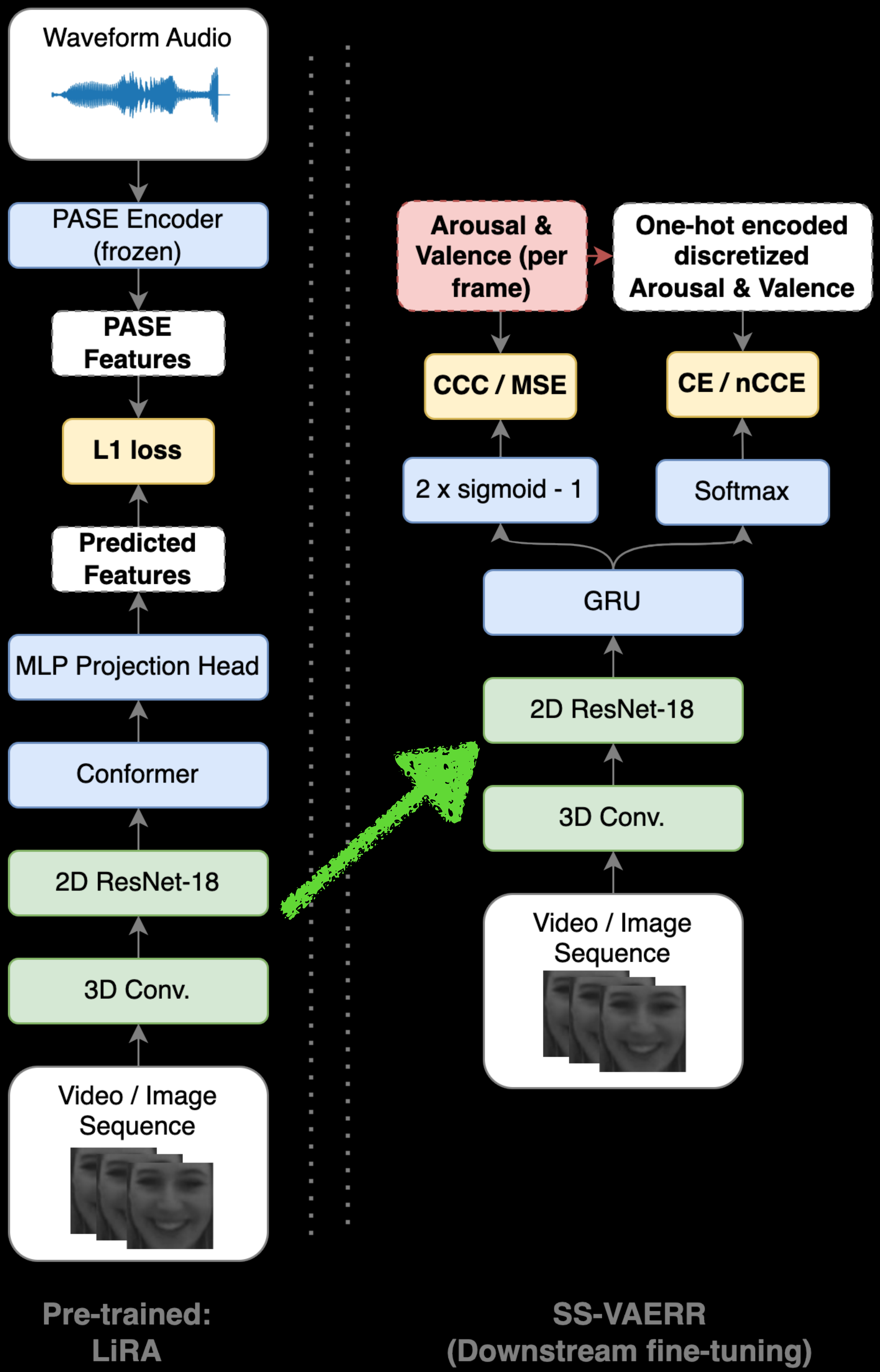
Training



1. Pretrain model (LiRA)
2. Initialise our SS-VAERR model with its weights
3. Fine-tune on a relevantly labelled dataset
4. Profit!

