Boundary-Aware Multi-Granularity Learning for Depression Severity Estimation

Zhihong Li[1], Benqiang Yang[1], Yaowei Wang[2], and Yun Yang[1](✉)

[1] National Pilot School of Software, Yunnan University, Kunming, China
zhihongli@stu.ynu.edu.cn, ybq614@126.com, yangyan19@hotmail.com
[2] School of Information Science & Engineering, Yunnan University, Kunming, China
wangyaowei@itc.ynu.edu.cn

Abstract. Depression severity estimation is formulated as continuous regression, yet clinical decisions rely on categorical severity grades. This granularity mismatch leads to boundary-agnostic optimization and ordinally unstructured representations. We propose Multi-Granularity Supervision, jointly leveraging continuous scores and ordinal grades as complementary signals. Our key theoretical contribution proves that ordinal supervision induces boundary-aware gradient amplification, where gradient magnitude peaks at clinical thresholds. We instantiate this through Ordinal Consistency Regularization for boundary awareness in prediction space, and Prototype-based Severity Contrastive Learning for ordinal structure in feature space. On AVEC 2013, AVEC 2014, DAIC-WOZ, and E-DAIC, we achieve MAE of 5.06, 5.15, 4.33, and 5.02 respectively, establishing new state-of-the-art across all benchmarks. Code is available at `https://github.com/aqlzh/BAMGL_Depression`.

Keywords: Depression Severity Estimation · Multi-Granularity Learning · Ordinal Regularization

1 Introduction

Depression severity estimation from behavioral signals has significant clinical value [1]. State-of-the-art methods formulate this as continuous regression optimizing point-wise losses like MSE [2]. However, this overlooks a fundamental characteristic of clinical assessment: the dual granularity structure.

1.1 The Granularity Mismatch Problem

Clinical depression assessment operates at dual granularities: fine-grained continuous scores such as those from the Beck Depression Inventory-II for precise symptom tracking, and coarse-grained severity grades ranging from minimal to severe for guiding treatment decisions. Current methods optimize exclusively at the fine-grained level, receiving no signal about clinical boundaries.

This manifests as two limitations shown in Fig. 1. **Limitation 1: Boundary-Agnostic Optimization.** MSE assigns identical gradients to errors of equal

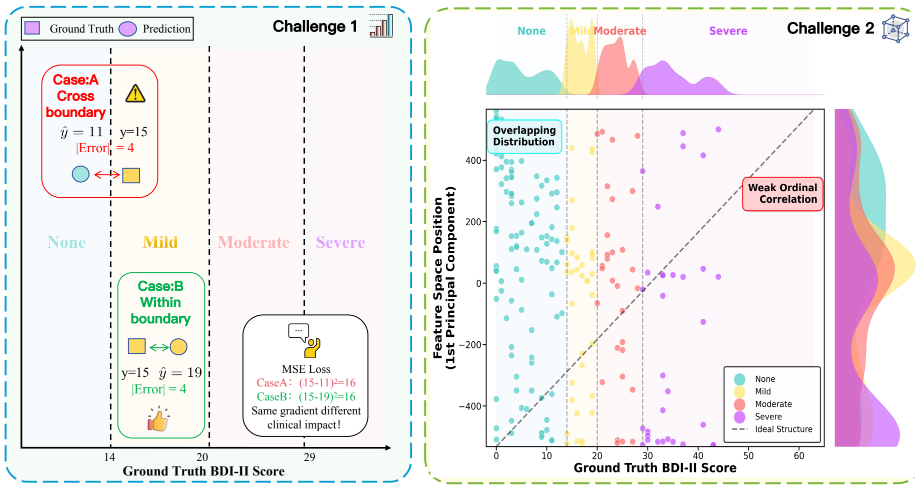

Fig. 1. Granularity mismatch problem. (a) Boundary-agnostic optimization: MSE assigns identical gradients despite different clinical impact. (b) Ordinally unstructured representations.

magnitude regardless of boundary crossing. For example, predicting 15 as 19 within the mild range receives the same MSE penalty as predicting 15 as 11 across the minimal–mild boundary, although the latter has greater clinical consequence. **Limitation 2: Ordinally Unstructured Representations.** Point-wise regression imposes no feature space constraints, allowing samples of similar severity to scatter arbitrarily without ordinal structure.

