

# Combining facial videos and biosignals for stress estimation during driving

Paraskevi Valergaki<sup>2</sup>, Vassilis C. Nicodemou<sup>2</sup>, Iason Oikonomidis<sup>2</sup>, Antonis Argyros<sup>1,2</sup>, and Anastasios Roussos<sup>2</sup>

<sup>1</sup> Computer Science Department, University of Crete, Heraklion, Greece

<sup>2</sup> Institute of Computer Science (ICS), Foundation for Research & Technology –

Hellas (FORTH), Heraklion, Greece

{vbalerg,nikodom,argyros,troussos}@ics.forth.gr

**Abstract.** Reliable stress recognition from facial videos is challenging due to stress’s subjective nature and voluntary facial control. While most methods rely on Facial Action Units, the role of disentangled 3D facial geometry remains underexplored. We address this by analyzing stress during distracted driving using EMOCA-derived 3D expression and pose coefficients. Paired hypothesis tests between baseline and stressor phases reveal that 41 of 56 coefficients show consistent, phase-specific stress responses comparable to physiological markers. Building on this, we propose a Transformer-based temporal modeling framework and assess unimodal, early-fusion, and cross-modal attention strategies. Cross-Modal Attention fusion of EMOCA and physiological signals achieves best performance (AUROC 92%, Accuracy 86.7%), with EMOCA-gaze fusion also competitive (AUROC 91.8%). This highlights the effectiveness of temporal modeling and cross-modal attention for stress recognition.

**Keywords:** Stress Recognition · Classification · Transformers.

## 1 Introduction

Stress recognition from facial videos is a problem of growing importance across domains such as human-computer interaction, affective computing [13], healthcare [28], and intelligent transportation systems [27]. In particular, driver stress monitoring has attracted significant attention due to its potential to improve road safety. Despite substantial progress, reliable video-based stress detection remains challenging. Stress is a complex process shaped by individual differences, contextual factors, and physiological reactivity. Moreover, facial expressions are partially voluntary and can be consciously suppressed or masked, further complicating visual inference [9].

Most existing facial stress recognition approaches rely on physiological signals, 2D appearance cues or Facial Action Units (AUs). While effective, these representations often entangle identity, pose, and expression, and provide limited insight into the underlying 3D facial dynamics associated with stress. Less works have examined deep 3D geometric facial features, despite evidence that head pose and subtle expression dynamics are sensitive stress indicators.

In this work, we study stress under distracted driving conditions using disentangled 3D expression and pose coefficients extracted with EMOCA [3] from infrared facial videos of a large-scale multimodal dataset. We first conduct a statistical analysis, performing paired t-tests between baseline and stressor phases to examine how 3D facial dynamics, physiological signals (heart rate, breathing rate, perinasal perspiration). Building on these findings, we propose a Transformer-based temporal modeling framework and systematically evaluate multimodal fusion strategies. Our results show that cross-modal attention fusion most effectively captures the interaction between facial dynamics and physiological responses, achieving state-of-the-art stress recognition performance.

Overall, this study highlights the importance of disentangled 3D facial geometry for stress analysis and provides a unified statistical and learning-based framework for multimodal stress recognition in driving scenarios.

## 2 Related Work

**Stress Recognition Datasets.** A number of datasets have been proposed for stress recognition, differing in modalities, elicitation protocols, and annotation strategies. Early datasets such as SUS [29] focus on unimodal audio recordings, without self-assessments. Video-only resources such as SADVAW [31] derive stress labels from external annotations of movie clips. Several datasets emphasize physiological sensing including WeSAD [26], and CLAS [22], which collect signals such as ECG, EDA, EMG, respiration, and acceleration under stress-inducing protocols such as driving tasks and audiovisual stimuli. More recent efforts have explored multimodal stress recognition. MuSE [14] and SWELL-KW [17] combine physiological measurements with audio and video data in laboratory settings involving public speaking or office-work scenarios, but are limited in scale, with recordings from fewer than 30 participants. UBFC-Phys [24], the Distracted Driving [30] and StressID [2] provide physiological signals and facial video from many subjects with the latter providing speech recordings, as well.

**Facial Action Unit-Based Stress Detection.** Recent work has explored facial action units (AUs) and expression-related cues as visual indicators of stress, often leveraging machine learning and deep learning models [33] [16], [35], [7], [15], [11], [32], [10], [18], [8], [4], [1]. In [11] the authors proposed a deep pipeline for AU detection, consisting of the steps of preprocessing (face detection, landmarking and 2D image registration), feature extraction (deep geometric and appearance features) and deep AU classification reporting a stress vs. neutral classification accuracy of 81.1 %. Using an attention-based two-level architecture (TSDNet), Zhang et al. [35] achieved 78.62% accuracy for face-only stress detection, with further gains (85.42% accuracy) when incorporating action motion cues. In [4] the authors reported a regression-based stress estimation model achieving a Pearson correlation of 0.539, highlighting the relevance of facial cues to perceived stress from facial videos of 240 participants. MTASR [34] is a pipeline of extracting rPPG signals from raw RGB videos and employing multi-task attentional learning for stress recognition which achieves 94.33% ac-

curacy for stress state and 83.83% for stress level recognition on UBFC-Phys. Koujan et al.[18] has demonstrated that disentangling expression from identity using 3D morphable face models achieves SOTA performance in recognizing basic emotions from in-the-wild images, while also supporting stress recognition from facial videos. Complementary studies [12] analyze 3D head pose dynamics estimated from facial landmarks and indicate that stress increases head mobility.

**Driving Behavior Classification.** FMDNet is a feature-attention-based multimodal driving behavior classification network [21] evaluated on the UAH-DriveSet fusing vehicle dynamics (acceleration, roll/pitch/yaw, speed) and roadside videos, but without any facial or behavioral cues from the driver. Vehicle-speed spectrograms are employed in [20] to categorize driving behaviors (normal, aggressive, drowsy). Comparably to the previous approach, Mou et al. [23] designed a dual-channel CNN-Transformer on the distracted driving dataset, integrating eye-related signals, physiological measurements and vehicle dynamics.

### 3 Methodology

#### 3.1 Dataset Description

We use the multimodal distracted driving dataset introduced by Taamneh et al. [30], which contains synchronized facial video, physiological measurements, gaze signals, driving behaviour traces, and detailed experimental annotations. The study involved sixty-eight volunteers, grouped into young (18–27 years) and older (above 60 years) participants, and was conducted in a driving simulator.

