# Affective Processes: Stochastic Modelling of temporal context for emotion and facial expression recognition
## Supplementary Material

Enrique Sanchez[1] Mani Kumar Tellamekala[2] Michel Valstar[2] Georgios Tzimiropoulos[1,3]
[1] Samsung AI Center, Cambridge, UK
[2] University of Nottingham, Nottingham, UK
[3] Queen Mary University London, London, UK
e.lozano@samsung.com {psxmkt, michel.valstar}@nottingham.ac.uk g.tzimiropoulos@qmul.ac.uk

## 1. Backbone

This document is intended to describe the backbone structure and training, left out of the main document due to lack of space. It also includes a formal definition of the main performance metrics used in the paper, as well as the Mean Squared Error reported for the Action Unit intensity estimation task. We first describe the architecture of the backbone (Sec. 1.1), and then the training details for each of the databases used in the paper (Sec. 1.2). Sec. 2 and Sec. 3 are devoted to describing the performance metrics used in the paper and to reporting the MSE results on DISFA and BP4D, respectively.

### 1.1. Architecture

The structure of the backbone for both Valence and Arousal and Action Unit recognition is depicted in Fig. 1. Note that both are depicted in the same figure for the sake of clarity, although the corresponding subnetworks consisting of the Emotion Head or the Action Units Head are trained independently using the task specific datasets. Both networks share a common module, referred to as Face Alignment Module, which is pre-trained for the task of facial landmark localisation, and kept frozen for the subsequent training steps. For both Valence and Arousal and Action Unit estimation, the backbone is decomposed into three main components, namely *a) Face Alignment Module*, *b) Task-specific Feature Module*, and *c) Task-specific Head*.

The *Face Alignment Module* is a lightweight version of the Face Alignment Network of [2]. It starts with a 2d convolutional layer (referred to as Conv2d) and a set of 4 convolutional blocks (ConvBlock, depicted in Fig. 2) that bring down the resolution of the input image from 256 to 64 and the number of channels from 3 to 128. This set of ConvBlocks is followed by an Hourglass, a four layer set of 128-channel ConvBlocks with skip connections, that aggregate the features at different spatial scales. The Hourglass is followed by another ConvBlock and two Conv2d layers that produce a set of 68 Heatmaps, corresponding to the position of the facial landmarks. In this paper, rather than using the facial landmarks to register the face, we directly concatenate the produced features at both an early and late stage of the network with the Heatmaps. The output is then a $128+128+68$ tensor of $64 \times 64$, resulting from concatenating the features computed after the fourth ConvBlock, the features computed after the last ConvBlock, and the produced Heatmaps. This way, the Heatmaps help the subsequent network locally attend to the extracted coarse and fine features [9, 12, 15]. The benefits of this approach are twofold: a) it dispenses with the need of registering the faces according to detected landmarks, and b) because of a) we can directly use the features from the Face Alignment Network and have shallower networks in the front-end for the subsequent tasks.

The *Task specific Feature Module* consists of a mere set of 4 ConvBlocks, each followed by a max pooling layer, that produce a tensor of $128 \times 4 \times 4$. To form the features $\mathbf{x}_t$ that will be used as input to our AP network, we further downsample that tensor through an average pooling operation with a $2 \times 2$ kernel. The $128 \times 2 \times 2$ output is flattened to form the 512-d feature vector $\mathbf{x}_t$.

The *Task specific head for Valence and Arousal* is composed of four independent Conv2d layers, each with $4 \times 4$ filters (i.e. equal to the spatial resolution of the input tensor). The first Conv2d layer is the corresponding Valence and Arousal classifier $\mathbf{W}$ mentioned in the main document. The output of this layer is a 2-d vector $\hat{\mathbf{y}}_t$, corresponding to the values of Valence and Arousal, respectively. In order to boost the performance of the network for the task of predicting the continuous values of Valence and Arousal ($\hat{\mathbf{y}}$), we approach the backbone training in a Multi-task manner (see below), where the goal is to also classify the *basic (dis-*

