Journal of Zhejiang University-SCIENCE B (Biomed & Biotechnol)
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.B2500705

Three-dimensional face reconstruction for emotional dynamics: a novel approach to distinguishing bipolar disorder from major depressive disorders

Tao DU[1,3], Jinchao GE[2], Hong LYU[1,3], Yutong ZHANG[1,4], Xin XU[5,6,7]✉, Shaohua HU[1,6,7,9]✉, Jingkai CHEN[1,8]✉

[1] *Department of Psychiatry, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China*

[2] *School of Computing and Information Technology, University of Wollongong, Northfields Avenue, Wollongong NSW 2522, Australia*

[3] *School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China*

[4] *Department of Computer Science, Loughborough University, Loughborough LE11 3TU, UK*

[5] *School of Public Health, The Second Affiliated Hospital of School of Medicine, Zhejiang University, Hangzhou 310058, China*

[6] *Nanhu Brain-computer Interface institute, Hangzhou 311100, China*

[7] *The State Key Lab of Brain-Machine Intelligence, MOE Frontier Science Center for Brain Science and Brain-Machine Integration, Zhejiang University School of Medicine, Hangzhou 310003, China*

[8] *The Zhejiang Key Laboratory of Precision Psychiatry, Hangzhou 310003, China*

[9] *Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou 310000, China*

Abstract: Major depressive disorder (MDD) and bipolar disorder (BD) are frequently misdiagnosed in clinical practice due to their overlapping symptoms, resulting in delayed treatment and an increased burden on patients. Reliable biomarkers for early differential diagnosis are currently lacking. To address this, we developed a clinical facial video dataset of patients with BD and MDD and introduced Adaptive Shape Model with Vision-to-Language (AsmV2L), a vision-language-guided fine-grained 3D face reconstruction framework that better preserves affect-related facial cues from monocular videos. Using valence-arousal (VA) representations estimated from these reconstructed 3D faces, we identified distinct affective patterns that differentiated the two disorders. In subject-level downstream evaluation, the emotional dynamic features extracted by our framework achieved 88.9% accuracy in distinguishing BD from MDD. Additionally, experiments on public benchmark datasets demonstrated that AsmV2L maintains competitive 3D reconstruction accuracy while preserving emotion-related semantic information more effectively.

Key words: Major depressive disorder; Bipolar disorder; 3D facial reconstruction; Emotion dynamics; Valence and arousal

---

✉ Jingkai CHEN, wzcjk922@163.com

Shaohua HU, dorhushaohua@zju.edu.cn

Xin XU, xuxinsummer@zju.edu.cn

Jingkai CHEN, https://orcid.org/0000-0002-1654-9466

Tao DU, https://orcid.org/0009-0001-0173-7974

Received Nov. 9, 2025; Revision accepted Apr. 19, 2026;

1 Introduction

Major depressive disorder (MDD) and bipolar disorder (BD) are among the leading causes of disability worldwide, yet their clinical differential diagnosis remains a significant challenge. The symptomatic presentation of depressive episodes in BD, in particular, overlaps with MDD, leading to misdiagnosis in up to 20% of BD patients (Wang et al., 2025; Chen et al., 2024). This can cause treatment delays and potentially trigger manic switches due to inappropriate antidepressant prescribing, potentially exacerbating the disease cycle and placing severe long-term burdens on patients' social functioning, quality of life, and economic status (Dai et al., 2025; Singh et al., 2025).

This dilemma arises because traditional diagnosis primarily relies on subjective clinical interviews and lacks objective, quantifiable biological markers. Despite advancements in identifying potential biomarkers in fields such as neuroimaging and molecular biology (Loosen et al., 2025; Shi et al., 2025; Kang et al., 2024; Preller et al., 2024; Sarmin et al., 2024), these approaches face significant barriers to widespread clinical adoption, including equipment cost, operational complexity, and limited specificity and sensitivity. Finding a non-invasive, low-cost, and objective auxiliary diagnostic tool is therefore urgently needed in the field of mental health.

Facial expressions, the most direct and comprehensive medium for conveying human emotions, hold significant clinical importance due to their dynamic nature. Research indicates that while MDD patients often exhibit "emotional blunting," characterized by reduced facial expressivity and slow responses, BD patients display "emotional instability," marked by intense fluctuations between emotional states (Jahromi et al., 2025; Martikkala et al., 2025; Steardo et al., 2025; Fagiolini et al., 2021). Early two-dimensional (2D) facial expression analysis methods utilizing feature point tracking and texture variations (e.g., Haar-like features for face detection, optical flow algorithms for tracking facial landmarks, and support vector machines (SVMs) or convolutional neural networks (CNNs) trained on 2D video datasets) have successfully captured basic emotion-related macro-expressions (Chen et al., 2025; Jin et al., 2025; Jiang et al., 2024; Younis et al., 2024). However, such methods are highly susceptible to interference from head pose, lighting, occlusion, and other factors, and struggle to precisely quantify the subtle three-dimensional (3D) movements of facial muscles, limiting their application to precision medicine.

Three-dimensional face reconstruction technology enables the recovery of high-fidelity 3D facial geometry from images captured in uncontrolled environments. Compared with 2D methods, 3D geometric information can provide more robust and precise measurements of key expression units, such as furrowed brows, flared nostrils, and raised mouth corners (Wang et al., 2024; Sun et al., 2023; Zheng et al., 2023). Early research primarily focused on fitting 3D morphable models to extract identity features under neutralized expressions or on analyzing the displacements of specific vertices to identify macro-expression categories (Cen et al., 2024; Zhu et al.,2024; Zhang et al., 2025). Recently, advancements in parametric face models and differentiable rendering techniques have allowed researchers to reconstruct high-fidelity dynamic 3D facial sequences from monocular videos and to further regress continuous emotional dimensions, such as valence and arousal. Some methods achieve good emotion reconstruction in controlled environments by jointly optimizing identity, expression, and pose parameters (Hu et al., 2025; Wang et al., 2024; Sadeghi et al., 2024; Yin et al., 2025). Text-guided 3D face reconstruction methods incorporate language priors into facial modeling, usually by using text as semantic guidance for attribute control or reconstruction optimization (Wu et al., 2024; Aneja et al., 2023). Additionally, some studies have performed affect analysis based on reconstructed 3D faces by first estimating parametric facial representations, such as 3D morphable model (3DMM) coefficients, and then regressing emotional states from these reconstructed features, with the aim of reducing the influence of identity, pose, and illumination variations while preserving expression-related cues for emotion analysis (Dong et al., 2024; Kollias et al., 2023; Kollias et al., 2021). However, effectively correlating the reconstructed 3D geometric features with clinically interpretable emotional dimensions, such as valence and arousal, and utilizing these correlations for disease differentiation, remains largely unexplored.

To address these challenges, this study integrates cutting-edge computer vision technology with clinical requirements in mental health. We propose an adaptive shape model with vision-to-language 3D face reconstruction framework (AsmV2L; see Fig. 1a), which introduces a fine-grained textual semantic chunking mechanism to achieve precise binding between text descriptions and local facial parameters, and constructs an adaptive weighted loss function based on pixel-vertex spatial mapping to strengthen geometric and semantic constraints on expression-sensitive regions. This approach thereby facilitates the high-precision reconstruction of details, even for extreme or subtle facial expressions. Building on this, we systematically identified the unique emotional dynamics patterns distinguishing patients with BD from those with MDD using a specialized facial video dataset, as shown in Fig. 1b. Finally, we conducted comprehensive evaluations across public benchmark datasets and clinical tasks, demonstrating that AsmV2L outperforms mainstream methods in both emotion recognition and assessment, while achieving reconstruction accuracy comparable to strong baseline models.