The experiment was designed to evaluate the effects of three distraction types: Cognitive, Emotional, and Sensorimotor. Multiple sensing modalities were recorded concurrently: a thermal Tau 640 infrared camera for perinasal perspiration, a FireWire CCD monochrome camera for facial video, a Shimmer3 GSR sensor for palm electrodermal activity, a Zephyr BioHarness 3.0 for heart rate and breathing rate, and the faceLAB eye-tracking system for gaze trajectories. Additional vehicular signals such as steering, braking, and lane position were logged.

In this work, we focus on two sessions: the Normal Driving (ND) serving as a baseline and the Sensorimotor Distraction (MD) loaded drive. The MD session consists of five alternating non-stressor and stressor phases (P1–P5). Sensorimotor stressors involve texting back words, sent one by one to the subject’s smartphone, administered during P2 and P4, with P1, P3, and P5 functioning as non-stress intervals.

#### 3.2 Feature Extraction with EMOCA

To obtain a detailed representation of facial behaviour during driving, we extract per-frame facial parameters using EMOCA [3,6], a SOTA emotion-driven monocular 3D face reconstruction framework built upon DECA [5]. EMOCA regresses FLAME [19] expression, pose, and shape parameters constrained by

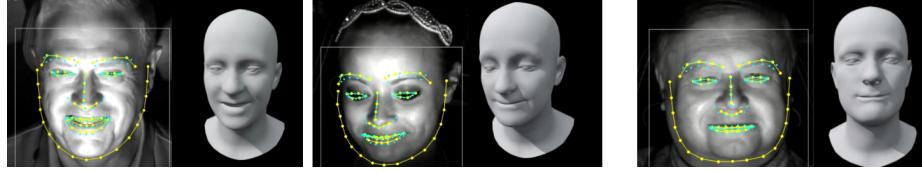


Fig. 1: *EMOCA feature visualization*. Left: original infrared video frames with MediaPipe facial landmarks projected on them. Right: FLAME mean-face mesh rendered using the corresponding EMOCA expression and pose coefficients. By projecting the estimated expression and pose parameters onto a fixed mean identity, facial dynamics are disentangled from subject-specific shape.

an emotion-consistency loss that enhances the fidelity of expression-related geometry.

For each frame of the Normal Driving (ND) and Sensorimotor Distraction (MD) videos, EMOCA outputs a 56-dimensional facial parameter vector  $\mathbf{x}_t = [\mathbf{e}_t \in \mathbb{R}^{50}, \mathbf{p}_t \in \mathbb{R}^6]$ , where  $\mathbf{e}_t$  encodes facial expression and  $\mathbf{p}_t$  represents global head pose (yaw, pitch, roll, jaw rotation, and translation). Although the FLAME model defines a 100-dimensional expression space, EMOCA regresses only the first 50 coefficients, following prior work (e.g., RingNet [25]), as these capture the dominant modes of expressive variation while improving generalization and reducing identity overfitting.

**Delta pose computation.** To model dynamic facial and head movements, we additionally compute frame-to-frame differences of both expression and pose parameters, i.e.,  $\Delta\mathbf{e}_t = \mathbf{e}_t - \mathbf{e}_{t-1}$  and  $\Delta\mathbf{p}_t = \mathbf{p}_t - \mathbf{p}_{t-1}$ , which emphasize abrupt motion changes such as rapid reorientations, tilts, and expressive transitions. To qualitatively assess parameter fidelity, the extracted coefficients are projected onto a mean FLAME mesh with identity parameters set to zero and rendered alongside the original videos with MediaPipe landmark overlays (Fig. 1). Visual inspection confirms the fidelity of reconstructions and the preservation of stress-relevant geometric cues.

**Gaze Dynamics** In order to complement facial features with oculomotor information, when gaze measurements are available, we derive frame-wise gaze dynamics from the 2D gaze position signals  $(x_t, y_t)$ , sampled at  $\Delta t = 1/30$  s. Horizontal and vertical gaze velocities are computed by discrete differentiation as  $v_t^x = (x_t - x_{t-1})/\Delta t$  and  $v_t^y = (y_t - y_{t-1})/\Delta t$ , with gaze speed defined as  $\|v_t\| = \sqrt{(v_t^x)^2 + (v_t^y)^2}$ . Gaze accelerations are computed analogously as finite differences of the velocity components,  $a_t^x = (v_t^x - v_{t-1}^x)/\Delta t$  and  $a_t^y = (v_t^y - v_{t-1}^y)/\Delta t$ , with magnitude  $\|a_t\| = \sqrt{(a_t^x)^2 + (a_t^y)^2}$ . Short-term temporal statistics are captured with rolling mean and standard deviation of gaze speed over 1 s and 3 s windows, as well as a 2 s gaze dispersion measure given by  $\|(\sigma_{x_t}, \sigma_{y_t})\|_2$ .

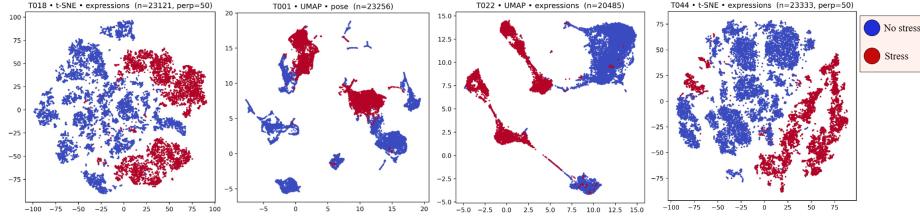


Fig. 2: Subject-wise low-dimensional embeddings of EMOCA features. From left to right: t-SNE embeddings of expression features for subject T018, UMAP embeddings of pose features for subject T001, UMAP embeddings of expression features for subject T022, and t-SNE embeddings of expression features for subject T044. Blue points denote non-stress and red points denote stress samples. Subject-wise embeddings reveal clear stress-related structure.

