ElderMTL: Multi-Task Affect Monitoring for Elderly Care

Maria Razzhivina[1] , **Shahane Tigranyan**[2,3] , **Aram Avetisyan**[1] ,
Ilya Makarov[1,4] and **Andrey Savchenko**[1,5]
[1]Trusted AI Research Center, RAS
[2]Russian-Armenian University
[3]IIAP NAS RA
[4]AXXX
[5]Sber AI Lab

Abstract

We present ElderMTL, a multi-task affect monitoring system designed for elderly care settings. The system simultaneously estimates Facial Action Units (FAUs), Valence-Arousal (VA) signals, and categorical emotions (FER) from video, capturing multiple layers of affective information. To improve sensitivity to subtle affective cues common in older adults, our approach incorporates age-conditioned physiological modeling, including baseline muscle adjustments and a dynamic AU co-activation graph. This enables the system to adapt to age-related changes in facial expression patterns, providing more reliable and interpretable emotion assessments. In a live demonstration, we showcase ElderMTL processing video streams, visualizing AU activations, affective state predictions, and interpretable insights that highlight age-specific affective dynamics. This work demonstrates that physiologically grounded, multi-task affective monitoring can provide meaningful, real-world support for elderly care.

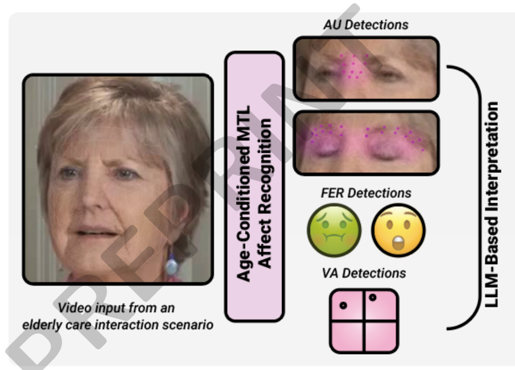

Figure 1: Overview of the ElderMTL demo pipeline.

1 Introduction

The global population is rapidly aging, increasing care demands while intensifying caregiver shortages and pressure on healthcare systems, particularly in geriatric facilities where continuous monitoring and emotional support are essential. Current monitoring approaches—such as periodic assessments or wearable sensors—are often intrusive, labor-intensive, and unsuitable for long-term, continuous observation of older adults.

Research in Human–Robot Interaction (HRI) and Human–Computer Interaction (HCI) has consistently emphasized the critical role of affective perception in facilitating effective human–machine interaction. In this context, affective computing aims to give computer systems the ability to automatically recognize, understand, and respond properly to human emotional cues [Günes and Churamani, 2023]. In broader HRI settings, real-time facial emotion recognition systems have been used to improve interaction by helping robots recognize common facial expressions and adjust their behavior accordingly. [Epifanic, 2023]. Furthermore, research has explored on-the-fly detection of engagement in spontaneous human–robot interactions, highlighting the relevance of temporal facial and behavioral dynamics during interaction [Ben Youssef *et al.*, 2020].

Despite these advances, most affect recognition methods are evaluated in controlled settings, trained on general-population data, or lack the latency and robustness required for real-time deployment in care environments. Facial expression recognition in real time remains challenging due to uncontrolled lighting, pose variations, and demographic biases in training data [Rawal and Stock-Homburg, 2022]. Parallel efforts in HRI, including multimodal affective datasets such as AFFECT-HRI, emphasize the need to better understand affective responses to robot behavior, but have not focused on continuous, video-only affect monitoring in elderly populations [Heinisch *et al.*, 2024]. Moreover, companion robot research underscores the importance of addressing the emotional needs of older adults, revealing the potential of socially assistive platforms to support well-being and engagement [Zeng *et al.*, 2025].

Motivated by this gap, we present **ElderMTL** (Elder Multitask Learning), a real-time, video-only affect recognition system tailored for geriatric care. ElderMTL is designed to monitor emotional states, engagement, and distress indicators during live interaction using standard RGB cameras, enabling scalable and accessible affective monitoring in real-world care environments. The overview of the proposed sys-

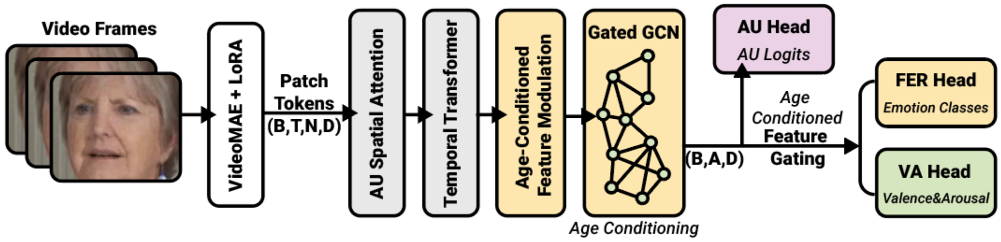

Figure 2: Age-aware multi-task architecture. Video features from a LoRA-adapted VideoMAE are processed by AU-conditioned spatial attention, temporal modeling, and an age-gated GCN to produce embeddings for AU, valence–arousal, and expression prediction.

tem is presented in Figure 1.