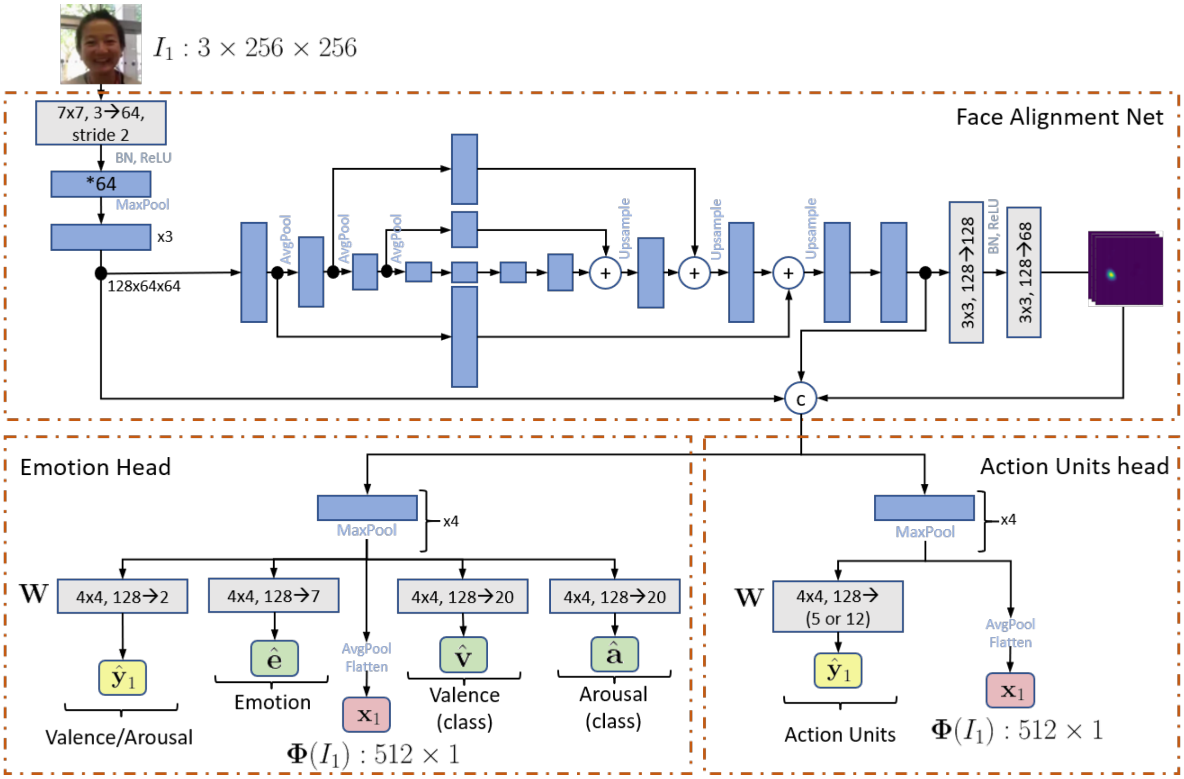

Figure 1. Architecture of the Backbone used in our AP pipeline described in the main document. For the sake of clarity, both the Emotion Head and Action Units Head are depicted together, despite these being different networks, trained separately. The grey modules represent 2-d convolutions (Conv2d), whereas the blue blocks represent Convolutional Blocks (ConvBlocks), described in Fig. 1. The $^*64$ inscribed in the first ConvBlock corresponds to a slightly different configuration that uses a skip connection to upsample the number of channels from 64 to 128. The backbone includes a Face Alignment Network, an Hourglass-like architecture that takes the input image $I$, and produces a set 68 Heatmaps corresponding to the position of the facial landmarks. The Hourglass comprises four layers of ConvBlocks with downsampling, and skip connections (for the sake of clarity we illustrate three layers, where each smaller block corresponds to halving the spatial resolution). As shown in Fig. 2, our ConvBlocks are of 128 channels, rather than the original 256 used in [2]. The Face Alignment Network is pre-trained and kept frozen, and returns a set of features resulting from concatenating the output of the last ConvBlock before the Hourglass, the output of the last ConvBlock of the network, and the produced Heatmaps. Then, the Emotion and Action Unit heads follow for each corresponding task. Both have a similar Feature Extraction Module, composed of $\times 4$ ConvBlocks followed by Average Pooling. The output of this module is a $128 \times 4 \times 4$ tensor, which is used as input to the corresponding classifiers, as well as to compute the final feature representation $\mathbf{x}_t$ that will be used along with $\hat{\mathbf{y}}_t$ as input to our proposed AP.