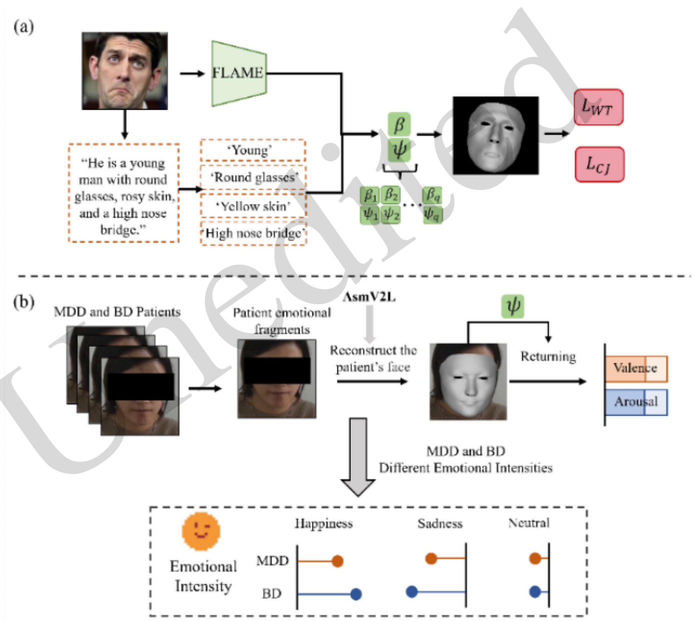

Fig. 1 Overall framework (a) Overview of the proposed AsmV2L framework; (b) Emotional dynamic analysis conducted on the constructed MDD and BD patient dataset. FLAME: Faces Learned with an Articulated Model and Expressions; L_{WT}: Language-guided Weighted Text-to-Image Loss; L_{CJ}: Joint Image Reconstruction Loss.

The main contributions of this study are summarized as follows:

(1) We propose AsmV2L, a vision-language-guided adaptive 3D face reconstruction framework that leverages fine-grained textual semantic chunking and pixel-to-vertex spatial mapping to achieve precise semantic binding and adaptive optimization for emotion-sensitive facial regions.

(2) We constructed a clinical facial video dataset comprising 130 patients with BD and 132 patients with MDD, and, based on reconstructed 3D facial dynamics, systematically characterized their

distinct valence-arousal patterns, providing potential objective cues for auxiliary differential diagnosis.

(3) We conducted comprehensive evaluations on public benchmark datasets and clinical tasks, demonstrating that AsmV2L achieves strong 3D reconstruction performance while outperforming mainstream methods in emotion recognition, assessment, and semantic alignment.

2 Experiment details

2.1 Benchmark and evaluation datasets

Training datasets: **FFHQ256** contains 70,000 high-quality aligned face images at 256 × 256 resolution, covering substantial diversity in age, ethnicity, and facial appearance. It was used to support large-scale learning of face appearance and geometry. **CelebA** contains 202,599 facial images from 10,177 identities, with rich pose variation and complex background conditions. It was used to improve the model's robustness across diverse real-world facial appearances. **AffectNet** contains approximately 440,000 manually annotated face images and provides both discrete emotion labels and continuous valence-arousal annotations. In this study, AffectNet was not only an important source for affect-related training supervision but also a public benchmark for emotion evaluation.

Emotion benchmark: AffectNet was used as the public benchmark for affective evaluation. For continuous emotion prediction, a four-layer MLP was trained to regress valence and arousal from the reconstructed expression-related representations. For discrete expression evaluation, expression labels were classified from the predicted 3DMM parameters. Additionally, following prior protocols, we trained an image-to-image UNet converter with AffectNet, using rendered images and randomly sampled occluded images as inputs, and employed it to compute the VGG-Loss between the reconstructed and original input images.

Reconstruction benchmark: REALY was used as the benchmark for 3D reconstruction evaluation. It contains 100 high-precision scanned neutral-expression faces and provides a standard protocol for regional geometric error analysis. Following this benchmark, we evaluated reconstruction accuracy in four local regions—the nose, mouth, forehead (including the eyes and eyebrows), and cheeks—under both frontal and lateral views.

Clinical dataset and evaluation protocol: We constructed a clinical facial video dataset for psychiatric affective analysis, collected data at the First Affiliated Hospital of Zhejiang University over a 19-month period. The dataset used for the present BD–MDD analysis includes 262 patients: 130 with BD and 132 with MDD (Clinical trial registration number: NCT05608135). All patients underwent structured interviews conducted by professionally trained mental health clinicians using the Mini-International Neuropsychiatric Interview (MINI), and diagnoses were established according to DSM-IV criteria. Exclusion criteria for the patient group were: (1) severe mental disorders other than BD and MDD, such as schizophrenia spectrum disorders or intellectual disability; (2) severe head trauma (loss of consciousness for more than 5 minutes), current or past epilepsy, intracranial hypertension, or other serious neurological conditions; (3) alcohol or substance abuse/dependence within the 6 months before testing; and (4) investigator-determined unsuitability for participation or refusal to participate. Each participant completed recordings under two ecologically meaningful scenarios: (1) watching a standardized emotion-elicitation video, and (2) participating in a natural doctor–patient conversation. The recording duration for each participant was approximately 20–40 min. Two downstream clinical tasks were designed. The first was valence-arousal (VA) dynamic analysis, in which we selected the three most frequent emotional segments: happy, sad, and neutral. For each category, 100 15-second segments were randomly sampled and preprocessed. The second task was BD–MDD differential classification, which involved constructing disease representations from expression-related dynamic 3DMM coefficients and VA dynamic statistics. All experiments were evaluated under a strict subject-level 5-fold cross-validation protocol.

2.2 Benchmark and evaluation datasets

Table 1 AsmV2L Implementation details.

Item	**Setting**
Overall framework	Text-guided parameter modulation branch + joint 2D-3D learning branch
2D-3D module	Pre-trained ResNet-18 encoder; 6×6×512 feature map; five upsampling layers; 192×192×32 dense 2D feature map; global 2D-3D representation dimension 1×1024
Pretraining	Expression, pose, and identity encoders pretrained using landmark loss and the MICA shape loss
Full-model training	40 epochs; Adam optimizer; initial learning rate 1e-4; batch size 12; input size 224×224
Runtime environment	Linux workstation with an NVIDIA GeForce RTX 4090 GPU
Clinical video	Videos decoded into RGB frames; frame-wise face detection/cropping and landmark alignment;

preprocessing	frames with failed detection, severe occlusion, or obvious motion blur discarded; retained crops resized to 224×224
Clip sampling	15-s segments sampled at 10 frames per second
Temporal regressor G(·)	LSTM encoder followed by a four-layer MLP; pretrained on AffectNet-related affective supervision and kept frozen during the clinical analysis

As shown in Table 1, AsmV2L was implemented in PyTorch and trained under a unified pipeline that covered model configuration, optimization protocol, and runtime environment. The overall framework comprises a text-guided parameter modulation branch and a joint 2D-3D learning branch. For the joint 2D-3D module, a pre-trained ResNet-18 was adopted as the convolutional encoder to produce a 6 × 6 × 512 feature map. The decoder comprised five upsampling layers and output a dense 2D facial feature map of size 192 × 192 × 32. The dimensionality of the global 2D-3D representation was set to 1 × 1024.