1.2 Our Approach: Multi-Granularity Supervision

We propose Multi-Granularity Supervision, bridging the granularity gap by treating coarse-grained severity structure as auxiliary supervision. Clinical severity grades encode boundary information invisible to point-wise losses. Injecting this ordinal knowledge as regularization, rather than task reformulation, yields representations respecting both granularities.

Our theoretical contribution proves ordinal supervision induces boundary-aware gradient amplification: gradient contribution from auxiliary loss is maximized at clinical thresholds, providing principled behavior where the model is most careful where clinically most consequential. Unlike CORAL [3] and SORD [4] that reformulate regression as ordinal classification, our method preserves continuous regression while augmenting with ordinal regularization. We instantiate this through Ordinal Consistency Regularization (OCR) for boundary-aware prediction and Prototype-based Severity Contrastive Learning (PSCL) for ordinal feature structure.

Contributions. (1) We identify the granularity mismatch problem with two concrete limitations. (2) We propose Multi-Granularity Supervision bridg-

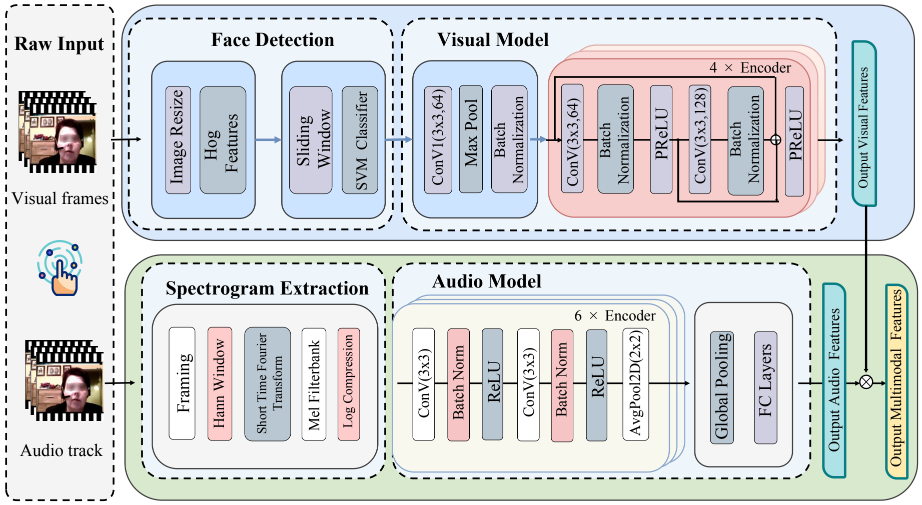

Fig. 2. Overview of the proposed multi-granularity supervision framework, where visual and audio features are extracted and integrated into multimodal representations for depression severity estimation under both fine-grained regression and coarse-grained ordinal supervision.

ing fine-grained regression and coarse-grained ordinal structure. (3) We analyze boundary-aware gradient amplification, characterizing how ordinal regularization concentrates learning at clinical thresholds. (4) We achieve state-of-the-art on four benchmarks: AVEC 2013, AVEC 2014, DAIC-WOZ, and E-DAIC.

2 Related Work

Depression Detection. Recent baselines, including PointTransform [2], AVA-DepNet [13], DepMGNN [19], FC-AEN [21], and SIMMA [23], mainly perform point-wise regression without explicit ordinal supervision.

Ordinal Regression. CORAL [3], SORD [4], and recent contrastive ordinal methods [5,7] reformulate severity estimation as discrete ordinal classification, weakening fine-grained continuous scoring; OCR instead preserves continuous regression while regularizing clinical-threshold learning.