*crete) emotion*, as well as the bin where both Valence and Arousal would lie in a discretised space. For the basic emotion (happiness, sadness, fear, anger, surprise, disgust and neutral), we include a second Conv2d which outputs the logits corresponding to each of the 7 target classes. For the discretised Valence ($\hat{\mathbf{v}}$) and Arousal ($\hat{\mathbf{a}}$), we use two Conv2d layers with 20 outputs each, i.e. we discretise the continuous space in 20 bins, and we treat the task of predicting the corresponding bin as a classification task (see below). Note that these extra heads, as well as the emotion head, are used to reinforce the learning of the regression head tasked with predicting $\hat{\mathbf{y}}$. Once the network is trained, the heads corresponding to the discrete emotion and the discretised Valence and Arousal are removed from the backbone.

The *Task specific head for Action Unit intensity estimation* is also composed of 4 ConvBlocks as for the Valence and Arousal head. The output features, a tensor of $128 \times 4 \times 4$ are also spatially downsampled with average pooling and flattened to form the input features to the AP $\mathbf{x}_t$. The Action Units classifier $\mathbf{W}$ is a Conv2d with a $4 \times 4$ filter that maps the $128 \times 4 \times 4$ into either the 5 or 12 target AUs, for BP4D and DISFA, respectively.

## 1.2. Training

**Data processing** The faces are first cropped according to a face bounding box, provided by the off-the-shelf face detector RetinaFace [4]. Given that the first block of the backbone is a Face Alignment Network that is used to provide the features to the subsequent networks, no face registration step is applied. During training, for image augmentation

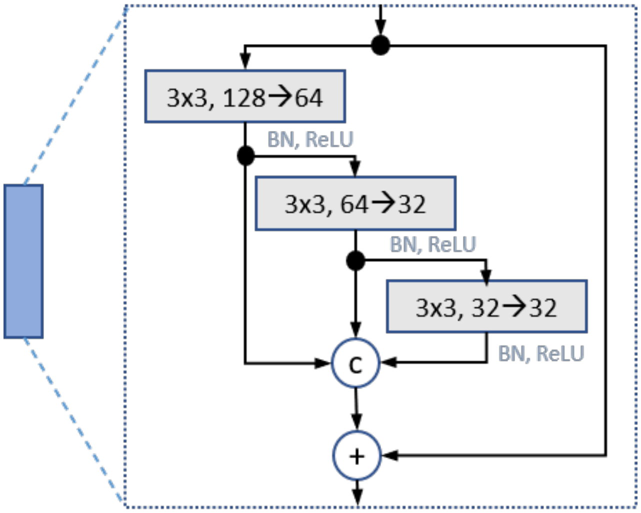

Figure 2. The Convolutional Block (ConvBlock), used in [2]. Instead of using 256 channels, we opt for a lighter version and choose 128 channels instead.

we applied random cropping (224×224), random horizontal flipping, random rotation($-20°$ to $+20°$), color jittering, and random gray scaling operations.

**Face Alignment Network** The Face Alignment Module was trained on the 300W-LP dataset [2] using standard Heatmap Regression, and was kept frozen afterwards.

**Valence and Arousal** As mentioned above, for *Valence and Arousal*, the network is trained in a Multi-task way. Let $\hat{\mathbf{y}} = (\hat{y}_v, \hat{y}_a)$ be the Valence and Arousal prediction, $\hat{\mathbf{e}} \in \mathbb{R}^7$ be the output of the discrete emotion layer, and $\hat{\mathbf{v}} \in \mathbb{R}^{20}$ and $\hat{\mathbf{a}} \in \mathbb{R}^{20}$ the output of the Valence and Arousal classes, respectively. We denote by $\mathbf{y}$ and $\mathbf{e}$ the corresponding Valence and Arousal and Emotion ground-truth values. The loss is defined as:

$$\begin{aligned}\mathcal{L} =& \lambda_{mse}\mathcal{L}_{mse} + \lambda_{ccc}\mathcal{L}_{ccc} \\ &+ \lambda_{xent-emo}\mathcal{L}_{xent-emo} \\ &+ \lambda_{xent-va}\mathcal{L}_{xent-va}\end{aligned} \tag{1}$$