LiRA: Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic. 2021. LiRA: Learning Visual Speech Representations from Audio through Self-supervision.

Training



1. Pretrain model (LiRA)
2. Initialise our SS-VAERR model with its weights
3. Fine-tune on a relevantly labelled dataset
4. Profit!

LiRA: Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic. 2021. LiRA: Learning Visual Speech Representations from Audio through Self-supervision.

Losses & Metrics

Full loss:

$$\boxed{L = w_{ccc}\, CCC + w_{mse}\, MSE + w_{ce}\, CE + w_{ncce}\, nCCE}$$

Concordance Coefficient:

$$\boxed{CCC(Y,\hat{Y}) = 2\frac{\mathbb{E}(Y-\mu_Y)(\hat{Y}-\mu_{\hat{Y}})}{\sigma_{\hat{Y}} + \sigma_Y + (\mu_{\hat{Y}} + \mu_Y)^2}}$$

MSE: $MSE(\hat{Y}, Y) = \mathbb{E}((Y-\hat{Y})^2)$

Cross-Entropy: $CE(Y,\hat{Y}) = -\sum_{l=1} Y_l log(\hat{Y}_l)$

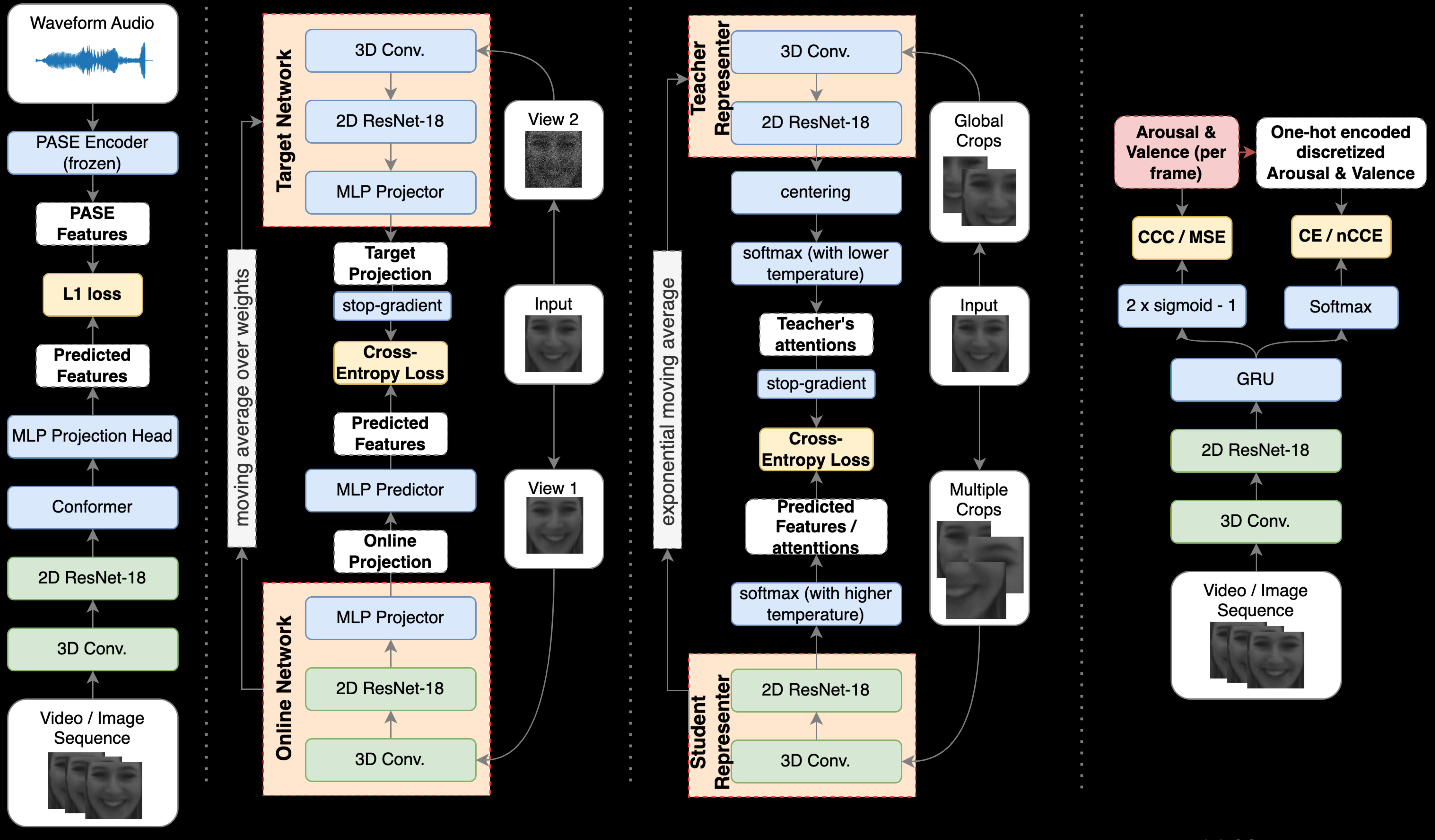
Cost-Sensitive Cross-Entropy:

$$nCCE(Y,\hat{Y}) = \frac{1}{F}\sum_{f=1}^{F} C_{norm}(Y_f, \hat{Y}_f) \sum_{l=1}^{L} Y_f{}^{(l)} \cdot log\hat{Y}_f^{(l)}$$

$$C_{norm}(Y_f, \hat{Y}_f) = 1 + \left\| \sum_{l=1}^{L} K^{(l)} (Y_f{}^{(l)} - \hat{Y}_f^{(l)}) \right\|_2$$

Duc Le, Zakaria Aldeneh, and Emily Mower Provost. 2017. Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent NeuralNetwork. In Interspeech 2017

Other Pre-train

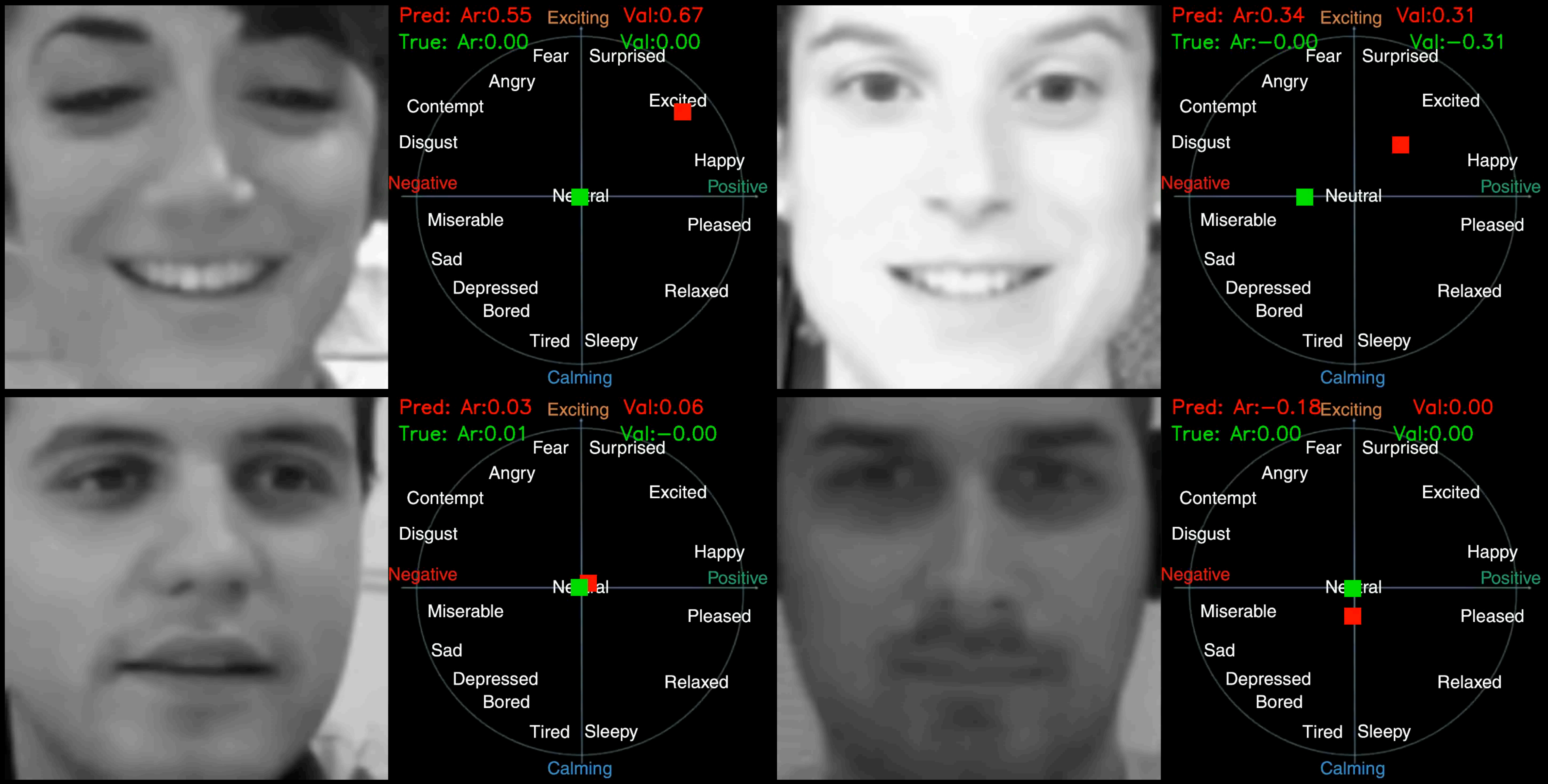


(a) LiRA (b) BYOL (c) DINO (d) SS-VAERR (Downstream fine-tuning)

LiRA: Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic. 2021. LiRA: Learning Visual Speech Representations from Audio through Self-supervision.
BYOL: Jean-Bastien Grill et al. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In NeurIPS 2020
DINO: Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transforn