### 3.3 Statistical Analysis

**Phase-wise MD–ND physiological effects.** Stress-related effects are quantified by comparing the sensorimotor distracted (MD) and normal driving (ND) conditions within five predefined phases. Phases P1–P4 are event-aligned using the cognitive drive (CD): stimulus episodes are detected as contiguous intervals where  $\text{Stimulus} \neq 0$  (defining P2 and P4), while P3 corresponds to a fixed detour segment (4400–5600 m), mapped to time via distance–time interpolation; when distance is unavailable, CD time boundaries are used. Phase P5 is time-based and spans a fixed 120 s recovery window immediately following the second stimulus episode. For each modality  $m \in \{\text{Breathing Rate, Heart Rate, Perinasal Perspiration}\}$ , phase  $p$ , and subject  $s$ , phase means  $\mu_{s,p,m}^{MD}$  and  $\mu_{s,p,m}^{ND}$  are computed by averaging valid samples within the phase window (after quality control and excluding phases with fewer than  $N_{\min}$  samples). Paired MD–ND effects are defined as  $\Delta_{s,p,m} = \mu_{s,p,m}^{MD} - \mu_{s,p,m}^{ND}$  and assessed at the group level via two-sided one-sample  $t$ -tests of  $\{\Delta_{s,p,m}\}_s$  against zero. An identical phase-wise paired-difference protocol is applied to each selected EMOCA expression and pose coefficient  $f$ , yielding  $\Delta_{s,p,f} = \mu_{s,p,f}^{MD} - \mu_{s,p,f}^{ND}$ , with significance assessed per  $(p, f)$  using two-sided one-sample  $t$ -tests.

**Qualitative analysis of feature distributions.** To inspect the structure of the feature space, we apply non-linear dimensionality reduction (t-SNE and UMAP) to EMOCA expression, pose, and gaze dynamics. Subject-wise embeddings reveal coherent separation between stressed and non-stressed samples relative to each individual’s baseline, whereas joint embeddings across subjects form subject-specific clusters without global stress separability (Fig. 2).

**Temporal smoothing and facial dynamics.** For each subject  $s$  and phase  $p$ , EMOCA expression and pose coefficients are treated as temporal signals  $X_{s,p}(t)$ . Each coefficient  $Y(t)$  is summarized by its mean level  $\mu = \frac{1}{T} \sum_t Y(t)$  and its velocity  $\nu = \frac{1}{T} \sum_t |\dot{Y}(t)|$ , where  $Y(t)$  may be unsmoothed or lightly smoothed using symmetric triangular convolution ( $k = 3$ ) or cubic smoothing

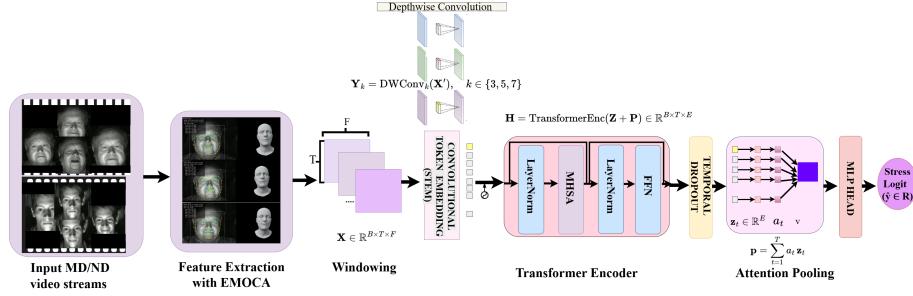


Fig. 3: Overview of the proposed visual stress recognition pipeline, where MD video streams are converted into EMOCA-based facial feature sequences augmented with temporal and MD–ND first-order difference cues, followed by Transformer-based temporal modeling and attention pooling for stress classification.

splines, and  $\dot{Y}(t)$  is obtained via finite differences. For each  $(s, p, f)$  we compute paired MD–ND deltas  $\Delta\mu_{s,p,f} = \mu_{s,p,f}^{MD} - \mu_{s,p,f}^{ND}$  and  $\Delta\nu_{s,p,f} = \nu_{s,p,f}^{MD} - \nu_{s,p,f}^{ND}$ , and test  $\{\Delta\mu_{s,p,f}\}_s$  and  $\{\Delta\nu_{s,p,f}\}_s$  against zero across subjects with one-sample  $t$ -tests (paired-difference formulation).

### 3.4 Transformer temporal modeling

We adopt the pipeline shown in Fig. 3, which combines Transformer-based temporal modeling with attention pooling for stress prediction. Raw timestamps are cleaned to enforce a strictly increasing temporal axis, and signals are segmented into non-overlapping windows of 9 s.

Let  $\mathbf{x}_{s,p,w}^d(t) \in \mathbb{R}^{F_0}$  denote the per-frame EMOCA coefficients for subject  $s$ , phase  $p$ , window  $w$ , and time index  $t$ , extracted from drive  $d \in \{\text{MD, ND}\}$ . Each frame comprises expression coefficients  $\mathbf{e}(t)$ , pose coefficients  $\mathbf{r}(t)$ , and first-order temporal differences  $\Delta\mathbf{x}_{s,p,w}^d(t) = \mathbf{x}_{s,p,w}^d(t) - \mathbf{x}_{s,p,w}^d(t-1)$  for  $t = 2, \dots, T$ . Accordingly, the frame-level MD visual sequence is

$$\mathbf{X}_{s,p,w}^{\text{MD}} = [\mathbf{e}_{s,p,w}^{\text{MD}}(t) \parallel \mathbf{r}_{s,p,w}^{\text{MD}}(t) \parallel \Delta\mathbf{x}_{s,p,w}^{\text{MD}}(t)]_{t=1}^T \in \mathbb{R}^{T \times F}. \quad (1)$$

To explicitly exploit the paired ND drive as a subject-specific baseline, we compute the window-level velocity-difference descriptor  $\Delta\mathbf{v}_{s,p,w} = \frac{1}{T-1} \sum_{t=2}^T (\Delta\mathbf{x}_{s,p,w}^{\text{MD}}(t) - \Delta\mathbf{x}_{s,p,w}^{\text{ND}}(t))$ , which is appended to the model input to emphasize stress-induced deviations relative to baseline dynamics, motivated by the significant MD–ND effects observed in stressor phases.

A multiscale convolutional stem extracts short-term temporal patterns using parallel depthwise 1D convolutions with kernel sizes 3, 5, and 7, whose outputs are concatenated and linearly projected to an embedding of dimension  $E$ :

$$\mathbf{Z}_0 = \text{Conv}_{1 \times 1}([\text{Conv}_3(\mathbf{X}) \parallel \text{Conv}_5(\mathbf{X}) \parallel \text{Conv}_7(\mathbf{X})]). \quad (2)$$

Learned positional embeddings are added, and the resulting sequence is processed by a stack of Transformer encoder layers with multi-head self-attention.