Prototype Learning. Unlike hard-assignment prototypes [6] for classification, PSCL employs *learnable* severity prototypes with soft ordinal weighting over continuous scores and explicit geometric ordering constraints for regression.

3 Method

Let $\mathbf{X} \in \mathbb{R}^{T\times D}$ denote input with T time steps. We predict continuous severity $y \in [0, Y_{\max}]$ while leveraging clinical thresholds $\{b_1, \ldots, b_{K-1}\}$ defining K severity grades. Fig. 2 and Fig. 3 illustrate our framework.

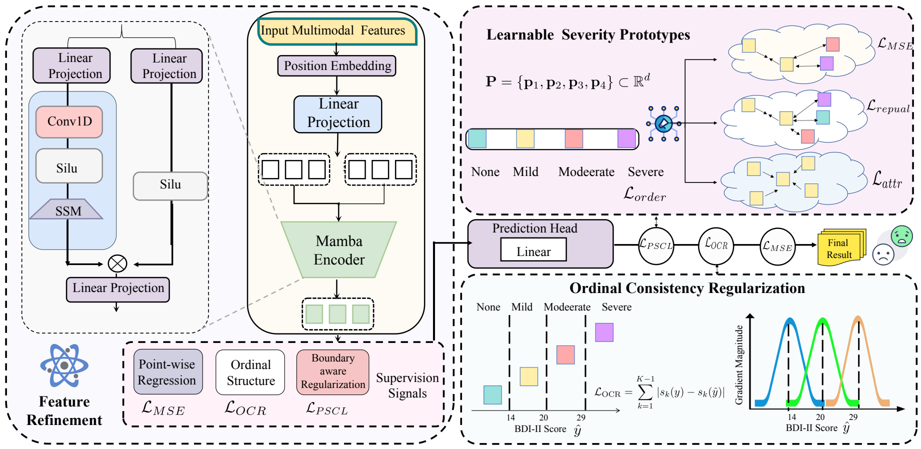

Fig. 3. Detailed architecture showing the Mamba encoder, prediction head, learnable severity prototypes, and ordinal consistency regularization.

3.1 Ordinal Consistency Regularization (OCR)

For clinical threshold b_k, we define differentiable soft assignment $s_k(v) = \sigma\left(\frac{v-b_k}{\tau}\right)$, where $\tau > 0$ controls transition sharpness. The vector $\mathbf{s}(v) = [s_1(v), \ldots, s_{K-1}(v)]^\top$ encodes cumulative ordinal membership. OCR penalizes squared discrepancies in ordinal assignments:

$$\mathcal{L}_{\text{OCR}} = \sum_{k=1}^{K-1} \left(s_k(y) - s_k(\hat{y})\right)^2 \tag{1}$$

Theoretical Analysis. We characterize OCR's gradient behavior.

Proposition 1 (Gradient Form). *The gradient of $\mathcal{L}_{OCR}$ with respect to $\hat{y}$ is:*

$$\frac{\partial \mathcal{L}_{OCR}}{\partial \hat{y}} = -\frac{2}{\tau} \sum_{k=1}^{K-1} (s_k(y) - s_k(\hat{y})) \cdot s_k(\hat{y})(1 - s_k(\hat{y})) \tag{2}$$

Theorem 1 (Boundary-Aware Gradient Amplification). *For each threshold b_k, the gradient magnitude $|s_k(\hat{y})(1 - s_k(\hat{y}))|$ achieves its unique maximum at $\hat{y} = b_k$:*

$$\max_{\hat{y}} |s_k(\hat{y})(1 - s_k(\hat{y}))| = \frac{1}{4} \quad \text{at} \quad \hat{y} = b_k \tag{3}$$

with exponential decay as $|\hat{y} - b_k|$ increases.