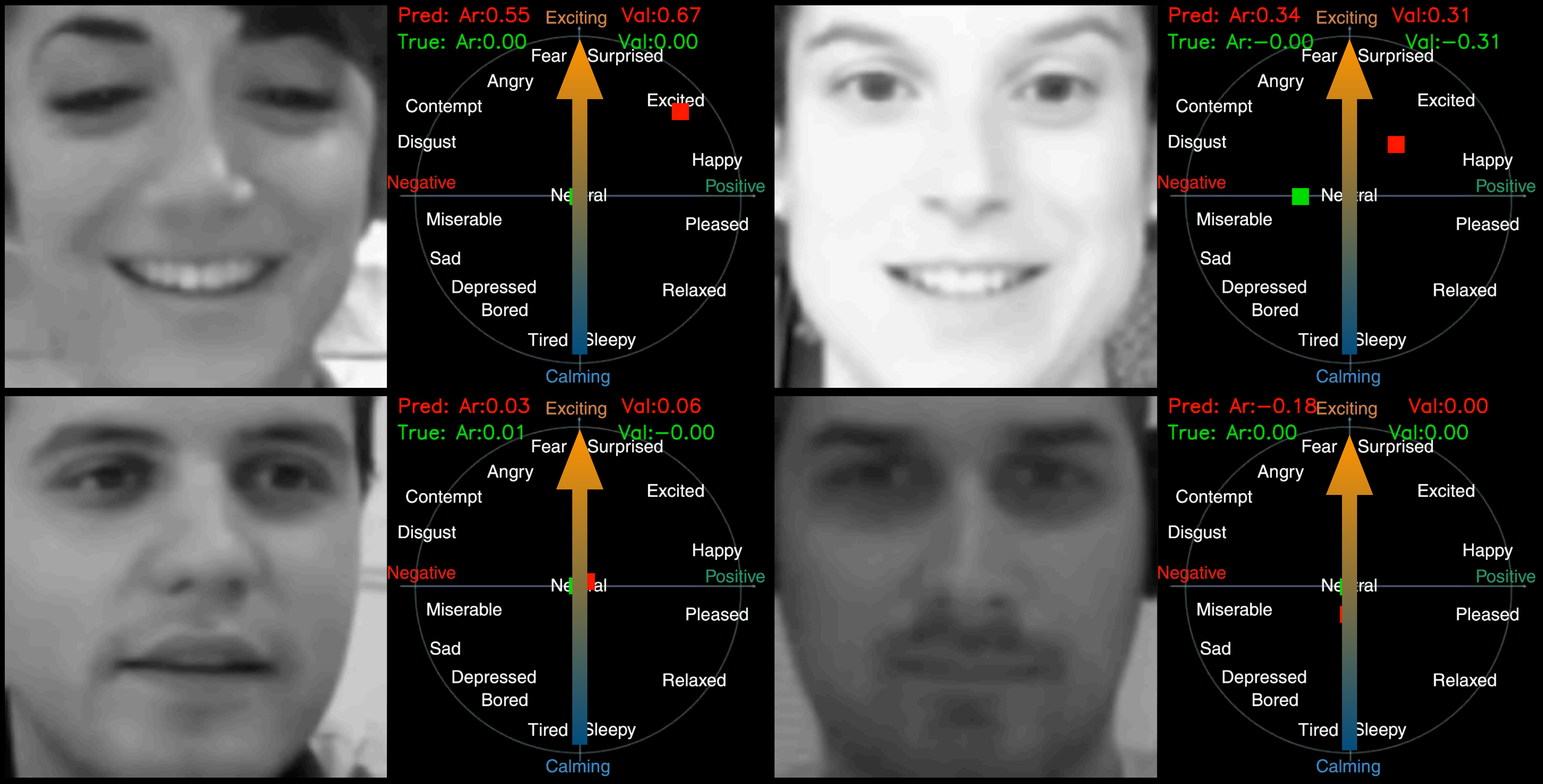
Video-only Apparent Emotional Reaction Recognition