We first pre-trained the expression, pose, and identity encoders using a landmark loss and the MICA shape loss to stabilize geometry- and identity-related representations. The full model was then trained for 40 epochs with the Adam optimizer, using an initial learning rate of 1e-4 and a batch size of 12. All input frames were resized to 224 × 224 pixels. Experiments were conducted on a Linux workstation equipped with an NVIDIA GeForce RTX 4090 GPU. For all competing methods, we either used publicly available implementations or re-implemented them under closely matched preprocessing and evaluation protocols to ensure a fair comparison.

For the clinical video validation, all videos were first decoded into RGB frames. Faces were detected and cropped frame-by-frame using a face detector, followed by facial landmark alignment. Frames with failed face detection, severe occlusion, or obvious motion blur were discarded. The retained face crops were resized to 224 × 224 pixels. For clip-based analysis, each 15-second segment was sampled at 10 frames per second. The temporal regressor described in Materials and methods took as input frame-level expression-related 3DMM coefficient sequences and predicted frame-level valence and arousal. This regressor consisted of an LSTM encoder followed by a four-layer MLP. It was pretrained on AffectNet-related affective supervision data and was kept frozen during the subsequent clinical analysis. Details of the AsmV2L method are provided in the Materials and methods section of the supplementary information (Fig. S1).

2.3 Evaluation metrics

Clinical evaluation metrics: The clinical evaluation involved two tasks: affective dynamics analysis and differential diagnosis between BD and MDD. For the affective dynamics analysis, we report the mean and standard deviation (mean ± SD) of frame-level valence and arousal estimates within each emotional segment category (happy/sad/neutral), enabling group-wise comparisons of affective baseline and variability between BD and MDD. For differential diagnosis, we adopted subject-level 5-fold cross-validation and report standard diagnostic metrics, including accuracy (ACC) and the area under the ROC curve (AUC) as primary measures. We also report sensitivity and specificity to characterize the trade-off between BD detection and false-positive control.

Emotion recognition and affect regression: To evaluate affect-related capability, we considered both continuous and discrete emotion settings. For continuous valence-arousal prediction, we report the concordance correlation coefficient for valence (V-CCC) and arousal (A-CCC), together with the corresponding root mean square errors (V-RMSE and A-RMSE). For discrete expression recognition, we report expression classification accuracy (E-ACC). In addition, to assess perceptual consistency between the reconstructed rendering and the original input image, we used VGG-Loss as an auxiliary metric reflecting reconstruction quality under affective facial variations.

3D reconstruction accuracy: To evaluate 3D facial reconstruction quality, we used region-wise geometric error in millimeters (mm) as the primary metric. Specifically, the distance between the reconstructed mesh and the reference scan was computed after alignment, and the mean ± standard deviation was reported for four facial regions, namely the nose, mouth, forehead/eyebrow region, and cheeks. Evaluations were conducted under both frontal and lateral views. This protocol enables a more fine-grained assessment of local reconstruction fidelity, especially in facial regions highly relevant to expression analysis.

3 Results

3.1 Results of emotion analysis of psychiatric patients

3.1.1 Valence-arousal dynamic analysis of BD and MDD

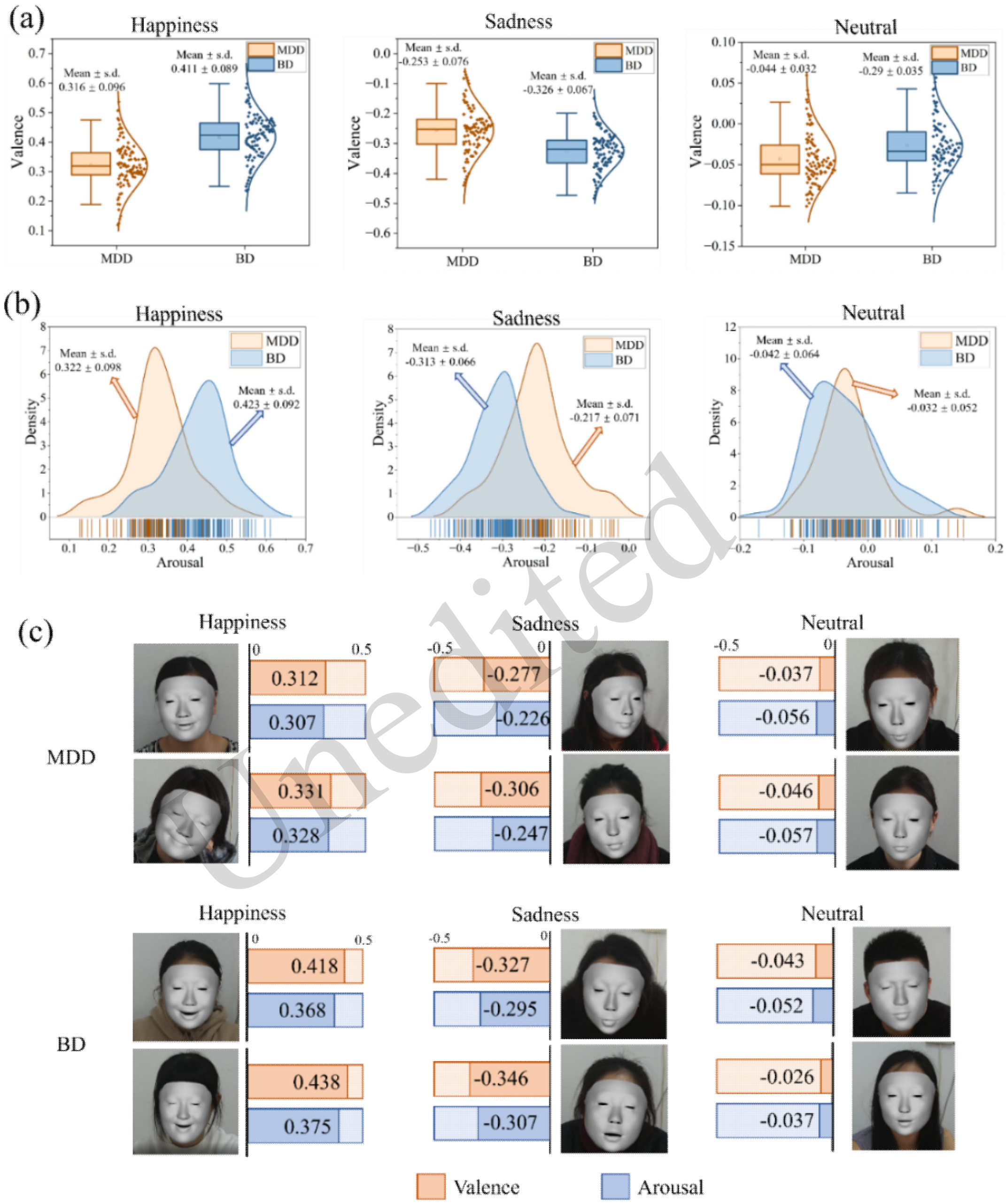

Fig. 2 Analysis of emotional valence and arousal in the constructed MDD and BD datasets; (a) Valence analysis of MDD and BD patients under happy, sad, and neutral conditions, respectively; (b) Arousal analysis of MDD and BD patients under happy, sad, and neutral conditions, respectively; (c) 3D reconstructed facial visualization and valence-arousal display for some patients.