where $\mathcal{L}_{mse} = \|\hat{\mathbf{y}} - \mathbf{y}\|$ is the standard MSE loss for Valence and Arousal, $\mathcal{L}_{ccc} = 1 - \frac{CCC(\hat{y}_v, y_v) + CCC(\hat{y}_a, y_a)}{2}$ is the CCC score between the predicted Valence and Arousal values and corresponding ground-truth, $\mathcal{L}_{xent-emo}$ is the standard cross entropy loss between the predicted emotion $\hat{\mathbf{e}}$ and the corresponding ground-truth $\mathbf{e}$. We define $\mathcal{L}_{xent-va} = \mathcal{L}_{xent-v} + \mathcal{L}_{xent-a}$, with $\mathcal{L}_{xent-v}$ the cross entropy loss between the 20-d output of the Valence head and the corresponding ground-truth bin, and equivalently $\mathcal{L}_{xent-a}$ for Arousal. The ground-truth bin results from uniformly discretising the Valence and Arousal spaces, which lie within the $[-1, 1]$ space, into 20 bins each.

The values of the loss weights are all set to 1 except for the MSE loss that is set to $\lambda_{mse} = 0.5$. For both SEWA and AffWild2, the training is performed for 20 epochs, using Adam with learning rate 0.0001, $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 0.000001. The learning rate is reduced by a factor of 10 after every 5 epochs.

For AffWild2 we used the sequences that were annotated with both discrete emotion and Valence and Arousal. Considering that SEWA has not been annotated with the basic emotions, we train our SEWA backbone by extending it with the sequences of AffWild2 containing such annotations. We backpropagate w.r.t. the emotion head using images from AffWild2, and w.r.t. the remaining heads using only images from SEWA. We apply the same 8:1:1 partition described in the paper, and choose the backbone according to the best validation CCC score.

**Action Units** For *Action Unit* intensity estimation, Mean Squared Error is used as the loss function to train the corresponding models in this work (for BP4D and DISFA). The AU intensities are normalised from -1 to 1 to align with the $\mathcal{L}_{reg}$ used in the AP framework described in the main document. Adam optimizer with a learning rate of 0.0003, $(\beta_1, \beta_2) = (0.9, 0.999)$, and an L2 weight decay of 0.00001 is used to train the Action Unit head. To tune the initial learning rate, cyclic learning rate scheduler with a cycle length of 2 is used. After training for 80 epochs, the best model is selected based on the highest ICC score on the validation set.

For BP4D, the model is trained using the official train/validation/test partitions. For DISFA, the model is trained using the three-fold cross validation method described in the main document, using exactly the same generated partitions.

## 2. Performance Metrics

For Valence and Arousal, we report the Concordance Correlation Coefficient [6], which is used to rank participants in the AVEC Challenge series [10]. It is a global measure of both correlation and proximity, and is defined as:

$$CCC(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}\rho_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}^2 + \sigma_{\hat{\mathbf{y}}}^2 + (\mu_{\mathbf{y}} - \mu_{\hat{\mathbf{y}}})^2}, \tag{2}$$

where $\mu$, $\sigma$, and $\rho$ refer to the mean value, (co-)variance, and Pearson Correlation Coefficient, respectively.

For Action Unit intensity, we follow the standard ranking criteria used in FERA challenges[13], and we report the Intra Class Correlation (ICC [11]). For an AU $j$ with ground-truth labels $\{y_i^j\}_{i=1}^N$, and predictions $\{\tilde{y}_i^j\}_{i=1}^N$, the ICC score is defined as $ICC^j = \frac{W^j - S^j}{W^j + S^j}$, with $W^j = \frac{1}{N}\sum_i \left((y_i^j - \hat{y}^j)^2 + (\tilde{y}_i^j - \hat{y}^j)^2\right)$, $S^j = \sum_i (y_i^j - \tilde{y}_i^j)^2$, and $\hat{y}^j = \frac{1}{2N}\sum_i (y_i^j + \tilde{y}_i^j)$.