Our model is RED DOT
Annotations are GREEN DOT



Video-only Apparent Emotional Reaction Recognition

Our model is RED DOT
Annotations are GREEN DOT



Arousal ~ Level of Excitation

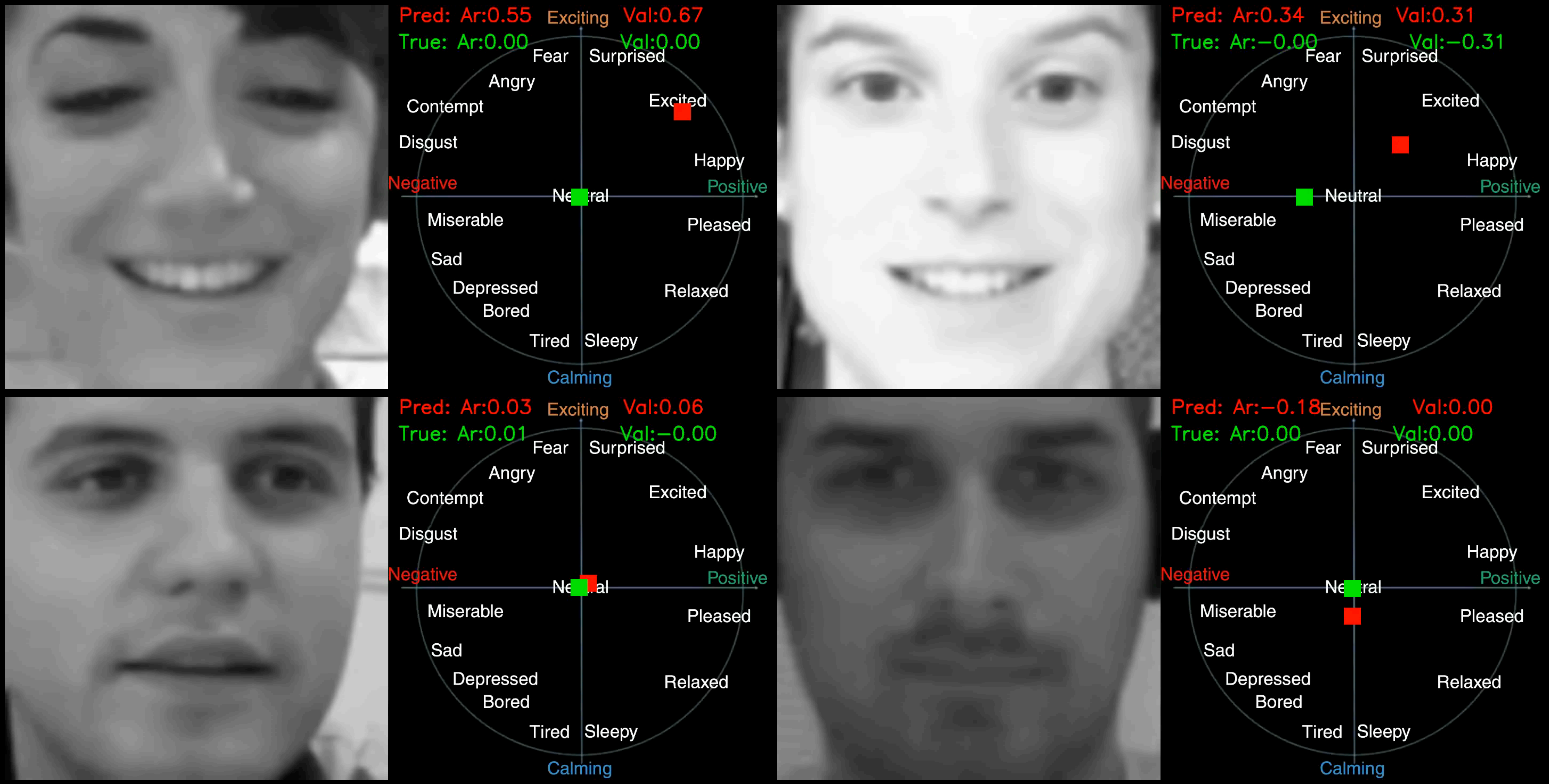
Video-only Apparent Emotional Reaction Recognition

Our model is RED DOT
Annotations are GREEN DOT

Valence ~ negativity/positivity of the reaction

Video-only Apparent Emotional Reaction Recognition

Our model is RED DOT
Annotations are GREEN DOT



Self-Supervised Video-only Apparent Emotional Reaction Recognition

TABLE I

OUR PROPOSED MODEL VS STATE-OF-THE-ART. RECOLA: RESULTS ARE PRESENTED ON DEVELOPMENT SET AS THE TEST SET IS NOT PUBLIC. FOR [2] (AP+DET.+ATT.) STANDS FOR AFFECTIVE PROCESSES WITH COMBINED LATENT AND DETERMINISTIC LAYERS WITH SELF-ATTENTION.