Proof. Let $g(\hat{y}) = s_k(\hat{y})(1-s_k(\hat{y}))$. By the chain rule, $g'(\hat{y}) = s_k'(\hat{y})(1-2s_k(\hat{y})) = 0 \Leftrightarrow \hat{y} = b_k$ since $s_k' > 0$. As $s(1-s)$ is strictly concave in $s \in (0,1)$ with $d^2/ds^2 = -2 < 0$, g attains unique global maximum $1/4$ at $\hat{y} = b_k$. As $|\hat{y} - b_k|$ increases, $s_k(\hat{y}) \to 0$ or 1, hence $g(\hat{y})$ decays exponentially.

Theorem 1 proves ordinal supervision concentrates learning signal at clinical boundaries. Near thresholds, maximum gradient ensures strongest correction; far from boundaries, diminished gradient allows MSE to dominate.

3.2 Prototype-based Severity Contrastive Learning (PSCL)

PSCL enforces ordinal geometry in feature space. We introduce K learnable prototypes $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_K\} \subset \mathbb{R}^d$ representing canonical severity embeddings. Given target y, soft assignment weights based on proximity to severity centers $c_k = (b_{k-1} + b_k)/2$ are:

$$w_k(y) = \frac{\exp(-|y - c_k|/\tau_p)}{\sum_{j=1}^{K} \exp(-|y - c_j|/\tau_p)} \tag{4}$$

PSCL employs three losses for attraction, repulsion, and ordering:

$$\begin{aligned}
\mathcal{L}_{\text{attr}} &= 1 - \textstyle\sum_{k=1}^{K} w_k(y) \cdot \cos(\mathbf{f}, \mathbf{p}_k), \\
\mathcal{L}_{\text{repul}} &= \textstyle\sum_{k=1}^{K} w_k(y) \cdot \sum_{j \neq k} \frac{1}{1+|k-j|} \cdot \max(0, \cos(\mathbf{f}, \mathbf{p}_j)), \\
\mathcal{L}_{\text{order}} &= \textstyle\sum_{k=1}^{K-1} \max(0, m - \|\mathbf{p}_k - \mathbf{p}_{k+1}\|_2).
\end{aligned} \tag{5}$$

where $\frac{1}{1+|k-j|}$ decays with ordinal distance, emphasizing adjacent grades.

The total PSCL loss is $\mathcal{L}_{\text{PSCL}} = \mathcal{L}_{\text{attr}} + \lambda_r \mathcal{L}_{\text{repul}} + \lambda_o \mathcal{L}_{\text{order}}$.

3.3 Multi-Granularity Training Objective

The complete objective integrates both granularities:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{MSE}}}_{\text{fine-grained}} + \underbrace{\lambda_{\text{ocr}} \mathcal{L}_{\text{OCR}}}_{\substack{\text{coarse-grained} \\ \text{(prediction)}}} + \underbrace{\lambda_{\text{pscl}} \mathcal{L}_{\text{PSCL}}}_{\substack{\text{coarse-grained} \\ \text{(feature)}}} \tag{6}$$

OCR and PSCL serve as regularizers. The primary task remains continuous regression; ordinal knowledge shapes optimization dynamics via Theorem 1 and feature geometry without altering output formulation.

4 Experiments

Datasets and Implementation. Using official splits, we evaluate visual/audio/audio-visual inputs on AVEC 2013/2014 [8,9] (BDI-II, 0–63) and DAIC-WOZ/E-DAIC [10] (PHQ-8, 0–24). Inputs use HOG+SVM faces and STFT-based log-Mel audio features. We use Mamba with AdamW (lr=10^{-5}, 100 epochs), λ_{ocr}=λ_{pscl}=0.1.

4.1 Main Results and Ablation Study

Tables 1 and 2 demonstrate state-of-the-art performance across all four benchmarks.

With the backbone fixed, Table 3 shows that PSCL, OCR, and the full objective consistently improve over MSE-only training, confirming loss-level gains without changing the architecture.

Table 1. Comparison on AVEC2013 and AVEC2014 test sets. "V" = visual unimodal; "Multimodal" = audio-visual. **Bold** = best; $\underline{\text{underlined}}$ = second-best.