| AU | 1 | 2 | 4 | 5 | 6 | 9 | 12 | 15 | 17 | 20 | 25 | 26 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGP-AE [5] | 0.51 | **0.32** | 1.13 | 0.08 | 0.56 | 0.31 | 0.47 | 0.20 | 0.28 | **0.16** | 0.49 | 0.44 | 0.41 |
| 2DC [7] | **0.32** | 0.39 | 0.53 | 0.26 | 0.43 | 0.30 | **0.25** | 0.27 | 0.61 | 0.18 | 0.37 | 0.55 | 0.37 |
| HR [9] | 0.41 | 0.37 | 0.70 | 0.08 | 0.44 | 0.30 | 0.29 | 0.14 | 0.26 | **0.16** | 0.24 | 0.39 | 0.32 |
| Ours backbone | 0.93 | 0.90 | 0.51 | 0.04 | 0.44 | 0.19 | 0.30 | 0.13 | **0.21** | 0.17 | **0.23** | 0.29 | 0.36 |
| BiGRU [3]$^\dagger$ | 0.85 | 0.79 | 0.48 | 0.06 | 0.47 | 0.19 | 0.34 | 0.18 | 0.23 | 0.21 | 0.30 | 0.40 | 0.37 |
| Self-Attn [14]$^\dagger$ | 0.76 | 0.71 | 0.52 | 0.04 | **0.42** | 0.17 | 0.35 | 0.14 | **0.21** | 0.19 | 0.28 | 0.36 | 0.34 |
| Ours AP | 0.68 | 0.59 | **0.40** | **0.03** | 0.49 | **0.15** | 0.26 | **0.13** | 0.22 | 0.20 | 0.35 | **0.17** | **0.30** |

Table 1. Results on the DISFA database (in MSE values) $^\dagger$ denotes in-house evaluation

## 3. Mean Squared Error Results

The additional Mean Squared Error results for DISFA and BP4D are reported in Table 1 and Table 2.

| **AU** | **6** | **10** | **12** | **14** | **17** | **Avg.** |
|---|---|---|---|---|---|---|
| CDL [1] | - | - | - | - | - | - |
| ISIR [8] | 0.83 | 0.80 | 0.62 | 1.14 | 0.84 | 0.85 |
| HR [9] | **0.68** | **0.80** | 0.79 | **0.98** | 0.64 | 0.78 |
| Ours backbone | 0.80 | 0.87 | 0.74 | 1.23 | 0.89 | 0.90 |
| BiGRU [3]$^\dagger$ | 0.79 | 0.85 | 0.76 | 1.19 | 0.78 | 0.87 |
| Self-Attn [14]$^\dagger$ | 0.82 | 0.88 | 0.70 | 1.22 | 0.80 | 0.88 |
| Ours AP | 0.72 | 0.84 | **0.60** | 1.13 | **0.57** | **0.77** |

Table 2. Results on the test partition of BP4D dataset (in MSE values) $^\dagger$ denotes in-house evaluation

## References

[1] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *IEEE Conf. and Workshops on Auto. Face and Gesture Recog.*, volume 6, pages 1–6. IEEE, 2015. 4

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Int. Conf. Comput. Vis.*, pages 1021–1030, 2017. 1, 2, 3

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 4

[4] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 2

[5] Stefanos Eleftheriadis, Ognjen Rudovic, Marc Peter Deisenroth, and Maja Pantic. Variational gaussian process auto-encoder for ordinal prediction of facial action units. In *ACCV*, pages 154–170. Springer, 2016. 4

[6] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989. 3

[7] Dieu Linh Tran, Robert Walecki, Stefanos Eleftheriadis, Bjorn Schuller, Maja Pantic, et al. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *Int. Conf. Comput. Vis.*, pages 3190–3199, 2017. 4

[8] Jeremie Nicolle, Kevin Bailly, and Mohamed Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *IEEE Conf. and Workshops on Auto. Face and Gesture Recog.*, volume 6, pages 1–6. IEEE, 2015. 4

[9] I. Ntinou, E. Sanchez, A. Bulat, M. Valstar, and G. Tzimiropoulos. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Transactions on Affective Computing*, pages 1–1, 2021. 1, 4

[10] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019. 3

[11] P.E. Shrout and J.L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 1979. 3

[12] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021. 1

[13] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *IEEE Conf. and Workshops on Auto. Face and Gesture Recog.*, volume 6, pages 1–8. IEEE, 2015. 3

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017. 4

[15] Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Fan-face: a simple orthogonal improvement to deep face recognition. In *AAAI Conference on Artificial Intelligence*, 2020. 1