Dep-MAP: A Multi-level Alignment Framework with Semantic Prototypes for Video-based Automatic Depression Assessment

Hao Wang[1, 2*], **Jiayu Ye**[3*], **Qingxiang Wang**[1, 2†]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[2]Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Jinan, China

[3]School of Computer Science, Guangdong University of Technology, Guangzhou, China

10431230232@stu.qlu.edu.cn, yejiayu97@outlook.com, wangqx@qlu.edu.cn

Abstract

Spatiotemporal analysis of facial behavior is a crucial method for evaluating the mental state of depression patients. However, in practice, depressed patients often display facial behaviors similar to healthy individuals due to masking tendencies. Additionally, facial expressions among depressed patients are also different, increasing the difficulty of assessment. To address this, we propose a video-based automatic depression assessment model Dep-MAP for complex facial behaviors of depression patients. Dep-MAP adopts a dual-branch architecture to extract visual features of facial behavior and capture corresponding emotional semantic features. Specifically, the extracted deep semantic features are clustered, resulting in semantically distinct prototype sets, where each severity group learns a set of discriminative facial behavior prototype representations, to suppress inter-class semantic confusion. Subsequently, we propose a semantic prototype-supervised contrastive learning method, which aligns latent semantics between shallow and deep features, realizing emotional semantic guidance and self-knowledge distillation for the visual feature branch, effectively suppressing intra-class difference. Then, we integrate key depression cues across multiple spatiotemporal scales via a multi-scale weighted fusion strategy, achieving automatic depression assessment. Experimental results demonstrate that Dep-MAP effectively identifies potential key frames in temporal sequences, and aggregates key frame representations with semantic consistency, achieving significantly superior state-of-the-art results on the AVEC2013 and AVEC2014 public datasets.

Code — https://github.com/QLUTEmoTechCrew/DepMap

Introduction

Nonverbal facial behavior is one of the primary channels of information transmission in interpersonal communication, and is also a key indicator for distinguishing differences in psychological and social adaptability between healthy individuals and depressed patients(Nowicki Jr and Duke 1994).

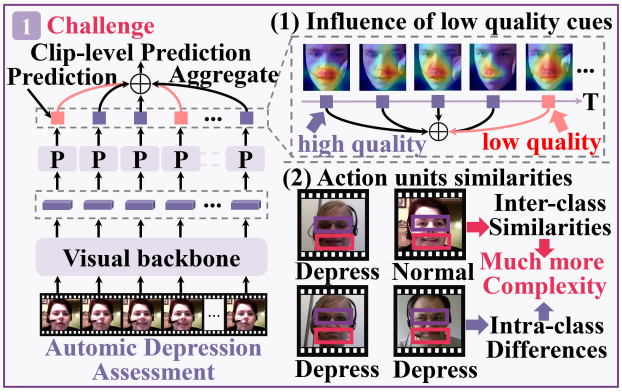

Figure 1: Challenges in existing video ADA methods.

Clinical studies have demonstrated that facial behaviors in depressed patients exhibit marked differences compared to those of healthy individuals (Di et al. 2021), typically characterized by reduced smiling, downward-curved mouth corners, furrowed brows, and a generally flat or blank facial expression. In recent years, automatic depression assessment (ADA) based on facial video has gained widespread attention(Wen et al. 2015). Early ADA research based on facial videos primarily relied on handcrafted facial features (Wen et al. 2015; Pampouchidou et al. 2016). Notably, the AVEC challenge series (Valstar et al. 2013, 2014) also adopted such handcrafted features as standard inputs to facilitate benchmarking of automatic depression recognition systems. However, such handcrafted features typically suffer from poor generalizability and limited sensitivity to depression-related cues, which promoted the development of deep learning-based ADA research. Conventional deep learning methods (Niu et al. 2022b; Shang et al. 2021) typically extract depression-related static features independently from each frame, and then aggregate information across all frames through feature-level or decision-level fusion, to perform video-level depression assessment. However, a major limitation of these methods is that they treat all video seg-

*These authors contributed equally.

†Corresponding author (wangqx@qlu.edu.cn).

ments equally, failing to account for the fact that crucial facial behavior cues are often sparse and unevenly distributed across the temporal dimension (Shangguan et al. 2022).