We used VA captured from the reconstructed 3D faces as the primary affective representation for differentiating BD and MDD. Compared with discrete emotion categories, VA provides a continuous and compact description of affective state, making it better suited to characterizing subtle temporal variations in emotional tone and activation in natural facial videos (Schneider et al., 2025; Dong et al., 2024; Kollias et al., 2023). As shown in Figs. 2a and 2b, the mean ± standard deviation of valence and arousal are reported for BD and MDD under the *happy*, *sad*, and *neutral* emotional segments.

In the happy emotion, BD patients show valence of 0.411±0.089 and arousal of 0.423±0.092, while MDD patients show valence of 0.316±0.096 and arousal of 0.322±0.098. BD patients exhibit higher valence and arousal, indicating emotional elation and pronounced emotional fluctuations, whereas MDD patients continue to display emotional blunting even in the happy state. In the sad condition, the valence and arousal of BD patients are -0.326±0.067 and -0.313±0.066, respectively, while for MDD patients they are -0.253±0.076 and -0.217±0.071, respectively. BD patients exhibit lower valence and arousal in this condition, indicating that they experience a stronger sense of despair during the depressive episode. There is no significant emotional difference between the two groups in the neutral condition, but both groups still predominantly show negative feelings, with BD patients demonstrating a stronger negative tendency. This indicates that both groups of patients exhibit a negative basic emotional tone, and the lower valence of BD patients may be related to their emotional instability.

As shown in Fig. 2c, we visualized the 3D faces and emotional valence and arousal of MDD and BD patients. The valence difference between positive and negative emotions is about 0.75 in BD patients, demonstrating more pronounced emotional fluctuations.

3.1.2 Differential classification of BD and MDD based on reconstructed affective dynamics

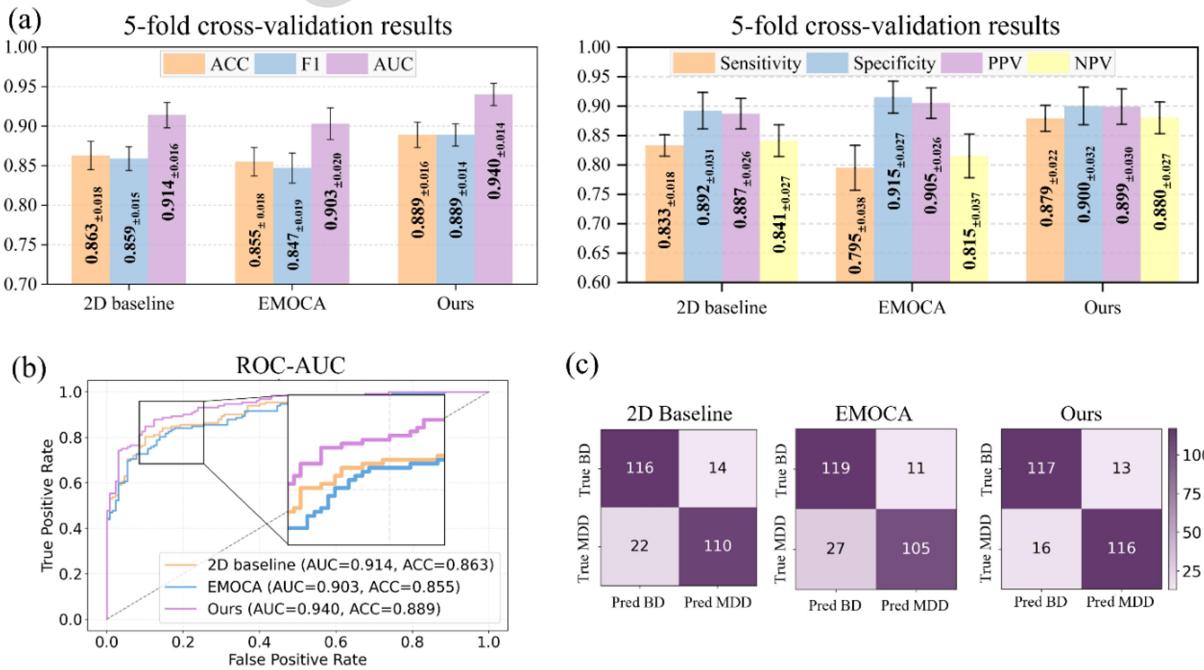

Fig. 3 BD/MDD differential classification results based on reconstructed affective dynamics. (a) Classification performance of three methods under 5-fold cross-validation; (b) ROC curves and AUC; (c) confusion matrices.

To examine whether reconstructed affective dynamics can support differential diagnosis, we further conducted a classification experiment to distinguish BD from MDD, comparing three methods: (i) a 2D affect-feature baseline, (ii) Emotion Driven Monocular Face Capture and Animation (EMOCA), and (iii) our method. For the 2D baseline, frame-level 2D affective descriptors were directly extracted from the raw videos and aggregated into subject-level representations for classification. For EMOCA and our method, we followed the procedure described in Materials and methods to construct disease representations based on expression-related dynamic 3DMM coefficients and VA sequences reconstructed from facial videos, and evaluated all methods under a strict 5-fold cross-validation protocol.

As shown in Fig. 3a, our method achieved the best overall performance, with an accuracy (ACC) of 0.889±0.016 and an AUC of 0.940±0.014, representing improvements of 2.6 and 2.6 percentage points over the 2D baseline, respectively. The ROC curves in Fig. 3b further demonstrate the superior discriminative ability of our method, indicating that the proposed 3D affective dynamic representation yields a clearer decision boundary. As shown in Fig. 3c, our method reduced the number of MDD→BD misclassifications to 16 and improved sensitivity by 4.6 percentage points compared with the 2D baseline. Notably, although EMOCA achieved higher specificity (0.915±0.027), its sensitivity was substantially lower (0.795±0.038), suggesting a more conservative decision tendency that led to more missed BD cases. Overall, these results indicate that the fine-grained 3D affective dynamics reconstructed by our framework can more effectively capture the contrast between the emotional blunting commonly observed in MDD and the affective instability more frequently seen in BD, thereby providing more discriminative auxiliary cues for differential diagnosis.

3.2 Results of 3D face reconstruction

We compared AsmV2L with several state-of-the-art 3D facial reconstruction methods, including DECA (Feng et al. 2021), EMOCA (Daněček et al., 2022), Deep3DFace (Deng et al., 2019), and 3DDFA-V3 (Wang et al., 2024). Deep3DFace used the latest PyTorch-based version for the comparison. DECA and EMOCA used coarsely tuned encoders because their fine-tuned versions tend to be affected by artifacts, thereby compromising emotion capture.

3.2.1 Evaluation of 3D facial reconstruction prediction

The quantitative results on the REALY benchmark are reported in Table 2. AsmV2L achieved average reconstruction errors of 1.441 mm and 1.456 mm under frontal and lateral views, respectively. These results are highly competitive with the strongest baseline, 3DDFA-V3, which achieved 1.436 mm and 1.442 mm under the same two views. More importantly, AsmV2L achieved the best reconstruction accuracy across several expression-relevant regions, especially the nose and cheek areas. For example, under the frontal view, AsmV2L achieved 1.452 mm on the nose and 1.106 mm on the cheek region, while under the lateral view, it achieved 1.024 mm on the cheek region. These findings suggest that the proposed semantics-aware region modulation and adaptive region-weighted optimization are particularly effective in preserving fine-grained local facial details while maintaining competitive global geometric fidelity.