1.1 FAU for Affective Monitoring

Facial affect recognition in elderly populations poses unique challenges not addressed by models trained on general adult datasets. Aging alters facial morphology and expression dynamics (e.g., wrinkles, reduced muscle tone, slower activations), causing appearance-based models to misinterpret age-related features as emotional cues and introduce systematic bias [Morey, 2025; Fan *et al.*, 2022; Park *et al.*, 2022].

Most existing approaches map facial appearance directly to emotion labels or affect dimensions. Although effective in controlled settings, such black-box models lack robustness and interpretability in real-world geriatric care environments characterized by lighting variation, pose changes, occlusions, and subtle expressions [Günes and Churamani, 2023].

Facial Action Units (FAUs), defined by the Facial Action Coding System (FACS), provide a physiologically grounded representation of facial muscle activity [Ekman and Friesen, 1978]. Because FAUs encode low-level muscle activations independently of emotion categories, they offer a more robust, interpretable, and generalizable description of facial behavior, supporting recognition of emotions, engagement, and distress-related states [Kollias and Zafeiriou, 2021; Ben Youssef *et al.*, 2020; Lucey *et al.*, 2011; Fan *et al.*, 2022].

Accordingly, we adopt an FAU-centered modeling framework for live human–robot and human–computer interaction, leveraging FAU dynamics to enable interpretable and reliable affect monitoring tailored to elderly users in real-world care settings.

2 Methodology

We propose **ElderMTL**, an age-aware multi-task framework for affect recognition in elderly care. As illustrated in Figure 2, the architecture comprises a transformer-based visual encoder, AU-specific spatial feature extraction, temporal aggregation, age-aware modulation, structured AU graph reasoning, and task-specific prediction heads. The model is trained using masked multi-task losses to jointly predict Facial Action Units (AUs), categorical facial expressions (FER), and Valence–Arousal (VA) from video sequences, while explicitly accounting for age-related facial variations.

Given an input frame x_t, a shared representation is extracted using Video-MAE expanded with lightweight adapter layers to mitigate domain shifts across datasets.

To obtain AU-specific spatial features, we introduce learnable AU queries $q_a \in R^D$ that attend to spatial tokens. For AU a at time t, cross-attention produces $h_{t,a} = \text{Attention}(q_a, z_t)$, enabling focus on discriminative facial regions. Temporal dynamics are modeled per AU by processing $h_{1,a}, \ldots, h_{T,a}$ with a temporal transformer, yielding an aggregated embedding H_a that captures subtle elderly-specific motion patterns. To account for baseline facial changes due to aging, age adaptation is introduced via a small learnable network $f_{\text{age}}(\cdot)$, $\tilde{H}_a = H_a + f_{\text{age}}(\text{age})$.

Structured dependencies among AUs are modeled using a symmetric learnable adjacency matrix A. Age modulation is incorporated by defining the effective adjacency as $A_{ij}^{\text{mod}} = \sigma(A_{ij}) \cdot g_i(\text{age}) \cdot g_j(\text{age})$, where σ is a sigmoid function and $g_i(\cdot)$ are age-conditioned scaling functions. Graph-enhanced embeddings are then computed as $H_a^{\text{graph}} = A^{\text{mod}} H_a W$, where W is a learnable projection matrix. An L_1 regularization term $\|A\|_1$ is applied to encourage sparsity and interpretable AU relationships.

The shared AU representations are fed into three task-specific heads. The AU head produces per-AU logits and is trained only when AU annotations are available. The FER head predicts categorical emotions using cross-entropy loss. The VA head outputs continuous valence and arousal values and is optimized using the Concordance Correlation Coefficient loss, defined as

$$\mathcal{L}_{\text{CCC}} = 1 - \frac{2\,\text{cov}(y, \hat{y})}{\text{var}(y) + \text{var}(\hat{y}) + (\bar{y} - \bar{\hat{y}})^2} \tag{1}$$

During training on multiple datasets with incomplete annotations, each loss is applied only when the corresponding labels are present. Masked losses and a multi-dataset curriculum are used to maintain balanced optimization. The total objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{AU}}\mathcal{L}_{\text{AU}} + \lambda_{\text{FER}}\mathcal{L}_{\text{FER}} + \lambda_{\text{VA}}\mathcal{L}_{\text{CCC}} + \lambda_{\text{sparse}}\|A\|_1 \tag{2}$$

where the coefficients λ control the contribution of each component.