Methods	**SEWA**		**RECOLA**	
	Arous.	**Val.**	**Arous.**	**Val.**
HO-CPConv [5]	0.520	0.750		
Affective Processes (AP+Det.+Att.) [2]	0.662	0.672		
Affective Processes Best [2]	0.640	0.750		
End-to-End Visual ResNet-50 [7]			0.371	0.637
TS-SATCN [3]			0.659	**0.690**
Baseline: 3Dconv+ResNet18+GRU From Scratch	0.588	0.609	0.344	0.538
Our SS-VAERR backbone	0.678	0.737	0.630	0.607
Our SS-VAERR (+ augmentations + composite loss)	**0.713**	**0.771**	**0.675**	0.626

Self-Supervised Video-only Apparent Emotional Reaction Recognition

TABLE II

COMPARISON OF THE PRETEXT TECHNIQUES ACROSS VARIOUS DATASETS FOR VIDEO-ONLY AERR.

		SEWA		RECOLA	
		Arous.	**Val.**	**Arous.**	**Val.**
PRETEXT TECHNIQUES	+ LIRA frozen	0.652	0.722	0.602	0.532
	+ LIRA fine-tuned	**0.678**	**0.737**	0.630	**0.607**
	+ Video-BYOL frozen	0.593	0.726	0.224	0.344
	+ Video-BYOL fine-tuned	0.604	**0.757**	0.307	0.446
	+ DINO-ResNet frozen	0.607	0.638	0.269	0.545
	+ DINO-ResNet fine-tuned	0.648	0.667	0.420	0.520

TABLE III

COMPARISON OF THE VARIOUS LOSSES FOR THE DOWNSTREAM TASKS WITH LIRA PRE-TRAINING. ONLY NON-ZERO LOSS-WEIGHTS ARE PRESENTED. 'AROUS.' AND 'VAL.' SUPERSCRIPTS SPECIFY THE LOSS APPLIED SPECIFICALLY TO EITHER AROUSAL OR VALENCE PREDICTIONS.

		SEWA				RECOLA			
		Fine-Tuned		**Frozen**		**Fine-Tuned**		**Frozen**	
		Arous.	**Val.**	**Arous.**	**Val.**	**Arous.**	**Val.**	**Arous.**	**Val.**
REGRESSION LOSSES	$w_{ccc} = 1$	0.678	**0.737**	0.652	0.722	0.630	0.607	0.560	0.603
	$w_{mse} = 1$	0.664	0.726	0.648	0.710	0.399	0.596	0.394	0.596
COMPOSITE LOSSES	$w_{ccc} = 0.5, w_{ce} = 0.5$	0.671	**0.735**	0.650	**0.747**	0.454	0.625	0.513	0.606
	$w_{ccc} = 0.5, w_{ce} = 0.25, w_{mse} = 0.25$	**0.716**	0.731	0.699	**0.747**	0.473	0.611	0.469	0.610
	$w_{ccc}^{Val.} = 1, w_{ccc}^{Arous.} = 0.66, w_{ce}^{Arous.} = 0.34$	0.631	0.663	0.659	0.709	**0.675**	0.626	0.640	**0.668**
	$w_{ccc}^{Val.} = 1, w_{ccc}^{Arous.} = 0.66, w_{ce}^{Arous.}, w_{mse}^{Arous.} = 0.17$	0.638	0.716	0.658	0.691	0.664	0.644	**0.655**	0.605
	$w_{ccc} = 0.5, w_{n\,cce} = 0.5$	0.633	0.667	**0.701**	0.741	0.669	0.655	0.614	0.661
	$w_{ccc} = 0.5, w_{n\,cce} = 0.25, w_{mse} = 0.25$	0.669	0.716	0.633	0.733	0.606	**0.669**	0.626	0.623

Future Research Directions

- Expand to complementary categorical labels
  - More socially familiar
  - Occasionally easier for conditional responses
- Sharing parameters of temporal component (e.g. GRU) from the pretext task
  - (Currently pretext tasks use different architecture to downstream above ResNet)
  - Is likely to further improve the performance
- Generative models for de-biasing the training data

Thank you for listening.

Questions?