Modal	Method	Year	AVEC2013		AVEC2014	
			MAE↓	RMSE↓	MAE↓	RMSE↓
V	MTDAN [11]	2023	6.14	8.08	6.35	7.93
	STA-DRN [12]	2023	6.15	7.98	6.00	7.75
	AVA-DepNet [13]	2023	5.97	7.26	5.59	6.83
	DepressionMLP [14]	2024	5.43	7.49	5.63	7.27
	PointTransform [2]	2024	5.62	7.36	$\underline{5.41}$	7.12
	SETD-Net [15]	2025	6.06	7.60	6.03	7.59
	MLGI [16]	2025	6.82	8.25	6.12	8.07
	MemRank [17]	2025	5.82	7.78	5.77	7.69
	LMTformer [18]	2025	6.12	7.75	6.05	7.97
	DepMGNN [19]	2025	$\underline{5.41}$	$\underline{7.08}$	5.45	$\underline{6.62}$
	Ours	–	**5.34**	**6.69**	**5.37**	**6.48**
Multi.	AVA-DepNet [13]	2023	6.23	7.99	$\underline{5.32}$	$\underline{6.83}$
	Self-attn+GFN [1]	2024	7.52	10.15	7.86	10.36
	CNN+BiLSTM+ADF [20]	2024	6.48	8.91	7.01	9.38
	FC-AEN [21]	2025	$\underline{5.11}$	$\underline{6.41}$	$\underline{5.32}$	7.00
	Ours	–	**5.06**	**6.34**	**5.15**	**6.39**

Table 2. Performance on DAIC-WOZ and E-DAIC test sets. "V" = visual; "A" = audio; "Multi." = multimodal; "–" = not reported. **Bold** = best; $\underline{\text{underlined}}$ = second-best.

Modal	DAIC-WOZ				E-DAIC			
	Method	Year	MAE↓	RMSE↓	Method	Year	MAE↓	RMSE↓
V	Rand. Forest [22]	2023	5.30	6.26	DPD Net [26]	2024	–	6.19
	AVA-DepNet [13]	2023	$\underline{4.78}$	$\underline{5.96}$	SIMMA [23]	2025	5.36	6.41
	SIMMA [23]	2025	4.98	6.04	FC-AEN [21]	2025	$\underline{5.35}$	**6.39**
	Ours	–	**4.41**	**5.13**	**Ours**	–	**5.20**	**6.39**
A	Rand. Forest [22]	2023	5.49	6.64	VQWTNet [24]	2024	5.38	6.29
	VQWTNet [24]	2024	5.38	6.36	SIMMA [23]	2025	5.33	6.47
	SIMMA [23]	2025	$\underline{4.88}$	$\underline{5.81}$	FC-AEN [21]	2025	$\underline{5.19}$	$\underline{6.21}$
	Ours	–	**4.37**	**5.52**	**Ours**	–	**4.87**	**5.98**
Multi.	SAGAN-M [25]	2022	$\underline{5.13}$	$\underline{5.78}$	MM [27]	2024	$\underline{5.21}$	–
	Rand. Forest [22]	2023	5.38	6.35	SIMMA [23]	2025	5.31	$\underline{6.37}$
	Ours	–	**4.33**	**5.33**	**Ours**	–	**5.02**	**6.05**

4.2 Analysis and Validation

Fig. 4 provides empirical validation of our theoretical findings. The OCR loss distribution shows elevated values around several severity transition regions, broadly consistent with the clinical thresholds at 14, 20, and 29, and supports the boundary-aware behavior analyzed in Theorem 1. PCA visualization demonstrates ordinal clustering in feature space with smooth severity transitions. Fig. 5 validates PSCL's effectiveness through uniform prototype attraction, strong inter-class separation, and satisfied ordinal constraints.

Table 3. Ablation study on AVEC2013 and AVEC2014 datasets.