Qualitative visual comparisons of reconstructed faces are provided in Fig. 4 Compared with the competing methods, AsmV2L more faithfully preserves both facial shape and expression-related appearance details, resulting in reconstructed faces that are visually more consistent with the original inputs.

Table 2. REALLY benchmark-based quantitative comparison.

Method	Front Veiw (mm) ↓					Side Veiw (mm) ↓				
	Nose Avg.±std.	Mouth Avg.±std.	Forehead Avg.±std.	Cheek Avg.±std.	Avg.	Nose Avg.±std.	Mouth Avg.±std.	Forehead Avg.±std.	Cheek Avg.±std.	Avg.
MGCNet	1.771±0.380	1.417±0.409	2.268±0.503	1.639±0.650	1.774	1.827±0.383	1.409±0.418	2.248±0.508	1.665±0.644	1.787
Deep3dFace	1.719±0.354	1.368±0.439	2.015±0.449	1.528±0.501	1.657	1.749±0.343	1.411±0.395	2.074±0.486	1.528±0.517	1.691
DECA	1.694±0.355	2.516±0.839	2.394±0.576	1.479±0.535	2.010	1.903±1.050	2.472±1.079	2.423±0.720	1.630±1.135	2.107
3DDFA-V2	1.903±0.517	1.597±0.478	2.447±0.647	1.757±0.642	1.926	1.883±0.499	1.642±0.501	2.465±0.622	1.781±0.636	1.943
3DDFA-V3	1.586±0.306	**1.238±0.373**	**1.810±0.394**	1.111±0.327	**1.436**	1.623±0.313	**1.205±0.366**	**1.864±0.424**	1.076±0.315	**1.442**
EMOCA	1.868±0.387	2.679±1.112	2.426±0.641	1.438±0.501	2.103	1.867±0.554	2.636±1.284	2.448±0.708	1.548±0.590	2.125
Ours	**1.452±0.335**	1.246±0.413	1.852±0.384	**1.106±0.315**	1.441	**1.608±0.346**	1.281±0.425	1.911±0.443	**1.024±0.307**	1.456

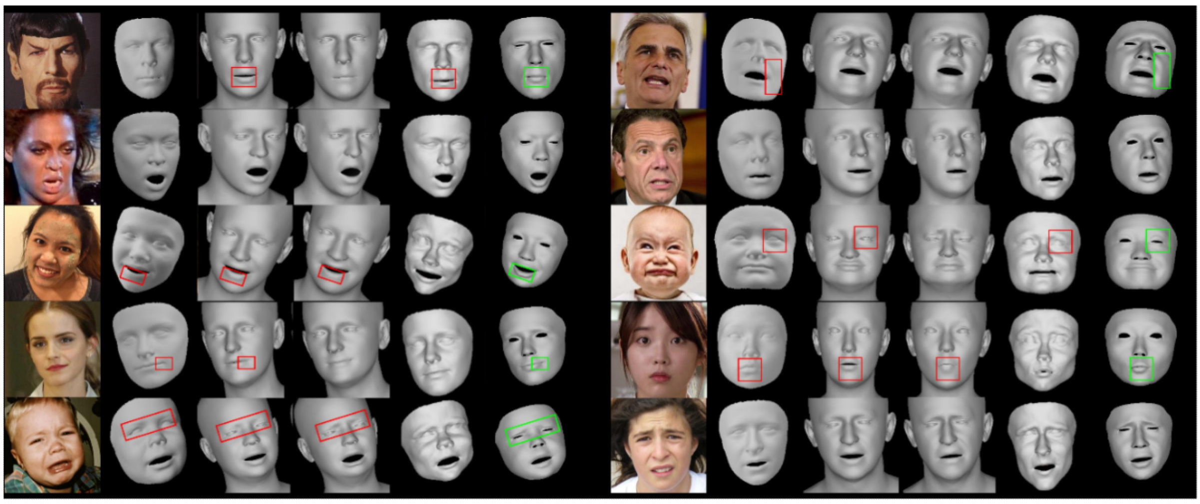

Fig. 4 Visual comparison of face reconstruction.

3.2.2 Evaluation of emotion recognition

To verify that the proposed framework can better preserve affect-related facial cues, we evaluated emotion recognition and affect regression on AffectNet. As shown in Fig. 5a, our method achieved the best performance in evaluations of both emotion recognition and the effects of rendered image reconstruction. The emotion recognition-related evaluations V-CCC, A-CCC, E-ACC, V-RMSE, and A-RMSE are 0.79, 0.71, 0.70, 0.30, and 0.29, respectively, and the reconstruction image-related evaluation VGG-Loss is 0.74. This may be related to our model's focus on local facial attributes, as these regions (such as the eyes and mouth) often play a decisive role in expression recognition.

In addition, we conducted a discrete emotion recognition evaluation on the AffectNet dataset, which includes eight emotions: neutral, happy, sad, surprised, fearful, disgusted, angry, and contemptuous. We compared DECA, EMOCA_v2, and our AsmV2L. The experimental results are presented in Fig 5b. Deca has an average accuracy of 58.50%, Emoca_v2 has an average accuracy of 66.63%, and our model has the highest average accuracy of 71.25%, indicating that our model has better emotion recognition ability. In addition, facial models generally perform better on positive emotions (happy), but relatively poorly on negative emotions (angry, sad, fearful, disgusted). However, AsmV2L also performs well on some negative emotions (disgusted, angry). Fig. 5c shows the t-SNE distribution of 3D facial expression features. AsmV2L forms obvious clusters and shows a similar trend to discrete emotion recognition.