3 Experimental Setup

We evaluate ElderMTL in a unified multi-task setting (FER, AU detection, and VA regression) using heterogeneous datasets with partial supervision and automatically estimated age metadata.

ElderReact [Ma *et al.*, 2019] contains elderly facial expressions with categorical emotion and valence annotations, supporting age-relevant affect modeling.

Method	FER (F1 ↑) ElderReact	FER (F1 ↑) AffWild2	FAU (F1 ↑) DISFA	FAU (F1 ↑) AffWild2	VA (CCC ↑) AffWild2
XGBoost [Ma *et al.*, 2019]	0.43	–	–	–	–
Single-task-CNN [Deng *et al.*, 2020]	–	0.42	–	**0.58**	**0.323**
EmotiEffNet(FS) [Savchenko, 2025]	–	0.43	–	–	–
BigSmall [Narayanswamy *et al.*, 2024]	–	–	0.42	–	–
ElderMTL (Ours)	**0.74**	**0.55**	**0.46**	0.51	0.217

Table 1: Performance comparison on facial expression recognition (FER), facial action unit (FAU) detection, and valence–arousal (VA) estimation. ElderMTL is optimized for real-time, interpretable affect recognition in geriatric HRI settings.

Model / Age Group	FER (F1 ↑)	FAU (F1 ↑)	Val (CCC ↑)	Aro (CCC↑)	Δ_{Y-M}	Δ_{Y-E}
Full Model (ElderMTL)						
Young	0.58	0.54	0.24	0.23	-	-
Middle	0.55	0.51	0.22	0.21	0.03	-
Elderly	0.52	0.48	0.19	0.18	-	0.06
Ablation: w/o Graph Module						
Young	0.56	0.52	0.22	0.21	-	-
Middle	0.52	0.48	0.20	0.19	0.04	-
Elderly	0.48	0.44	0.17	0.16	-	0.08

Table 2: Cross-age evaluation on AffWild2. ElderMTL reduces performance degradation with age compared to the ablated model, indicating effective age-aware AU structure modeling.

AffWild2 [Kollias and Zafeiriou, 2020] provides large-scale in-the-wild annotations for FER, VA, and multiple AUs, enabling robust multi-task training under unconstrained conditions.

DISFA [Mavadati *et al.*, 2013] offers high-quality frame-level AU annotations in controlled settings, strengthening structural AU supervision.

Since demographic annotations are incomplete or inconsistent across datasets, we estimate age for all videos using MiVOLO [Gonçalves *et al.*, 2024], a state-of-the-art vision-language age estimator. Age is normalized to $[0, 1]$ and injected into (1) AU-level muscle bias modeling, (2) AU graph modulation, and (3) global affect feature conditioning.

All AU annotations are aligned into a shared 12-AU representation; missing AUs are handled via binary masking during loss computation. Datasets lacking FER or VA labels contribute only to the corresponding supervised objectives, enabling partial-label multi-task learning within a shared backbone.

Video clips are uniformly sampled to 16 frames and resized to 224×224. We use AdamW with weight decay and a learning rate equal to 3e-5.

To assess age robustness, we partition samples into Young (< 40), Middle ($40 - 59$), and Elderly (≥ 60) groups and report Macro-F1 for FER, mean AU-F1, CCC for valence and arousal, and absolute performance gaps between age groups.

3.1 Demonstration

In our live demonstration, ElderMTL processes video streams and visualizes predictions for Action Units, Valence-Arousal, and Emotion. To illustrate interpretability and applicability, we integrate a frozen local instruction-tuned language model (Llama 3.1 8B Instruct [Huang *et al.*, 2024]). Importantly, it is not trained or fine-tuned on our data, and it is explicitly instructed to reason strictly from detected signals. Its purpose is to show how structured affect outputs can be translated into natural language insights for caregiver support. A recording of our demonstration is available at the following link: https://youtu.be/ZIZS9w760BA

4 Results

We evaluate ElderMTL in a unified multi-task setting (FER, AU detection, VA regression) on ElderReact, AffWild2, and DISFA. As shown in Table 1, ElderMTL achieves the strongest FER and AU detection performance across datasets, reaching 0.74 FER F1 on ElderReact and 0.55 on AffWild2, while obtaining 0.46/0.51 AU F1 (DISFA/AffWild2) and 0.217 VA CCC. These results highlight the benefit of joint multi-task learning with age-aware conditioning.

Cross-age evaluation (Table 2) shows controlled degradation with age ($\Delta_{Y-M} = 0.03$, $\Delta_{Y-E} = 0.06$ for FER). Removing the age-aware graph module consistently reduces performance, particularly for elderly subjects, confirming the importance of structured, age-modulated AU modeling.