			AVEC2013		AVEC2014	
Baseline	OCR	PSCL	MAE↓	RMSE↓	MAE↓	RMSE↓
✓	×	×	5.80	6.89	5.94	7.37
✓	×	✓	5.52	6.78	5.59	7.30
✓	✓	×	5.22	6.65	5.30	6.83
✓	✓	✓	**5.06**	**6.34**	**5.15**	**6.39**
Wilcoxon signed-rank			$W = 945$, $p = 0.001$		$W = 976$, $p < 0.001$	
Paired t-test			$t(49) = 2.87$, $p = 0.006$		$t(49) = 3.59$, $p < 0.001$	
Bootstrap 95% CI			[0.36, 1.98]†		[0.59, 2.02]†	

Note: †CI excludes zero, indicating significant improvement.

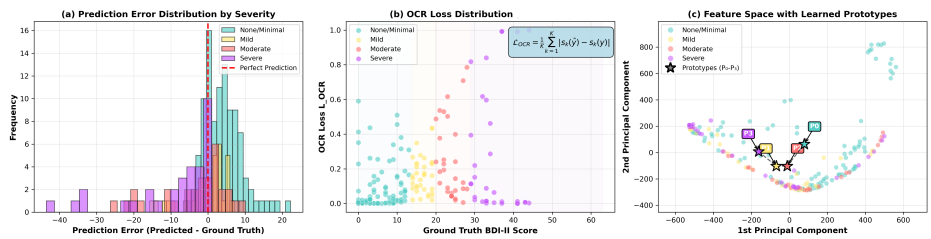

Fig. 4. Empirical analysis on AVEC 2014 validating Theorem 1. (a) Prediction errors are concentrated near zero. (b) OCR loss shows elevated values around several severity transition regions, broadly consistent with the clinical thresholds at 14, 20, and 29. (c) PCA visualization shows ordinal clustering with learned prototypes.

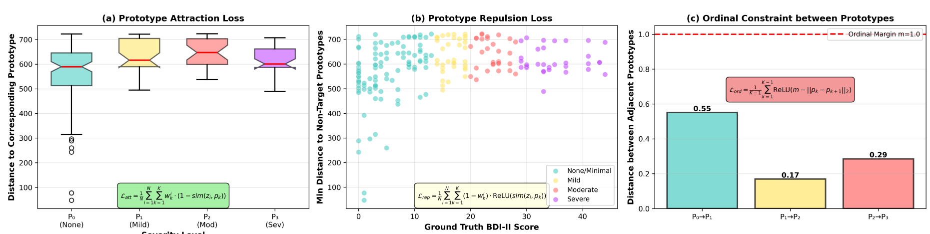

Fig. 5. PSCL component analysis on AVEC 2014. (a) Attraction loss. (b) Repulsion loss. (c) Ordinal constraint satisfaction with margin m=1.0.

Table 4 demonstrates that moderate regularization weights (λ_{ocr}=λ_{pscl}=0.1) achieve optimal performance. Excessive regularization (λ=0.5) compromises results, confirming that OCR and PSCL serve as auxiliary signals complementing rather than dominating the primary regression objective.

Table 4. Parameter sensitivity analysis on AVEC2013 and AVEC2014. Moderate regularization (λ=0.1) achieves optimal balance; over-regularization (λ=0.5) degrades performance.

		AVEC2013				AVEC2014			
λ_{OCR}	λ_{PSCL}	MAE↓	RMSE↓	CCC↑	PCC↑	MAE↓	RMSE↓	CCC↑	PCC↑
0.05	0.05	5.40	6.55	0.80	0.85	5.62	7.02	0.78	0.80
0.05	0.1	5.53	6.75	0.78	0.85	5.69	7.19	0.78	0.79
0.5	0.5	5.70	7.03	0.75	0.83	5.51	6.73	0.79	**0.85**
0.1	0.1	**5.06**	**6.34**	**0.81**	**0.86**	**5.15**	**6.39**	**0.82**	**0.85**

