# Guided Interpretable Facial Expression Recognition via Spatial Action Unit Cues (#29)

Soufiane Belharbi[1*] & Marco Pedersoli[1] & Alessandro Lameiras Koerich[1] & Simon Bacon[2] & Eric Granger[1]

[1]LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
[2]Dept. of Health, Kinesiology & Applied Physiology, Concordia University, Montreal, Canada

## Context / Problem

☞ Facial expression recognition (FER) models lack **interpretability**

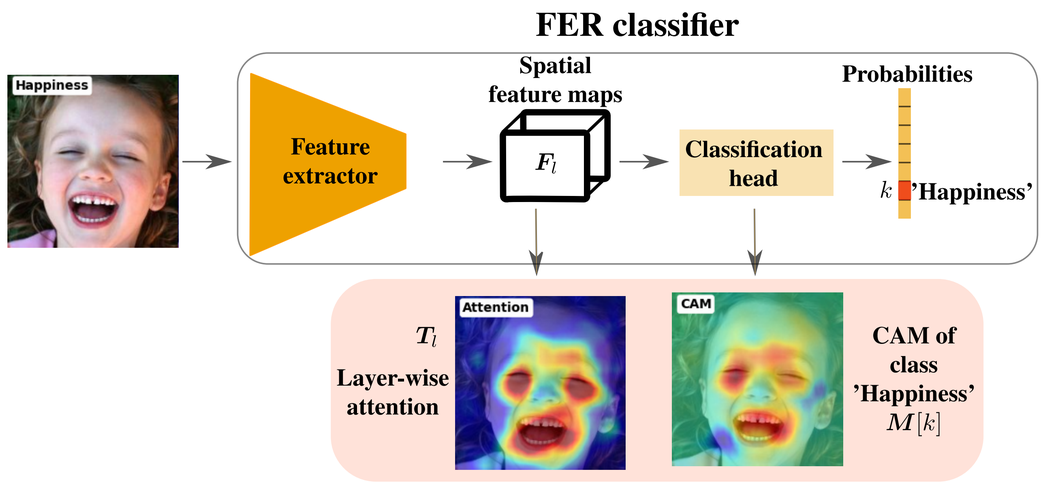


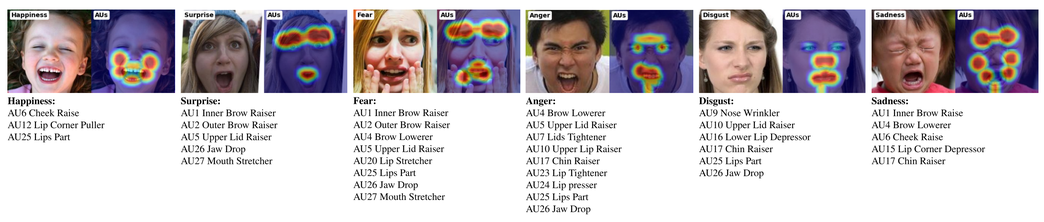

*Figure 1. Poor interpretability of FER classifiers.*

☞ Our goal: Design an interpretable deep classifier for FER

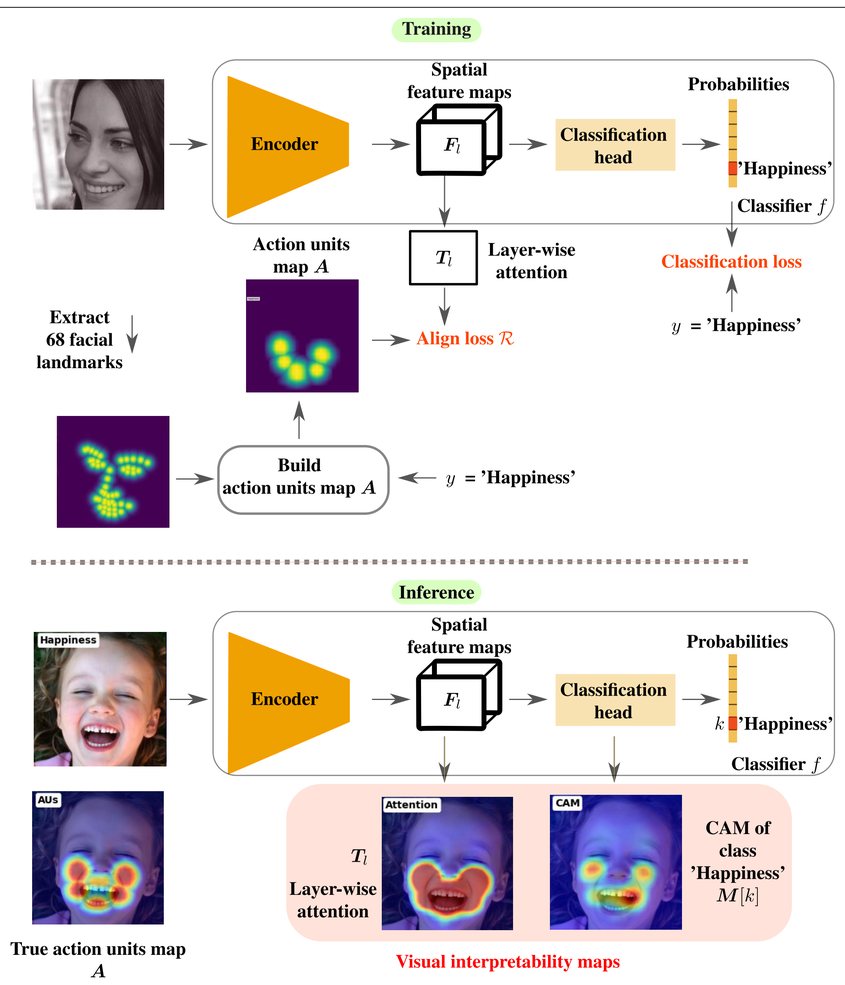
- ☞ Using similar cues used by experts: Facial Action Units (**AUs**)
- ☞ Allow the model to provide visual interpretability
- ☞ Maintain high classification performance