A fixed-length window representation is obtained via attention-based temporal pooling, with weights  $a_t = \text{softmax}(\mathbf{v}^\top \tanh(W\mathbf{z}_t))$  and pooled embedding  $\mathbf{h} = \sum_{t=1}^T a_t \mathbf{z}_t \in \mathbb{R}^E$ , emphasizing stress-relevant temporal segments. The embedding is passed to a lightweight feed-forward head and trained using binary cross-entropy with logits.

For all experiments, subject-wise normalization is applied per cross-validation split: per-feature statistics are estimated exclusively from the training windows of each split and used to normalize the corresponding training, validation, and test data, preventing information leakage across subjects.

**Early fusion.** As a strong multimodal baseline, we implement an *early-fusion* strategy where EMOCA coefficients are concatenated with the secondary modality (either biosignals or gaze-dynamics) at the input level to form a single per-frame feature vector; the same windowing and normalization protocol is applied.

**Cross-modal attention fusion.** To model inter-modal interactions, we employ a cross-modal attention architecture (Fig. 4) with two modality-specific encoders (EMOCA and the paired modality), each comprising a convolutional stem and a Transformer encoder. The latent sequences are fused via bidirectional cross-attention (EMOCA  $\leftarrow$  modality-B and modality-B  $\leftarrow$  EMOCA), followed by attention pooling per stream and concatenation of pooled representations for classification. The same feature construction is used across fusion strategies, including EMOCA (expression, pose, and temporal differences) and the appended window-level MD–ND difference descriptors  $\Delta_1^{(w)}$  for EMOCA and physiological signals; in early fusion, these descriptors are concatenated at the input.

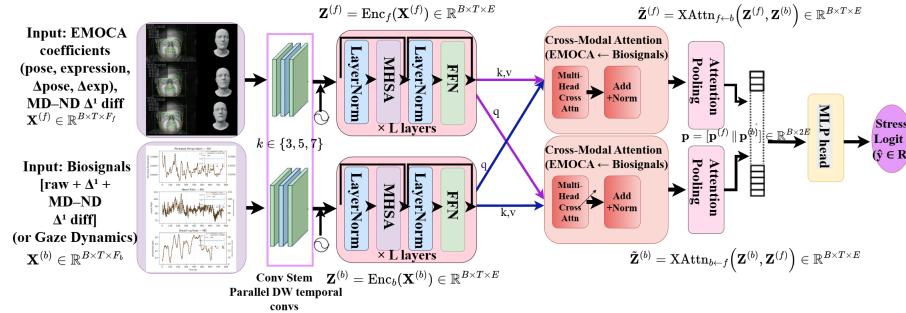


Fig. 4: Cross-modal attention fusion architecture, where EMOCA and biosignal (or gaze-dynamics) streams are independently encoded with convolutional stems and Transformer encoders, fused via bidirectional cross-attention, aggregated using attention pooling, and finally concatenated for stress prediction.

Table 1: Low  $p$ -values are therefore expected primarily during the stressor phases (P2, P4), with non-significant effects during non-stress phases. In addition to the physiological signals, we report representative EMOCA coefficients selected among the most stress-correlated dimensions (identified via PCA loadings), illustrating that both biosignals and 3D facial features exhibit consistent, phase-dependent responses to the sensorimotor stressor.

Signal	P1	P2 (Stressor)	P3	P4 (Stressor)	P5
Breathing Rate	$1.668 \times 10^{-1}$	<b>2.02e-12</b>	$4.036 \times 10^{-1}$	<b>5.36e-15</b>	<b>0.00744</b>
Heart Rate	<b>0.00107</b>	<b>1.75e-10</b>	$2.015 \times 10^{-1}$	<b>3.34e-08</b>	6.529
Perinasal Perspiration	$8.623 \times 10^{-2}$	<b>2.99e-04</b>	$7.907 \times 10^{-1}$	<b>0.005998</b>	5.452
Expression03 (exp_03)	$7.98 \times 10^{-2}$	<b>5.72e-08</b>	$7.9 \times 10^{-1}$	<b>4.66e-07</b>	9.05
Expression18 (exp_18)	$2.056 \times 10^{-1}$	<b>7.85e-13</b>	$1.32 \times 10^{-2}$	<b>1.25e-13</b>	<b>0.00037</b>
Expression20 (exp_20)	<b>0.00087</b>	<b>4.29e-11</b>	$5.105 \times 10^{-1}$	<b>1.00e-13</b>	9.818
Expression40 (exp_40)	$1.05 \times 10^{-3}$	<b>4.48e-12</b>	$7.244 \times 10^{-1}$	<b>3.05e-10</b>	6.74
Pose0 (pose_00)	$4.05 \times 10^{-3}$	<b>1.23e-10</b>	$1.335 \times 10^{-1}$	<b>8.34e-11</b>	2.23

## 4 Experiments

### 4.1 Statistical Analysis: Physiological and Visual Stress Trackers

**Visual stress trackers (EMOCA coefficients).** We applied the MD–ND differencing strategy described in Section 3.2 to the EMOCA facial features. For each subject  $s$ , phase  $P_i$ , and coefficient  $m$ , we computed the paired difference  $\Delta x_{s,m}(P_i) = \bar{x}_{s,m}(\text{MD}, P_i) - \bar{x}_{s,m}(\text{ND}, P_i)$  and assessed group-level effects using two-sided one-sample  $t$ -tests against zero across subjects. Statistical significance is reported using the convention \* ( $p < 0.05$ ), \*\* ( $p < 0.01$ ), and \*\*\* ( $p < 0.001$ ). Figure 5 summarizes the phase-wise effects. Under the strict criterion  $p < 0.001$ , 24/50 expression coefficients show significant modulation in at least one stressor phase (P2 or P4), including 18/50 significant in both P2 and P4, while 2/6 pose coefficients also exhibit stress-related effects. Using the more permissive threshold  $p < 0.05$ , 28/50 expression coefficients are significant in both stressor phases with an additional 3/50 significant in either P2 or P4, and 3/6 pose coefficients show consistent modulation. Perinasal perspiration, heart rate, and breathing rate capture structured physiological stress responses (Table 1).