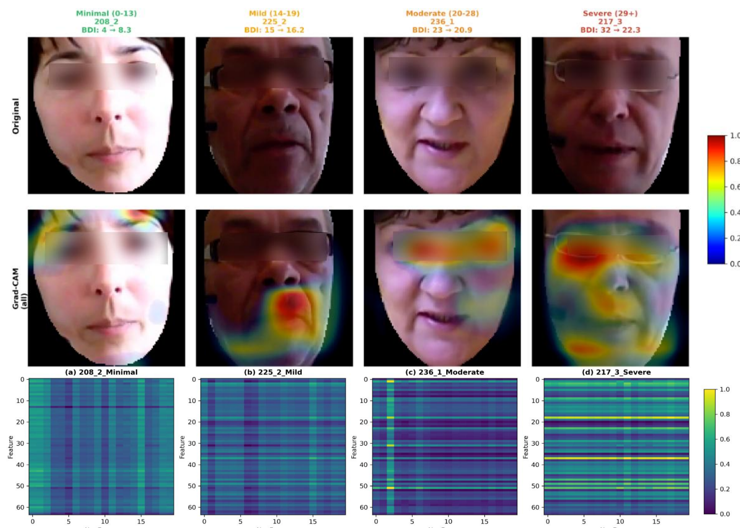

Fig. 6. Qualitative analysis across severity levels. Grad-CAM attention and temporal activations exhibit ordinal progression correlating with clinical severity.

4.3 Qualitative Results

Fig. 6 visualizes model interpretability across severity levels. Grad-CAM highlights clinically-relevant facial regions (eyes, mouth) with attention intensity increasing by severity. Temporal activations exhibit clear ordinal structure, empirically confirming boundary-aware learning.

5 Conclusion

We identified the granularity mismatch problem and proposed Multi-Granularity Supervision to bridge fine-grained regression with coarse-grained clinical structure. Our theoretical analysis shows that ordinal regularization concentrates learning at clinical thresholds, and results on four benchmarks validate the boundary-aware design. Broader validation across larger and more diverse cohorts remains future work.

Acknowledgments. Supported by the National Natural Science Foundation of China (No. 62366055), Yunnan Fundamental Research Projects (No. 202401BC070006), Yunnan Univ. Graduate Practice and Innovation Project (No. ZC-252512430), and Yunnan Provincial Dept. of Education Science Research Fund (No. 2026Y0248).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fan, H., Zhang, X., Xu, Y., Fang, J., Zhang, S., Zhao, X., Yu, J.: Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Inf. Fusion **104**, 102161 (2024)
2. Niu, M., Li, M., Fu, C.: PointTransform networks for automatic depression level prediction via facial keypoints. Knowl.-Based Syst. **297**, 111951 (2024)
3. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit. Lett. **140**, 325–331 (2020)
4. Diaz, R., Marathe, A.: Soft labels for ordinal regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4738–4747 (2019)
5. Wang, X., Tan, T., Gao, Y., Marcus, E., Han, L., Portaluri, A., Zhang, T., Lu, C., Liang, X., Beets-Tan, R., Teuwen, J., Mann, R.: Ordinal learning: Longitudinal attention alignment model for predicting time to future breast cancer events from mammograms. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15001, pp. 155–165. Springer, Cham (2024)
6. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 4077–4087 (2017)
7. Saleem, A., Lewis, J.R., Gilani, S.Z.: A hybrid contrastive ordinal regression method for advancing disease severity assessment in imbalanced medical datasets. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15972, pp. 14–23. Springer, Cham (2025)
8. Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 3–10 (2013)
9. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.: AVEC 2014: 3D dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 3–10 (2014)
10. Ringeval, F., Schuller, B., Valstar, M., 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 (AVEC), pp. 3–12 (2019)
11. Zhang, S., Zhang, X., Zhao, X., Fang, J., Niu, M., Zhao, Z., Yu, J., Tian, Q.: MTDAN: A lightweight multi-scale temporal difference attention networks for automated video depression detection. IEEE Trans. Affect. Comput. **15**(3), 1078–1089 (2024)