☞ Proposal:

- ☞ Build **AU** heatmaps for training: Requires only image class label
- ☞ Align layer-wise attention with **AU** heatmaps




## Proposal: Architecture




## Proposal: Training Loss

☞ Total loss: classification, and spatial **AU** alignment,

$$\min_{\boldsymbol{\theta}} \quad -\log(f(\boldsymbol{X};\boldsymbol{\theta})_y) + \lambda(1-\mathcal{R}(\boldsymbol{T}_l,\boldsymbol{A}))\ , \tag{1}$$

where,

$$\mathcal{R}(\boldsymbol{T}_l,\boldsymbol{A}) = \frac{\sum(\boldsymbol{T}_l \odot \boldsymbol{A})}{\|\boldsymbol{T}_l\|_2\|\boldsymbol{A}\|_2}\ , \tag{2}$$

is the cosine similarity between the **AU** map $\boldsymbol{A}$ and the layer-wise attention $\boldsymbol{T}_l$ that is computed as,

$$\boldsymbol{T}_l = \frac{1}{n}\sum_{j=0}^{n}\boldsymbol{F}_l[j]\ , \tag{3}$$

over the layer $l$ and its spatial features $\boldsymbol{F}_l$.

## Empirical Results: RAF-DB & AffectNet Datasets

☞ Classification (**CL**) and CAM-localization (**CAM-COS**) performance on **RAF-DB** and **AffectNet** test sets with and without **AUs** across methods:

| Method | RAF-DB | | | | AffectNet | | | |
|---|---|---|---|---|---|---|---|---|
| | CL | | CAM-COS | | CL | | CAM-COS | |
| | w/o AU | w/ AU | w/o AU | w/ AU | w/o AU | w/ AU | w/o AU | w/ AU |
| CNN-based | | | | | | | | |
| CAM *(cvpr,2016)* | 88.20 | **88.95** | 0.55 | **0.70** | 60.88 | **62.37** | 0.56 | **0.69** |
| WILDCAT *(cvpr,2017)* | 88.26 | **88.85** | 0.52 | **0.69** | 59.88 | **61.62** | 0.62 | **0.80** |
| GradCAM *(iccv,2017)* | 88.39 | **88.85** | 0.55 | **0.74** | 60.77 | **62.08** | 0.53 | **0.75** |
| GradCAM++ *(wacv,2018)* | 87.84 | **89.14** | 0.60 | **0.82** | 60.22 | **62.45** | 0.66 | **0.83** |
| ACoL *(cvpr,2018)* | 87.94 | **88.68** | 0.54 | **0.67** | 58.28 | **61.48** | 0.55 | **0.65** |
| PRM *(cvpr,2018)* | 88.13 | **88.88** | 0.48 | **0.59** | 57.77 | **60.97** | 0.52 | **0.75** |
| ADL *(cvpr,2019)* | 87.45 | **88.65** | 0.50 | **0.63** | 57.88 | **61.25** | 0.54 | **0.66** |
| CutMix *(eccv,2019)* | 88.39 | **88.59** | 0.55 | **0.57** | 58.74 | **59.88** | 0.56 | **0.58** |
| LayerCAM *(ieee,2021)* | 87.90 | **88.88** | 0.60 | **0.84** | 60.77 | **62.45** | 0.66 | **0.83** |
| Transformer-based | | | | | | | | |
| TS-CAM *(iccv,2021)* | 86.70 | **88.00** | 0.58 | **0.71** | 58.99 | **59.54** | 0.57 | **0.58** |
| APViT *(ieee,2022)* | 91.00 | **91.03** | –– | –– | 60.62 | **62.28** | –– | –– |