PCA of the EMOCA expression and pose coefficients revealed `pose_00` and `exp_40` as the most stress-correlated components (Fig. 6); we therefore visualize the top five stress-related components by rendering the FLAME mean face at  $\pm 3\sigma_1$  (Fig. 7).

**1D Convolution and Spline Embeddings.** Figure 8 compares phase-wise MD–ND significance patterns obtained from velocity-based representations under different temporal operators. Mean-level features exhibit sparse and weak effects, whereas velocity features markedly enhance discrimination of the stressor phases (P2 and P4). Without temporal smoothing (finite-difference velocity,  $k=1$ ), all 56 coefficients are significant in both P2 and P4, with 37 showing

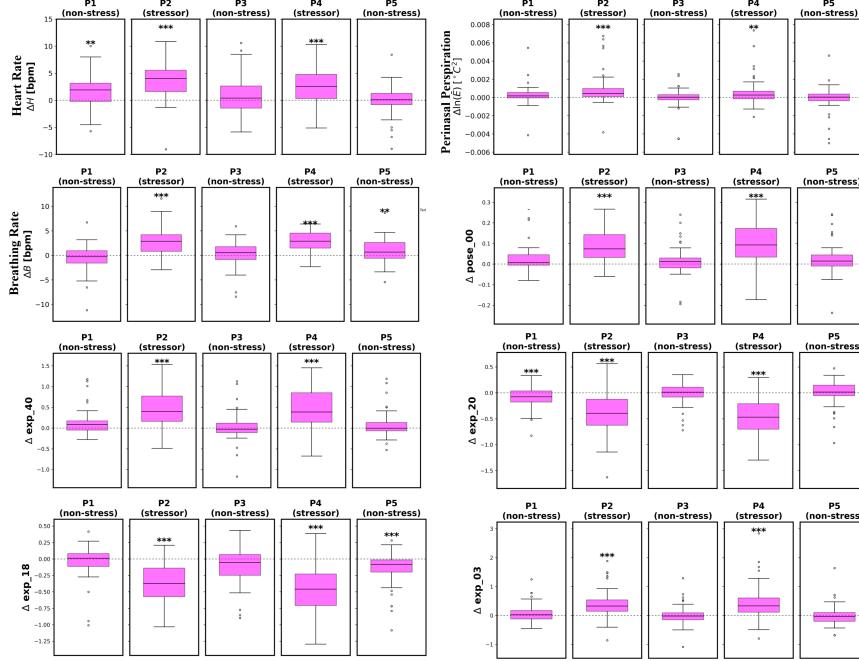


Fig. 5: Phase-wise MD-ND differences highlighting stress-related separability during the stressor phases P2 and P4. Established physiological markers (heart rate, perinasal perspiration) and selected EMOCA facial coefficients ( $\Delta \text{pose}_{00}$ ,  $\Delta \text{exp}_{40}$ ,  $\Delta \text{exp}_{20}$ ,  $\Delta \text{exp}_{18}$ ,  $\Delta \text{exp}_{03}$ ), chosen via PCA for their high correlation with the stress label, exhibit pronounced deviations from baseline, indicating their potential as reliable facial stress trackers.

exclusive stress-phase selectivity and 13 exhibiting significance in exactly one additional non-stress phase. Lightweight triangular convolution ( $k=3$ ) preserves stress discrimination but reduces selectivity, while cubic spline smoothing yields the most conservative patterns, increasing significance in both stressor phases (41 coefficients) at the cost of reduced sensitivity and minimal leakage to non-stress phases. Overall, raw velocity without smoothing provides the most informative representation, indicating that stress is primarily reflected in rapid frame-to-frame facial dynamics. A logistic-regression probe confirms the predictive value of EMOCA velocity features, achieving an AUROC of 83.6% for P2/P4 versus P1/P3/P5 (Table 2).

#### 4.2 LDA stress axis and geometric visualization

To identify which facial components are most predictive of stress, we applied binary Linear Discriminant Analysis (LDA) to the standardized coefficients. Because the problem is two-class, LDA yields a single discriminant direction  $w$  that

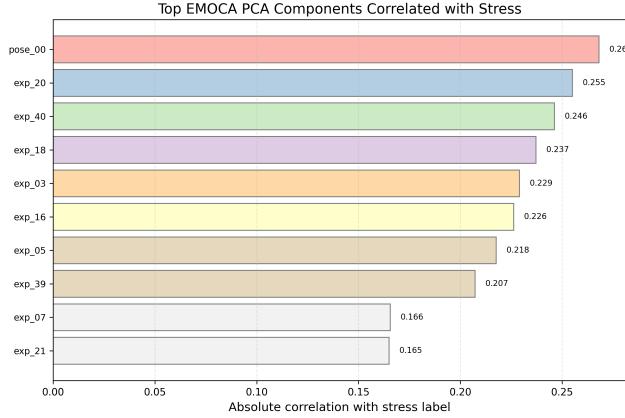


Fig. 6: Top EMOCA PCA components most correlated with stress. The strongest effects are observed for the global pose component `pose_00` and the expression component `exp_20`.

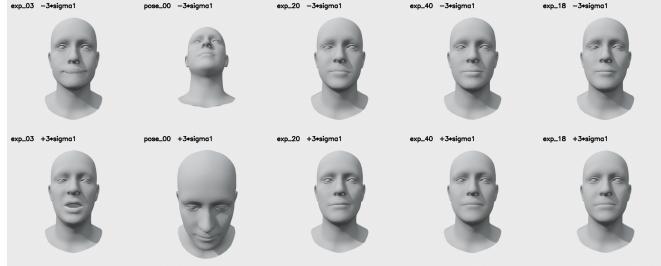


Fig. 7: Top-5 stress-correlated PCA components visualized on the mean face. Components are rendered at  $-3 \cdot \sigma_1$  (top) and  $+3 \cdot \sigma_1$  (bottom), where  $\sigma_1$  is a shared visualization scale corresponding to the largest PCA standard deviation.

maximally separates stress from no-stress frames, producing for each sample a scalar projection  $z_i = w^\top x_i^{\text{std}}$ , which effectively serves as a latent action unit for stress. To visualize the effect of this axis on 3D geometry, we perturbed the mean FLAME face by moving the EMOCA coefficients by  $\pm 3\sigma$  along  $w$  and rendered the resulting shapes. As shown in Fig. 9 the LDA stress axis is primarily associated with lower-face deformations (mouth, jaw) and head pose variation, indicating that these components contribute most strongly to discriminating stress from non-stress states.