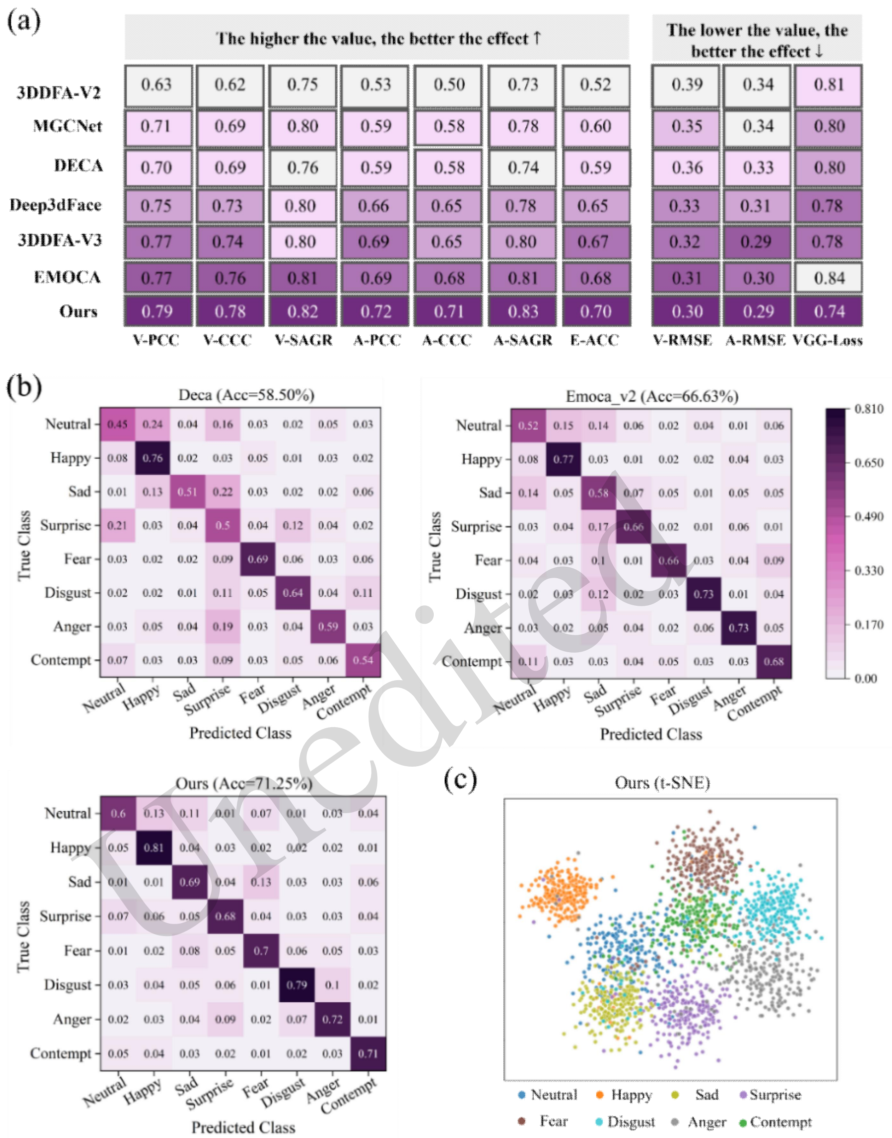

Fig. 5 Results of emotion recognition. (a) Emotion results on the AffectNet dataset. (b) Eight-category emotion classification results. (c) t-SNE distribution of facial expression features.

3.3 Results of the ablation study

To clarify the impact of each core module in the AsmV2L model on 3D facial reconstruction accuracy, this section presents an ablation study that quantitatively analyzes the function of different modules. The experiment strictly follows the protocol described in Section 3.1, with VGG Loss and expression classification accuracy (E-ACC) adopted as evaluation metrics. Detailed results are shown in Table 3.

The baseline model without any auxiliary loss terms performs worst, with a VGG Loss as high as 0.99 and an E-ACC of only 0.47, indicating significant pixel-level differences between the reconstructed and real images. When the image reconstruction loss module is first introduced based on the baseline model, the model performance is significantly improved: VGG Loss decreases from 0.99 to 0.85, a decrease of 14.1%, indicating that this loss effectively reduces the difference between reconstructed and real images by constraining pixel-level errors; E-ACC increases by 25.5%, from 0.47 to 0.59, demonstrating that the image reconstruction loss not only optimizes geometric structures but also indirectly improves the extraction effect of expression features. Subsequently, by retaining the image reconstruction loss, the text reconstruction loss module is introduced, further optimizing the model's reconstruction accuracy: VGG Loss decreases to 0.81, and E-ACC increases to 0.62. This result demonstrates that the text reconstruction loss can guide the model to more accurately focus on key facial features in text descriptions by establishing a connection between textual semantics and local facial regions. However, compared with the initial improvement from the image reconstruction loss, the optimization range of the text reconstruction loss is relatively limited, reflecting the complementarity between the two in terms of mechanism.

Table 3 Component Ablation

baseline	Image Reconstruction	Text Reconstruction	Area weighting	VGG Loss↓	E-ACC↑
√				0.99	0.47
√	√			0.85	0.59
√	√	√		0.81	0.62
√	√	√	√	0.72	0.70

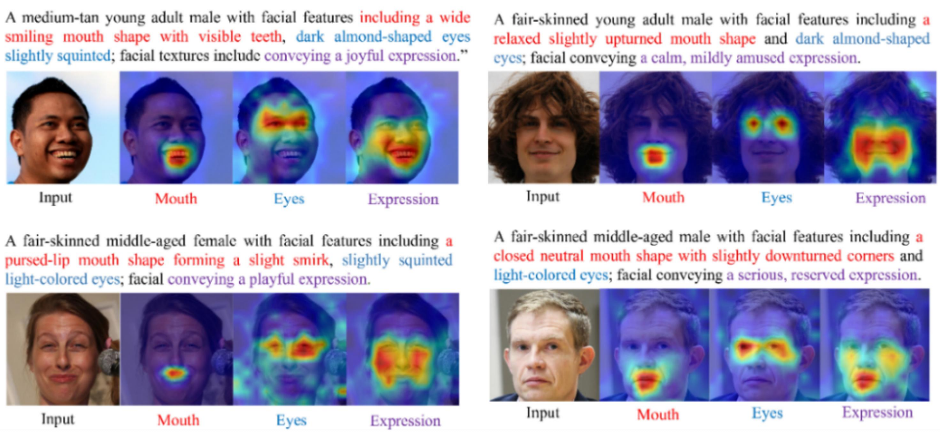

Fig. 6 Visualization of keyword-level attention heatmaps for fine-grained semantic-region binding.

Finally, to retain both the image reconstruction loss and the text reconstruction loss, the adaptive regional weighting module is introduced, leading to a significant leap in the model's reconstruction accuracy: VGG Loss decreases to 0.72, and E-ACC increases to 0.70. This result demonstrates that the adaptive regional weighting module can dynamically adjust the optimization weights of different regions based on differences in the contributions of various facial regions to reconstruction accuracy and expression features, thereby strengthening the constraints on expression-sensitive regions.

To verify that the proposed text-guided modulation can learn meaningful fine-grained semantic binding, we visualized keyword-level attention heatmaps for representative facial descriptions, as shown in Fig 6. The proposed fine-grained text chunking strategy establishes semantically meaningful correspondences between linguistic cues and local facial regions. In particular, expression-level semantics typically induce broader activations spanning multiple expression-sensitive regions, as facial expressions are jointly conveyed by multiple local muscle groups.

4 Discussion

This study proposed a fine-grained adaptive 3D face reconstruction method, guided by vision-language (AsmV2L), to accurately extract emotional dynamic features from monocular videos and assist in the early differentiation of MDD and BD. Experimental results showed that AsmV2L maintained competitive 3D reconstruction performance on public benchmarks, while achieving favorable results in emotion recognition and semantic alignment tasks. Compared with conventional 3D face reconstruction frameworks, the proposed method better preserved expression-sensitive local facial cues, thereby providing a more reliable basis for affect-related representation learning and downstream psychiatric analysis. Based on the self-constructed clinical facial video dataset, we not only systematically revealed the differentiated dynamic patterns of emotional valence and arousal between BD and MDD patients, but also further validated the discriminative value of these reconstructed affective representations in a subject-level downstream classification task. Specifically, BD patients exhibited “emotional instability,” with more pronounced fluctuations between positive and negative affective states, whereas MDD patients showed “emotional blunting,” characterized by persistently reduced affective reactivity even under positive emotional stimulation. In addition, the subject-level differential classification results further indicated that the emotional dynamic features extracted by AsmV2L could effectively support the distinction between BD and MDD. These findings suggest that reconstructed 3D facial affect dynamics represent not only descriptive markers of emotional differences, but also potentially useful objective cues for auxiliary differential assessment, especially in the clinically challenging scenario of distinguishing BD depressive episodes from MDD.