Several methods (Shangguan et al. 2022; Ye et al. 2024) have adopted sparse key signal measurement approaches to extract depression-related expression patterns from frame-level facial features. Although these methods enhance the ability to perceive key frames, they often overlook depression-related emotional semantic information. To address this, recent studies have increasingly incorporated emotional semantic information associated with facial action units, guiding models to explicitly capture the semantic differences in facial actions between healthy individuals and depressed patients (Niu et al. 2024; Fu et al. 2025).

However, depression-related facial behaviors display notable complexity at the semantic level, on one hand, depressed individuals and healthy subjects may exhibit similar facial behaviors under different semantic contexts; on the other hand, depressed patients may present distinct facial expressions even within the same semantic context. In real-world settings, the subjectivity of clinical diagnosis and patient deception (Hook and Andrews 2005) further exacerbate intra-class differences and inter-class similarity issues at the semantic level, resulting in failure of semantic guidance or feature ambiguity. Additionally, these methods do not effectively integrate correlations between multi-level semantic and visual information, potentially causing shifts in distance metrics during discrimination and weakening the sensitivity of model to depression-related features. Therefore, as illustrated in Figure 1, existing methods still confront significant challenges in modeling depression-related facial behavior cues in real-world face videos.

To this end, we propose a novel semantic-guided prototype clustering method designed for sparse key signal measurement. This method partitions deep semantic features into clusters that are semantically distinct, and dynamically updates prototypes for different depression levels using the empirical means of each cluster. In this way, the prototypes can capture discriminative facial behavior features that distinguish them from other semantic categories, thereby forming semantically diverse prototype representations; meanwhile, each potential key action unit-related feature in facial images has the chance to be assigned to the most responsive semantic prototype, ensuring that each class-specific prototype representation covers a comprehensive set of relevant features, effectively suppressing inter-class semantic confusion. Based on this, we further propose Dep-MAP, which employs a semantic prototype-supervised contrastive learning method, to align the latent semantics between shallow and deep features, enabling emotional semantic guidance and self-distillation within the visual feature branch, thereby effectively suppressing intra-class difference. Finally, Dep-MAP integrates discriminative cues across multiple spatiotemporal scales via a wavelet-based multi-scale fusion strategy and spatiotemporal context-weighted module, for automatic depression assessment. The main contributions and innovations of this paper are as follows:

- We propose a semantic-guided prototype clustering method that maps facial behavior features to a set of semantically diverse prototypes, to model discriminative facial expressions corresponding to different depression levels, enhancing the discriminability and semantic consistency of the prototypes, without the need for explicit category labels or handcrafted semantic groupings.
- We propose a cross-layer contrastive learning strategy supervised by semantic prototypes, which aligns the latent semantics between shallow visual features and deep semantic representations, guiding the visual branch to learn depression-related emotional semantics, and facilitating semantic-level self-distillation.
- We reveal a key limitation of existing ADA methods: the facial behaviors of depressed patients exhibit both inter-class differences and intra-class similarities with those of healthy individuals. To address this, we propose Dep-MAP, the first ADA network that integrates multi-level semantic alignment and implicit key frame perception. The network suppresses inter-class differences and intra-class similarities in the semantic space and achieves state-of-the-art performance on benchmark datasets.

Related Work

Facial behavior based ADA: Human facial behavior serves as an effective biomarker for depression (Kupfer, Frank, and Phillips 2012; Ye et al. 2023). Several studies (Niu et al. 2022b; Shang et al. 2021) have explored dynamic facial behavior features of individuals with different depression severity levels using deep learning models. These methods often extract static features from individual frames or facial regions, and aggregate all frame-level features or predictions, to perform automatic depression assessment. However, such approaches often overlook the emotional uncertainty inherent in facial behaviors. Zhou et al. (Zhou et al. 2020) proposed modeling the ordinal relationship between depression scores and visual features through label distribution learning and metric learning. Another approach involves regularization-based methods, where He et al. (He et al. 2022) proposed a squared ranking regularization module to handle noisy labels in automatic depression estimation. To identify sparse yet crucial depression-related cues from temporal context, Shangguan et al. (Shangguan et al. 2022) proposed a weakly supervised learning approach, employing multiple instance learning with max pooling to capture the most representative depressive cues. Ye et al. (Ye et al. 2024) proposed a sparse attention encoder, which adaptively extracts key depressive cues from sequential context, for automatic depression recognition.