### 4.3 Classification Metrics

All reported results were obtained using 5-fold subject-wise cross-validation to ensure no data leakage between sets. Each window was assigned a binary stress

Table 2: Logistic-regression probe for stress-phase classification (P2/P4 vs. P1/P3/P5) using phase-level EMOCA velocity features. Evaluation is performed on MD drives only, under subject-wise cross-validation. Results are reported as mean  $\pm$  standard deviation across folds.

Method	AUROC	AUPRC	Accuracy
Raw velocity	$0.730 \pm 0.056$	$0.637 \pm 0.088$	$0.682 \pm 0.027$
Triangular 1D conv ( $k=5$ )	<b><math>0.836 \pm 0.082</math></b>	<b><math>0.762 \pm 0.110</math></b>	<b><math>0.750 \pm 0.044</math></b>

label based on the proportion of stressed frames it contained: windows with a stress ratio greater than 0.4 were labeled as stress. Performance was computed at the window level, and metrics including AUROC, AUPRC, F1, Accuracy, and Balanced Accuracy were averaged across folds.

**Early-Fusion Extensions.** We optimized temporal and training hyperparameters and obtained our strongest visual-only performance using non-overlapping 9s windows, dropout = 0.2, and early stopping within 20 epochs. This configuration achieved a mean AUROC of  $0.908 \pm 0.015$  and accuracy of  $0.841 \pm 0.017$  under 5-fold subject-wise cross-validation (Table 4). The visual representation includes expression and pose coefficients, their first-order temporal differences, and window-level MD–ND differences computed from these dynamics, enabling the model to capture stress-induced deviations relative to baseline driving.

Early fusion does not outperform the visual-only baseline (AUROC  $0.901 \pm 0.017$  for EMOCA+Bio), indicating that facial dynamics already capture the dominant stress-related information under this setting.

**Cross-Modal Attention Fusion.** We introduce a cross-modal attention architecture that processes visual and non-visual signals with modality-specific temporal encoders, followed by bidirectional cross-attention. This design enables each modality to attend selectively to temporally relevant cues in the other before aggregation via attention pooling and final classification.

Cross-modal fusion consistently outperforms both single-modality and early-fusion baselines when using the full set of visual facial descriptors and all available physiological signals, combined with subject-wise normalization. As shown in Table 4, EMOCA combined with physiological signals via cross-attention achieves the strongest overall performance, with AUROC  $0.92 \pm 0.04$ , F1  $0.866 \pm 0.054$ , accuracy  $0.866 \pm 0.05$ , and balanced accuracy  $0.87 \pm 0.05$ . Cross-modal fusion with gaze yields competitive but slightly lower performance (AUROC  $0.918 \pm 0.043$ ), while combinations excluding facial dynamics perform substantially worse.

**Comparisons.** To ensure a fair comparison with established stress-classification approaches, we evaluated our model against two literature-aligned baselines under the same 5-fold subject-wise protocol. First, we implemented *StressID-style* traditional ML pipelines using non-overlapping windows and per-window mean and standard deviation features, training SVM and kNN classifiers

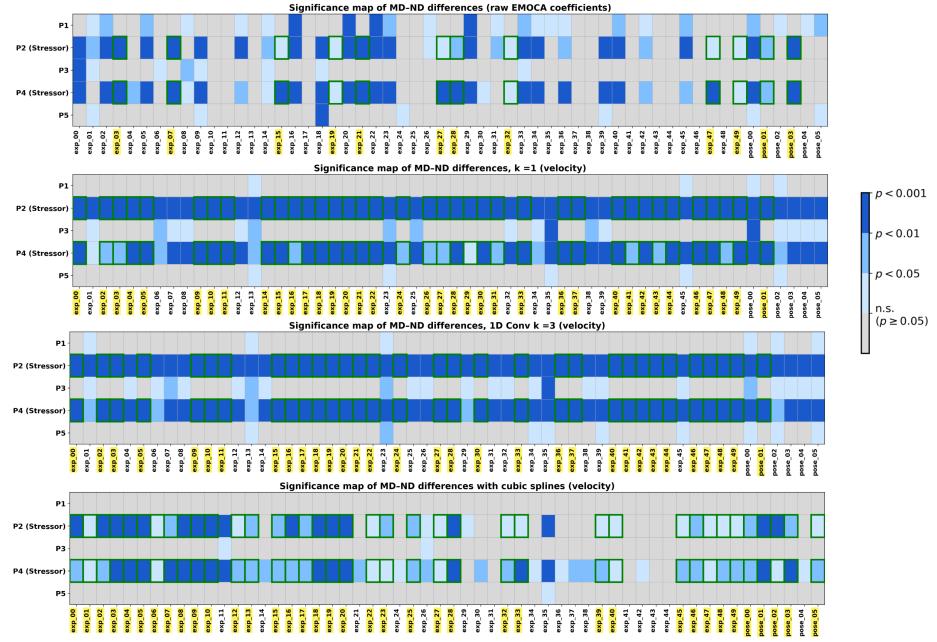


Fig. 8: Phase-wise significance maps. From top to bottom: raw EMOCA mean levels; velocity without temporal smoothing ( $k=1$ ); velocity with lightweight triangular 1D convolution ( $k=3$ ); and velocity after cubic spline smoothing. Velocity-based representations markedly enhance discrimination of the stressor phases (P2, P4), while increased smoothing yields more conservative but selective effects. Color encodes raw  $p$ -values; green outlines and highlighted yellow denote coefficients significant in only P2 and P4.

with subject-wise cross-validation. Second, we evaluated a *Giannakakis-style* [11] fully-connected baseline adapted for binary stress prediction, using identical normalization and splits.

As shown in Table 3, traditional SVM-based baselines are competitive, with EMOCA (SVM) achieving AUROC  $0.893 \pm 0.020$  and early-fusion EMOCA+Bio (SVM) reaching AUROC  $0.893 \pm 0.021$ . Our visual-only Transformer slightly improves upon these results (AUROC  $0.908 \pm 0.015$ ), while maintaining higher accuracy and balanced accuracy. In contrast, fully-connected baselines perform consistently worse under the same subject-wise protocol.