12. Pan, Y., Shang, Y., Liu, T., Shao, Z., Guo, G., Ding, H., Hu, Q.: Spatial-temporal attention network for depression recognition from facial videos. Expert Syst. Appl. **237**, 121410 (2024)
13. Pan, Y., Shang, Y., Shao, Z., Liu, T., Guo, G., Ding, H.: Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. IEEE Trans. Affect. Comput. **15**(3), 828–836 (2024)
14. Niu, M., Li, Y., Tao, J., Zhou, X., Schuller, B.W.: DepressionMLP: A multi-layer perceptron architecture for automatic depression level prediction via facial keypoints and action units. IEEE Trans. Circuits Syst. Video Technol. **34**(9), 8924–8938 (2024)
15. Huang, Z., Chen, Y., Ren, F., Hu, M., Liu, J.: SETD-Net: Multi-scale spatiotemporal enhancement and multi-granularity temporal difference network for depression detection. Biomed. Signal Process. Control **112**, 108900 (2026)
16. Niu, M., Shao, Z., He, Y., Tao, J., Schuller, B.W.: Multimodal local global interaction networks for automatic depression severity estimation. IEEE Trans. Circuits Syst. Video Technol. **36**(2), 2649–2664 (2025)
17. Li, Y., Wei, Z., Guo, G., Zhou, X.: MemRank: Memory-augmented similarity ranking for video-based depression severity estimation. IEEE Trans. Affect. Comput. **16**(3), 2062–2073 (2025)
18. He, L., Zhao, J., Zhang, J., Jiang, J., Qi, S., Wang, Z., Wu, D.: LMTformer: facial depression recognition with lightweight multi-scale transformer from videos. Appl. Intell. **55**(2), 195 (2025)
19. Wu, Z., Zhou, L., Li, S., Fu, C., Lu, J., Han, J., Zhang, Y., Zhao, Z., Song, S.: DepMGNN: Matrixial graph neural network for video-based automatic depression assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence **39**(2), 1610–1619 (2025)
20. Zhang, X., Li, B., Qi, G.: A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method. Biomed. Signal Process. Control **96**, 106552 (2024)
21. Wang, Y., Lin, Z., Teng, Y., Cheng, Y., Jiang, H., Yang, Y.: FC-AEN: Fully convolutional adaptive ensemble network for automatic depression detection. IEEE Trans. Comput. Soc. Syst. **13**(2), 1381–1396 (2025)
22. Gupta, R.K., Sinha, R.: An investigation on the audio-video data based estimation of emotion regulation difficulties and their association with mental disorders. IEEE Access **11**, 74324–74336 (2023)
23. Wang, Y., Lin, Z., Teng, Y., Cheng, Y., Jiang, H., Yang, Y.: SIMMA: Multimodal automatic depression detection via spatiotemporal ensemble and cross-modal alignment. IEEE Trans. Comput. Soc. Syst. **12**(5), 3548–3564 (2025)
24. Han, Z., Shang, Y., Shao, Z., Liu, J., Guo, G., Liu, T., Ding, H., Hu, Q.: Spatial-temporal feature network for speech-based depression recognition. IEEE Trans. Cogn. Dev. Syst. **16**(1), 308–318 (2024)
25. Zhao, Z., Wang, K.: Unaligned multimodal sequences for depression assessment from speech. In: 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 3409–3413 (2022)
26. He, M., Bakker, E.M., Lew, M.S.: DPD (DePression Detection) Net: a deep neural network for multimodal depression detection. Health Inf. Sci. Syst. **12**(1), 53 (2024)
27. Teng, S., Chai, S., Liu, J., Tateyama, T., Lin, L., Chen, Y.W.: Multi-modal and multi-task depression detection with sentiment assistance. In: 2024 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–5 (2024)