Overall, the results demonstrate that incorporating age-conditioned modulation and structured AU reasoning improves affect recognition accuracy and mitigates age-related performance bias.

5 Conclusion

We present ElderMTL, an age-aware multi-task framework for affect recognition in elderly care that jointly models Facial Action Units, Valence–Arousal, and categorical emotions with age-conditioned modulation and structured AU dependencies. The approach addresses age-specific facial dynamics and achieves competitive, age-robust performance across tasks. A live demonstration illustrates its applicability for affect monitoring in assistive elderly care systems.

Ethical Statement

ElderMTL is intended to support, not replace, human caregivers, and its predictions should be treated as decision-support signals warranting human verification rather than clinical assessments. Since the system continuously monitors a vulnerable population via video, deployment requires informed consent and appropriate safeguards for sensitive data.

Acknowledgments

The work was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation (agreement dated June 20, 2025 No. 139-15-2025-011, identifier 000000C313925P4G0002).

References

[Ben Youssef *et al.*, 2020] Atef Ben Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. On-the-fly detection of user engagement decrease in spontaneous human-robot interaction. *International Journal of Social Robotics*, 2020. arXiv:2004.09596.

[Deng *et al.*, 2020] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020.

[Ekman and Friesen, 1978] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[Epifanic, 2023] Jelena Epifanic. Real-time facial emotion recognition for human-robot interaction. *Proceedings of the MEi:CogSci Conference*, 2023. Available online.

[Fan *et al.*, 2022] Yiming Fan, Hewei Wang, Xiaoyu Zhu, Xiangming Cao, Chuanjian Yi, Yao Chen, Jie Jia, and Xiaofeng Lu. Fer-pcvt: facial expression recognition with patch-convolutional vision transformer for stroke patients. *Brain Sciences*, 12(12):1626, 2022.

[Gonçalves *et al.*, 2024] Daniel Gonçalves, Rafael Lopes, and Jaime S. Cardoso. Mivolo: Multi-input vision-language model for age and gender estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[Günes and Churamani, 2023] Hatice Günes and Nikhil Churamani. Affective computing for human-robot interaction research: Four critical lessons for the hitchhiker. 2023. arXiv:2303.18176.

[Heinisch *et al.*, 2024] Judith S Heinisch, Jérôme Kirchhoff, Philip Busch, Janine Wendt, Oskar von Stryk, and Klaus David. Physiological data for affective computing in hri with anthropomorphic service robots: the affect-hri data set. *Scientific Data*, 11(1):333, 2024.

[Huang *et al.*, 2024] Kunal Chawla Huang, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, et al. The llama 3 herd of models. *preprint*, 2024.

[Kollias and Zafeiriou, 2020] Dimitrios Kollias and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild2 database and challenge, deep architectures, and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1470–1479, 2020.

[Kollias and Zafeiriou, 2021] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[Lucey *et al.*, 2011] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 57–64. IEEE, 2011.

[Ma *et al.*, 2019] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M. Girard, and Louis-Philippe Morency. Elderreact: A multimodal dataset for recognizing emotional response in aging adults. In *Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI)*, pages 349–357, New York, NY, USA, 2019. Association for Computing Machinery.

[Mavadati *et al.*, 2013] Sadegh Mavadati, Mohammad H. Mahoor, Marian S. Bartlett, and Phuc Trinh. Disfa: A spontaneous facial action intensity database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 151–158, 2013.

[Morey, 2025] Francesc Xavier Gayà Morey. *Advances in Explainable AI for Affective Computing and Activity Recognition in Elderly*. PhD thesis, Universitat de les Illes Balears, 2025.

[Narayanswamy *et al.*, 2024] Girish Narayanswamy, Yujia Liu, Yuzhe Yang, Chengqian Ma, Xin Liu, Daniel McDuff, and Shwetak Patel. Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7914–7924, 2024.

[Park *et al.*, 2022] Hyungjoo Park, Youngha Shin, Kyu Song, Channyeong Yun, and Dongyoung Jang. Facial emotion recognition analysis based on age-biased data. *Applied Sciences*, 12(16):7992, 2022.

[Rawal and Stock-Homburg, 2022] N. Rawal and R.M. Stock-Homburg. Facial emotion expressions in human–robot interaction: A survey. *International Journal of Social Robotics*, 2022.

[Savchenko, 2025] Andrey V Savchenko. Hsemotion team at abaw-8 competition: Audiovisual ambivalence/hesitancy, emotional mimicry intensity and facial expression recognition. *arXiv preprint arXiv:2503.10399*, 2025.

[Zeng *et al.*, 2025] Hui Zeng, Yuxin Sheng, and Jinwei Zhu. Companion robots supporting the emotional needs of the elderly: Research trends and future directions. *Information*, 2025.