## 5 Conclusions

This work investigates stress estimation under distracted driving by jointly analyzing disentangled 3D facial dynamics, physiological signals, and gaze behavior. Phase-wise analysis revealed that several EMOCA-derived expression

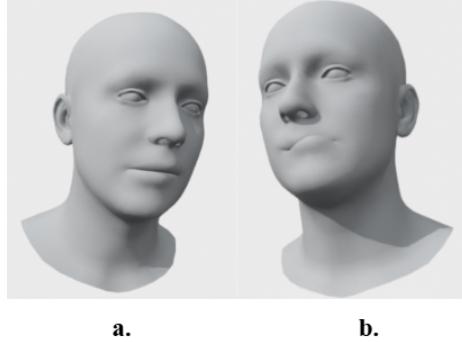


Fig. 9: Visualization of the LDA stress discriminant direction in EMOCA space. The EMOCA coefficients are perturbed by  $-3\sigma$  (a) and  $+3\sigma$  (b) along the LDA axis and rendered on the mean FLAME face.

Table 3: Comparison of stress classification performance between the proposed methodology and prior baseline approaches. All results are reported as mean  $\pm$  standard deviation over 5-fold subject-wise cross-validation.

Modality	AUROC	AUPRC	F1	Accuracy	Balanced Acc.
<i>Ours</i>					
EMOCA	<b>0.908</b> $\pm$ 0.015	0.915 $\pm$ 0.022	<b>0.848</b> $\pm$ 0.025	<b>0.8412</b> $\pm$ 0.0168	<b>0.8407</b> $\pm$ 0.0183
EMOCA + Bio EF	0.901 $\pm$ 0.017	<b>0.918</b> $\pm$ 0.024	0.832 $\pm$ 0.039	0.831 $\pm$ 0.036	0.835 $\pm$ 0.033
<i>StressID reported traditional ML baselines (9s windows; mean+std statistics) [2]</i>					
EMOCA (SVM)	0.893 $\pm$ 0.020	0.903 $\pm$ 0.026	0.842 $\pm$ 0.029	0.829 $\pm$ 0.027	0.826 $\pm$ 0.026
EMOCA (kNN)	0.848 $\pm$ 0.037	0.843 $\pm$ 0.050	0.803 $\pm$ 0.0366	0.790 $\pm$ 0.0372	0.789 $\pm$ 0.0373
EMOCA + Bio EF (SVM)	0.893 $\pm$ 0.021	0.903 $\pm$ 0.026	0.840 $\pm$ 0.027	0.827 $\pm$ 0.024	0.824 $\pm$ 0.022
EMOCA + Bio EF (kNN)	0.856 $\pm$ 0.033	0.849 $\pm$ 0.042	0.812 $\pm$ 0.025	0.797 $\pm$ 0.027	0.795 $\pm$ 0.027
<i>Fully-connected baselines (Giannakakis et al. MLP [11])</i>					
EMOCA	0.871 $\pm$ 0.023	0.889 $\pm$ 0.032	0.822 $\pm$ 0.031	0.810 $\pm$ 0.028	0.809 $\pm$ 0.026
EMOCA + Bio EF	0.8757 $\pm$ 0.0206	0.8921 $\pm$ 0.0276	0.8148 $\pm$ 0.0352	0.8051 $\pm$ 0.0280	0.8046 $\pm$ 0.0252

and pose coefficients show consistent, phase-selective stress modulations comparable to physiological markers. Stress was more strongly encoded in temporal dynamics—especially velocity-based descriptors—than in static features, with convolutional smoothing outperforming spline alternatives. We introduced a Transformer-based temporal modeling framework and evaluated unimodal, early-fusion, and cross-modal attention strategies. Cross-Modal Attention Fusion of EMOCA and physiological signals achieved the best results (AUROC  $0.92 \pm 0.04$ , Accuracy  $0.866 \pm 0.05$ ), with gaze fusion also competitive (AUROC  $0.918 \pm 0.04$ ). Benchmarks against literature-aligned baselines, including StressID-style SVM/kNN and MLP models, confirmed that none matched the proposed approach, underscoring the importance of temporal modeling and explicit inter-modal interaction.

Table 4: Stress classification performance using single-modality, early-fusion, and cross-modal attention inputs. Evaluation is performed with 5-fold subject-wise cross-validation. Results are reported as mean  $\pm$  standard deviation.

Modality	AUROC	AUPRC	F1	Accuracy	Balanced Acc.
EMOCA	0.908 $\pm$ 0.015	0.915 $\pm$ 0.022	0.848 $\pm$ 0.025	0.8412 $\pm$ 0.0168	0.8407 $\pm$ 0.0183
Bio (PP, HR, BR)	0.527 $\pm$ 0.054	0.568 $\pm$ 0.023	0.439 $\pm$ 0.324	0.510 $\pm$ 0.041	0.498 $\pm$ 0.003
Early Fusion(EMOCA + Bio)	0.901 $\pm$ 0.017	0.918 $\pm$ 0.024	0.832 $\pm$ 0.039	0.831 $\pm$ 0.036	0.835 $\pm$ 0.033
Early Fusion(EMOCA + Gaze)	0.863 $\pm$ 0.047	0.866 $\pm$ 0.074	0.780 $\pm$ 0.067	0.780 $\pm$ 0.052	0.785 $\pm$ 0.048
Cross-Modal (EMOCA + Bio)	<b>0.92 <math>\pm</math> 0.04</b>	0.91 $\pm$ 0.05	<b>0.866 <math>\pm</math> 0.054</b>	<b>0.866 <math>\pm</math> 0.05</b>	<b>0.87 <math>\pm</math> 0.05</b>
Cross-Modal (EMOCA + Gaze)	0.918 $\pm$ 0.043	<b>0.92 <math>\pm</math> 0.04</b>	0.85 $\pm$ 0.05	0.855 $\pm$ 0.047	0.857 $\pm$ 0.046
Cross-Modal (Gaze + Bio)	0.65 $\pm$ 0.039	0.64 $\pm$ 0.06	0.55 $\pm$ 0.08	0.591 $\pm$ 0.03	0.59 $\pm$ 0.03