The innovation of AsmV2L lies in two aspects: first, through fine-grained textual semantic chunking and local parameter binding, it achieves precise modeling of key expression regions; second, it constructs an adaptive weighted loss function based on pixel-vertex mapping, thereby strengthening geometric and semantic constraints on expression-sensitive regions. Ablation experiments further verified the synergistic effect of image reconstruction loss, text alignment loss, and the regional weighting mechanism in improving reconstruction accuracy and semantic consistency.

While AsmV2L performs excellently in multiple evaluations, this study has certain limitations. First, although the clinical dataset is substantial in size, its generalization ability still needs further verification on larger-scale, multi-center data, given its collection from a single medical center. Second, the regression of emotional dimensions still relies on pre-trained downstream networks; future work could explore end-to-end emotion-geometry joint modeling frameworks to further enhance the interpretability and discriminative power of emotional features. Furthermore, the quantitative mapping of emotional dimensions to clinical symptoms remains to be explored in larger samples and at finer temporal scales. Looking ahead, we will delve deeper into the following directions: first, expanding to multimodal data fusion, including speech and physiological signals, to build a more comprehensive emotional assessment system; second, exploring a universal emotional dynamic modeling framework across the disease spectrum.

5 Conclusions

This paper addressed the challenges in the early differentiation of MDD and BD by proposing a fine-grained 3D face reconstruction framework, AsmV2L, guided by vision-language. This framework, through local binding of text semantics and facial regions, joint learning of 2D-3D spatial mapping, and adaptive weighted multimodal loss design, achieved high-precision 3D reconstruction and semantic alignment under complex expressions. We systematically analyzed the emotional dynamic patterns of BD and MDD patients in our self-constructed clinical video dataset, finding that BD patients exhibited significant emotional fluctuations, whereas MDD patients exhibited persistent low mood and a slower response. In public benchmark tests and multiple downstream tasks, AsmV2L outperformed existing mainstream methods, achieving leading levels of geometric reconstruction error and demonstrating stronger capabilities in emotion recognition and semantic consistency. This study confirms the significant potential of emotion dynamic features derived from 3D face reconstruction for differentiating mental disorders, offering a new concept for developing non-invasive, objective, and low-cost auxiliary diagnostic tools. In the future, we will continue to optimize the model's robustness in complex environments, extend it to additional types of mental disorders, and explore its integration with multimodal data, such as neuroimaging and genomics, to further enhance its clinical applicability and interpretability.

Materials and methods

Detailed methods are provided in the electronic supplementary materials of this paper.

Data availability statement

The data provided in this study are available from the corresponding author upon reasonable request.

Acknowledgments

This work was supported by the National Key Research and Development Program of China 2023YFC2506200), the National Natural Science Foundation of China (82571735), the Key R&D Program of Zhejiang Province (2024C03098；2025C02109；2025C01104).

Author contributions

Tao Du: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing. Jinchao Ge: Methodology, Software, Investigation, Writing – Original Draft, Writing – Review & Editing. Hong Lyu: Data Curation, Formal Analysis, Investigation, Writing – Review & Editing. Yutong Zhang: Formal Analysis, Investigation, Writing – Review & Editing. Xin Xu: Resources, Conceptualization, Supervision, Writing – Review & Editing. Shaohua Hu: Resources, Funding acquisition, Conceptualization, Supervision, Writing – Review & Editing. Jingkai Chen: Resources, Conceptualization, Methodology, Supervision, Writing – Original Draft, Writing – Review & Editing.

Compliance with ethics guidelines

Tao Du, Jinchao Ge, Hong Lyu, Yutong Zhang, Xin Xu, Shaohua Hu and Jingkai Chen declare that they have no conflicts of interest.

The use of private datasets strictly adheres to the privacy protection policies of the hospital and relevant ethical review guidelines to ensure patient privacy and security. The use of this privacy dataset was approved for use by the hospital (Ethics Approval Number: #2019-1181 and #2021-382) and complies with all relevant laws and regulations.

Declaration on the use of generative AI tools

No generative AI tools were used in the preparation of this manuscript.

References

Chen, L., Xu, Y.-Y., Lin, J.-Y., et al., 2024. The prevalence and clinical correlates of suicide attempts in patients with bipolar disorder misdiagnosed with major depressive disorder: Results from a national survey in China. Asian Journal of Psychiatry, 93:103958.

https://doi.org/10.1016/j.ajp.2024.103958

Wang, P., Bai, Y.L., Xiao, Y., et al., 2025. Aberrant network topological structure of sensorimotor superficial white-matter system in major depressive disorder. Journal of Zhejiang University-SCIENCE B, 26(1):39-51.

https://doi.org/10.1631/jzus.B2300880

Singh, B., Swartz, H.A., Cuellar-Barboza, A.B., et al., 2025. Bipolar disorder. The Lancet, 406(10506):963-978.

https://doi.org/10.1016/S0140-6736(25)01140-7

Dai, C., Fu, Y.Y., Li, X.W., et al., 2025. Clinical efficacy and safety of vortioxetine as an adjuvant drug for patients with bipolar depression. Journal of Zhejiang University-SCIENCE B, 26(1):26-38.

https://doi.org/10.1631/jzus.B2400470

Sarmin, N., Roknuzzaman, A.S.M., Mouree, T.Z., et al., 2024. Evaluation of serum interleukin-12 and interleukin-4 as potential biomarkers for the diagnosis of major depressive disorder. Scientific Reports, 14(1):1652.

https://doi.org/10.1038/s41598-024-51932-9

Kang, X., Liu, X., Chen, S., et al., 2024. Major depressive disorder recognition by quantifying EEG signal complexity using proposed APLZC and AWPLZC. Journal of Affective Disorders, 356:105-114.

https://doi.org/10.1016/j.jad.2024.03.169

Preller, K.H., Scholpp, J., Wunder, A., et al., 2024. Neuroimaging Biomarkers for Drug Discovery and Development in Schizophrenia. Biological Psychiatry, 96(8):666-673. https://doi.org/10.1016/j.biopsych.2024.01.009

Loosen, A.M., Kato, A., Gu, X., 2025. Revisiting the role of computational neuroimaging in the era of integrative neuroscience. Neuropsychopharmacology, 50(1):103-113.

https://doi.org/10.1038/s41386-024-01946-8

Shi, B., Yu, B., Chen C., et al., 2025. Guidelines for Cognitive Clinical Diagnosis and Treatment of Coronary Artery Disease Complicated with Depression and Anxiety: December 2025 Update. Heart and Mind 9(6):p 462-471.

DOI: 10.4103/hm.HM-D-25-00151

Fagiolini, A., Florea, I., Loft, H., et al., 2021. Effectiveness of Vortioxetine on Emotional Blunting in Patients with Major Depressive Disorder with inadequate response to SSRI/SNRI treatment. Journal of Affective Disorders, 283:472-479.

https://doi.org/10.1016/j.jad.2020.11.106

Steardo, L., D'Angelo, M., Monaco, F., et al., 2025. Decoding neural circuit dysregulation in bipolar disorder: Toward an advanced paradigm for multidimensional cognitive, emotional, and psychomotor treatment. Neuroscience & Biobehavioral Reviews, 169:106030.

https://doi.org/10.1016/j.neubiorev.2025.106030

Jahromi, G.G, Rezaei, N, 2025. Connecting the Dots: NLRP3 Inflammasome as a Key Mediator in the Intersection of Depression and Cardiovascular Disease – A Narrative Review. Heart and Mind 9(1):p 48-60.