☞ Attention-localization (**ATT-COS**) (at layer 5) performance over **RAF-DB** and **AffectNet** test sets with and without **AUs**:

| Methods / Case | RAF-DB | | AffectNet | |
|---|---|---|---|---|
| | w/o AU | w/ AU | w/o AU | w/ AU |
| CNN-based | | | | |
| CAM*(cvpr,2016)* | 0.57 | **0.85** | 0.64 | **0.82** |
| WILDCAT *(cvpr,2017)* | 0.47 | **0.85** | 0.61 | **0.81** |
| GradCAM *(iccv,2017)* | 0.63 | **0.85** | 0.65 | **0.82** |
| GradCAM++ *(wacv,2018)* | 0.52 | **0.87** | 0.65 | **0.82** |
| ACoL *(cvpr,2018)* | 0.46 | **0.84** | 0.60 | **0.81** |
| PRM *(cvpr,2018)* | 0.43 | **0.85** | 0.55 | **0.82** |
| ADL *(cvpr,2019)* | 0.51 | **0.85** | 0.65 | **0.83** |
| CutMix *(eccv,2019)* | 0.51 | **0.80** | 0.57 | **0.82** |
| LayerCAM *(ieee,2021)* | 0.52 | **0.86** | 0.65 | **0.82** |
| Transformer-based | | | | |
| TS-CAM *(iccv,2021)* | 0.55 | **0.88** | 0.48 | **0.79** |
| APViT *(ieee,2022)* | 0.38 | **0.85** | 0.45 | **0.84** |

☞ Visual Results: Better interpretability

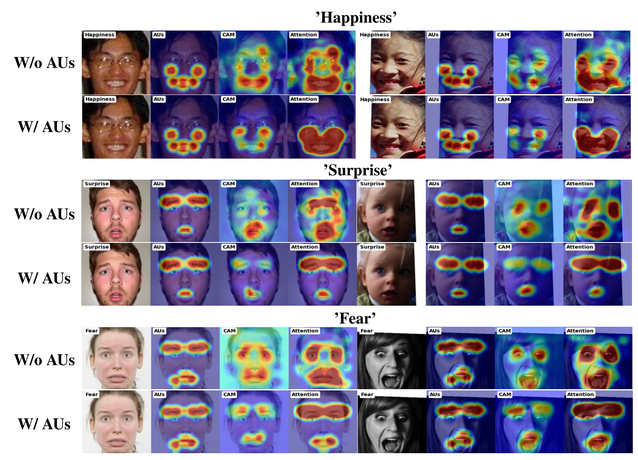



*Figure 2. Illustration of interpretability prediction over **RAF-DB** test samples using CAM method with and without action units alignment. From left to right: Input image, true action units map $\boldsymbol{A}$, CAM $\boldsymbol{M}[k]$, attention $\boldsymbol{T}_5$.*

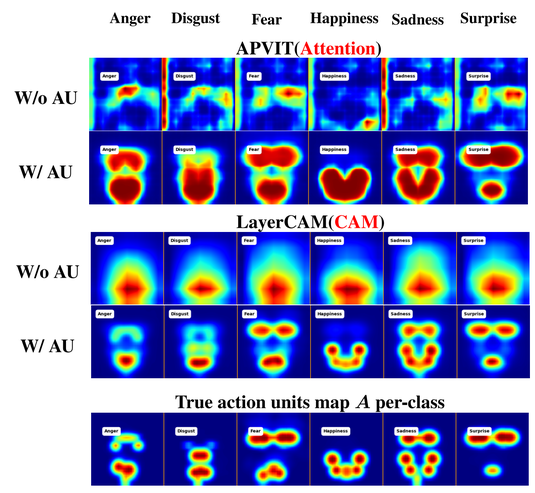



*Figure 3. Illustration of **per-class average attention and CAM maps over all test set of **RAF-DB** with and without **AU** alignment. Expressions from left to right: 'Anger', 'Disgust', 'Fear', 'Happiness', 'Sadness', 'Surprise'.*

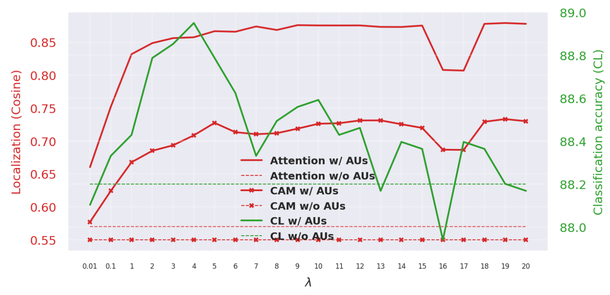



*Figure 4. Ablation on the **RAF-DB** test set: impact of $\lambda$ over classification and localization (interpretability) performance.*