## References

1. Almeida, J., Rodrigues, F.: Facial expression recognition system for stress detection with deep learning. pp. 256–263 (01 2021)
2. Chaptoukaev, H., Strizhikova, V., Panariello, M., Dalpaos, B., Reka, A., Manera, V., Thümmler, S., Ismailova, E., W., N., bremond, f., Todisco, M., Zuluaga, M.A., M. Ferrari, L.: Stressid: a multimodal dataset for stress identification. In: NeurIPS. vol. 36, pp. 29798–29811 (2023)
3. Danecik, R., Black, M.J., Bolkart, T.: EMOCA: Emotion driven monocular face capture and animation. In: CVPR. pp. 20311–20322 (2022)
4. Ding, D., Xu, W., Liu, X., Zhu, T.: Facial video based stress detection for enhancing ecological validity. Acta Psychologica **255** (2025)
5. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. SIGGRAPH **40**(8) (2021)
6. Filntisis, P.P., Retsinas, G., Paraperas-Papantonio, F., Katsamanis, A., Roussos, A., Maragos, P.: Visual speech-aware perceptual 3d facial expression reconstruction from videos. arXiv:2207.11094 (2022)
7. Gavrilescu, M., Vizireanu, N.: Predicting depression, anxiety, and stress levels from videos using the facial action coding system. Sensors **19**(17) (2019)
8. Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P., Marias, K., Tsiknakis, M.: Stress and anxiety detection using facial cues from videos. Biomedical Signal Processing and Control **31**, 89–101 (2017)
9. Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., Tsiknakis, M.: Review on psychological stress detection using biosignals. IEEE Trans. on Affective Computing **13**, 440–460 (2019)
10. Giannakakis, G., Koujan, M.R., Roussos, A., Marias, K.: Automatic stress detection evaluating models of facial action units. In: FG. p. 728–733 (2020)
11. Giannakakis, G., Koujan, M.R., Roussos, A., Marias, K.: Automatic stress analysis from facial videos based on deep facial action units recognition. Pattern Anal. Appl. p. 521–535 (Aug 2022)
12. Giannakakis, G.A., Manousos, D., Chaniotakis, V., Tsiknakis, M.: Evaluation of head pose features for stress detection and classification. BHI pp. 406–409 (2018)
13. Hazer-Rau, D., Zhang, L., Traue, H.C.: A workflow for affective computing and stress recognition from biosignals. Engineering Proceedings **2**(1) (2020)
14. Jaiswal, M., Bara, C.P., Luo, Y., Burzo, M., Mihalcea, R., Provost, E.M.: Muse: a multimodal dataset of stressed emotion. In: Int'l Conf. on Language Resources and Evaluation (2020)

15. Jaiswal, S., Valstar, M.: Deep learning the dynamic appearance and shape of facial action units. In: WACV. pp. 1–8 (2016)
16. Jeon, T., Bae, H.B., Lee, Y., Jang, S., Lee, S.: Deep-learning-based stress recognition with spatial-temporal facial information. Sensors **21**(22) (2021)
17. Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M.A., Kraaij, W.: The swell knowledge work dataset for stress and user modeling research. In: ICMI (2014)
18. Koujan, M.R., Alharbawee, L., Giannakakis, G., Pugeault, N., Roussos, A.: Real-time facial expression recognition “in the wild” by disentangling 3d expression from identity. In: FG. p. 24–31 (2020)
19. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. SIGGRAPH **36**(6), 194:1–194:17 (2017)
20. Liu, W., Gong, Y., Zhang, G., Lu, J., Zhou, Y., Liao, J.: Glmdrivenet: Global-local multimodal fusion driving behavior classification network. Eng. Appl. AI (2024)
21. Liu, W., Lu, J., Liao, J., Qiao, Y., Zhang, G., Zhu, J., Xu, B., Li, Z.: Fmdnet: Feature-attention-embedding-based multimodal-fusion driving-behavior-classification network. IEEE Trans. on Comp. Social Systems **11**(5) (2024)
22. Markova, V., Ganchev, T., Kalinkov, K.: Clas: A database for cognitive load, affect and stress recognition (01 2020)
23. Mou, L., Chang, J., Zhou, C., Zhao, Y., Ma, N., Yin, B., Jain, R., Gao, W.: Multimodal driver distraction detection using dual-channel network of cnn and transformer. Expert Systems with Applications **234**, 121066 (2023)
24. Sabour, R.M., Benezeth, Y., De Oliveira, P., Chappé, J., Yang, F.: Ubfc-phys: A multimodal database for psychophysiological studies of social stress. IEEE Trans. on Affective Computing **14**(1), 622–636 (2023)
25. Sanyal, S., Bolkart, T., Feng, H., Black, M.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: CVPR (Jun 2019)
26. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: ICMI. p. 400–408 (2018)
27. Siam, A.I., Gamel, S.A., Talaat, F.M.: Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques. Neural Computing and Applications **35**, 12891–12904 (2023)
28. Sinhal, A., Sinhal, A., Sinhal, A.: Stress monitoring in healthcare: An ensemble machine learning framework using wearable sensor data (2025)
29. Steeneken, H.J.M., Hansen, J.H.L.: Speech under stress conditions: overview of the effect on speech production and on system performance. ICASSP **4** (1999)
30. Taamneh, S., Tsiamyrtzis, P., Deosta, M., Buddharaju, P., Khatri, A., Manser, M., Ferris, T., Wunderlich, R., Pavlidis, I.: A multimodal dataset for various forms of distracted driving. Scientific Data **4**, 170110 (08 2017)
31. Tran, T.D., Kim, J., Ho, N.H., Yang, H.J., Pant, S., Kim, S.H., Lee, G.S.: Stress analysis with dimensions of valence and arousal in the wild. Applied Sciences **11**(11) (2021)
32. Viegas, C., Lau, S.H., Maxion, R., Hauptmann, A.: Towards independent stress detection: A dependent model using facial action units. In: CBMI. pp. 1–6 (2018)
33. Wang, X., Zhang, T., Chen, C.: Pau-net: Privileged action unit network for facial expression recognition. IEEE Trans. on Cognitive and Developmental Systems **PP**, 1–1 (01 2022)
34. Xu, J., Song, C., Yue, Z., Ding, S.: Facial video-based non-contact stress recognition utilizing multi-task learning with peak attention. IEEE Journal of Biomedical and Health Informatics **28**(9), 5335–5346 (2024)

35. Zhang, H., Feng, L., Li, N., Jin, Z., Cao, L.: Video-based stress detection through deep learning. *Sensors* **20**(19) (2020)