DOI: 10.4103/hm.HM-D-24-00076

Martikkala, A., Baryshnikov, I., Granroth-Wilding, H., et al., 2025. Temporal variations of depressive symptoms in patients with bipolar, borderline personality, and major depressive disorder: an ecological momentary assessment study. Journal of Psychiatric Research, 191:313-322.

https://doi.org/10.1016/j.jpsychires.2025.09.054

Younis, E.M.G., Mohsen, S., Houssein, E.H., et al., 2024. Machine learning for human emotion recognition: a comprehensive review. Neural Computing and Applications, 36(16):8901-8947.

https://doi.org/10.1007/s00521-024-09426-2

Chen, X., Lu, Y., Cue, J.M., et al., 2025. Classification of schizophrenia, bipolar disorder and major depressive disorder with comorbid traits and deep learning algorithms. Schizophrenia, 11(1):14.

https://doi.org/10.1038/s41537-025-00564-7

Jiang, X., Cao, B., Li, C., et al., 2024. Identifying misdiagnosed bipolar disorder using support vector machine: feature selection based on fMRI of follow-up confirmed affective disorders. Translational Psychiatry, 14(1):9.

https://doi.org/10.1038/s41398-023-02703-z

Jin, R., Zhou, R., Zhang, D., 2025. Recent advances in antibody optimization based on deep learning methods. Journal of Zhejiang University-SCIENCE B, 26(5):409-420.

https://doi.org/10.1631/jzus.B2400387

Wang, J., Yu, C., Li, H., 2024. Multi-modal Feature Guided Detailed 3D Face Reconstruction from a Single Image. In: Liu, Q., Wang, H., Ma, Z., et al. (Eds.), Pattern Recognition and Computer Vision, p.356-368.

https://doi.org/10.1007/978-981-99-8432-9_29

Sun, J., Wang, X., Wang, L., et al., 2023. Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20991-21002.

https://doi.org/10.1109/CVPR52729.2023.02011

Zheng, Y., Wang, Y., Wetzstein, G., et al., 2023. PointAvatar: Deformable Point-Based Head Avatars from Videos. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21057-21067.

https://doi.org/10.1109/CVPR52729.2023.02017

Cen, M., Shen, H., Zhao, W., et al., 2024. A Survey of Text-guided 3D Face Reconstruction. In: 2024 3rd International Conference on Image Processing and Media Computing (ICIPMC), 82-87. https://doi.org/10.1109/ICIPMC62364.2024.10586613

Zhu, C.A., Joslin, C., 2024. A review of motion retargeting techniques for 3D character facial animation. Computers & Graphics, 123:104037.

https://doi.org/10.1016/j.cag.2024.104037

Zhang, Q., Feng, J., 2025. Disentangled Geometry and Appearance for Efficient Multi-View Surface Reconstruction and Rendering. arXiv preprint: arXiv:2508.17436. https://arxiv.org/abs/2508.17436

Hu, R., Wang, X., Zhao, C., 2025. Identity aware 3D face reconstruction from in-the-wild images. Neurocomputing, 641:130299.

https://doi.org/10.1016/j.neucom.2025.130299

Wang, Z., Zhu, X., Zhang, T., et al., 2024. 3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1672-1682.

https://doi.org/10.1109/CVPR52733.2024.00165

Sadeghi, M., Richer, R., Egger, B., et al., 2024. Harnessing multimodal approaches for depression detection using large language models and facial expressions. npj Mental Health Research, 3(1):66.

https://doi.org/10.1038/s44184-024-00112-8

Yin, G., Yuan, J., Chen, Y., et al., 2025. Schizophrenia recognition based on three-dimensional adaptive graph convolutional neural network. Scientific Reports, 15(1):4067.

https://doi.org/10.1038/s41598-024-84497-8

Wu, Y., Meng, Y., Hu, Z., Li, L., Wu, H., Zhou, K., Xu, W., Yu, X., 2024. Text-guided 3D face synthesis: From generation to editing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1260-1269.

https://doi.org/10.1109/CVPR52733.2024.00126

Aneja, S., Thies, J., Dai, A., Nießner, M., 2023. ClipFace: Text-guided editing of textured 3D morphable models. SIGGRAPH '23: ACM SIGGRAPH 2023 Conference Proceedings, Article 70:1-11.

https://doi.org/10.1145/3588432.3591566

Dong, L., Wang, X., Setlur, S., Govindaraju, V., Nwogu, I., 2025. Ig3D: Integrating 3D face representations in facial expression inference. Computer Vision -- ECCV 2024 Workshops, 404-421.

https://doi.org/10.1007/978-3-031-91581-9_29

Kollias, D., Tzirakis, P., Baird, A., Cowen, A., Zafeiriou, S., 2023. ABAW: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW):5889-5898.

https://doi.org/10.1109/CVPRW59228.2023.00626

Kollias, D., Zafeiriou, S., 2021. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. arXiv preprint, arXiv:2103.15792.

https://doi.org/10.48550/arXiv.2103.15792

Li, T., Bolkart, T., Black, M.J., et al., 2017. Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph., 36(6):194.

https://doi.org/10.1145/3130800.3130813

Karras, T., Laine, S., Aila, T., 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4396-4405.

https://doi.org/10.1109/CVPR.2019.00453

Liu, Z., Luo, P., Wang, X., et al., 2015. Deep Learning Face Attributes in the Wild. In: 2015 IEEE International Conference on Computer Vision (ICCV), 3730-3738. https://doi.org/10.1109/ICCV.2015.425

R, M.B., Tewari, A., Seidel, H.P., et al., 2021. Learning Complete 3D Morphable Face Models from Images and Videos. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3360-3370.

https://doi.org/10.1109/CVPR46437.2021.00337

Chai, Z., Zhang, H., Ren, J., et al., 2022. REALY: Rethinking the Evaluation of 3D Face Reconstruction. arXiv preprint: arXiv:2203.09729.

https://arxiv.org/abs/2203.09729

Tellamekala, M.K., Sümer, Ö., Schuller, B.W., et al., 2024. Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?. IEEE Transactions on Affective Computing, 15(2):535-548.

https://doi.org/10.1109/TAFFC.2023.3280530

Feng, Y., Feng, H., Black, M.J., et al., 2021. Learning an animatable detailed 3D face model from in-the-wild images. ACM Trans. Graph., 40(4):88.

https://doi.org/10.1145/3450626.3459936

Daněček, R., Black, M., Bolkart, T., 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20279-20290.

https://doi.org/10.1109/CVPR52688.2022.01967

Deng, Y., Yang, J., Xu, S., et al., 2019. Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 285-295.

https://doi.org/10.1109/CVPRW.2019.00038

Wang, Z., Zhu, X., Zhang, T., et al., 2024. 3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1672-1682.

https://doi.org/10.1109/CVPR52733.2024.00165

Retsinas, G., Filntisis, P.P., Daněček, R., et al., 2024. 3D Facial Expressions through Analysis-by-Neural-Synthesis. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2490-2501.

https://doi.org/10.1109/CVPR52733.2024.00241

Schneider H, Pavlitska S, Gremmelmaier H, Zöllner M (2025). Datasets for valence and arousal inference: A survey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 5703–5710.

https://doi.org/10.48550/arXiv.2510.00738

Supplementary information

Fig. S1; Materials and methods