Expression semantic learning: Expression semantics consist of diverse muscle deformations and texture features, which contribute significantly to facial behavior analysis. Yang et al. (Yang et al. 2021) noted that each action unit encodes specific semantics associated with facial expressions, encompassing the variation patterns of expression features (Vemulapalli and Agarwala 2019), and their intrinsic relationships are crucial for guiding feature representation learning (Li et al. 2019). In the ADA field, semantic information is commonly extracted and modeled via emotion map-

ping (Parikh, Sadeghi, and Eskofier 2024), visual semantic tokens (Ray et al. 2019), and attention-driven methods (Fu et al. 2025), to enhance feature representation ability. Thus, latent semantics derived from facial action perception can accurately capture expression similarities and differences (Ruan et al. 2021), thereby enhancing model sensitivity to micro-expression variations and atypical representations, significantly boosting robustness in depression assessment. However, these latent semantic features remain challenged by several issues, mainly stemming from the complexity and ambiguity of semantic representations. To address this, we explore a semantic-guided prototype clustering method and a semantic prototype-supervised cross-layer contrastive learning methods, simultaneously suppressing inter-class similarity and intra-class difference in the semantic space, enhancing the discriminative ability of depression-related facial behaviors and the generalization of the model.

Proposed Method

Overview: As shown in Figure 2, Given a facial video sequence. Initially, visual encoder extracts spatial-scale facial feature maps from each frame and constructs the 2D feature set $F = \{f_1, f_2, \cdots, f_T\}$ representing the entire video. Subsequently, we construct an emotional semantic mapping space using these frame-level feature maps of hashing coding mechanism. Specifically, we employ the CLIP (Radford et al. 2021) to extract textual semantic embeddings T as query and visual semantic embeddings V as key in the attention mechanism; while the visual features f_t serve as the value, using channel attention mechanism to facilitate cross-modal information fusion and semantic information guidance. As a result, we obtain multi-level emotional semantic features, which explicitly model the emotional semantic distribution of each frame in video, as follows:

$$E_t = \gamma \sum_{i=1}^{\mathcal{C}} \left(\frac{exp(T_i, V_i) \cdot f_{t,i}}{\sum_{i=1}^{\mathcal{C}} exp(T_i, V_i)}\right) + f_t, \tag{1}$$

here, γ is a learnable parameter that maps the CLIP semantic embeddings to the target depression assessment task, $exp(T_i, V_i)$ measures the similarity between textual semantic embeddings and visual semantic embeddings on the i-th channel and C denotes the total number of channels.

Subsequently, we perform semantic disambiguation prototype clustering module(SDPM), dynamically updating prototypes for different depression levels based on the empirical means of each cluster, resulting in a semantic prototype set $P_A^K \in \mathbb{R}^{A \times K \times L}$, which includes A classes corresponding to various depression severity levels, where each prototype P_a^K represents distinctive facial cues specific to that class, effectively suppressing inter-class similarity, as illustrated below:

$$P_a^K \leftarrow P_a^K + (1 - \beta)\overline{F}. \tag{2}$$

Based on the set of semantically distinct prototypes, we calculate the cosine similarity between each frame semantic feature f_t and its assigned prototype to obtain logits representing frame-level semantic consistency, and implicitly identify potential key frames within the temporal sequence based on these logits, thereby effectively avoiding interference from low-confidence frames in the final prediction. Subsequently, multi-scale representation learning module integrates discriminative depression cues from all key frames across multiple spatiotemporal scales, to achieve depression assessment.

During training, we employ the Mean Squared Error loss to compare predictions y_{pred} with ground truth y, while using the contrastive learning loss from KDM to achieve multi-level latent semantic alignment at the semantic level. This is formulated as follows:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \|y_i - y_i^{\text{pred}}\|_2^2 + \lambda \sum_{t=1}^{T} \sum_{j=1}^{J} \mathcal{L}_{j,t}^{\text{cl}}, \tag{3}$$

here, $||.||_2$ denotes L2 norm; J is the number of multi-scale features and λ controls the strength of the contrastive term.

Semantic Disambiguation Prototype Cluster

As shown in Figure 2(b), to alleviate the inter-class semantic confusion in facial behaviors within videos, SDPM aims to extract the most discriminative facial behavior cues from frame-level features, and construct a set of semantically distinct prototypes, to characterize typical behavioral expressions corresponding to varying depression severities.

Initially, we apply convolution and average pooling to the feature map E_t of the t-th frame derived from CLIP-guided deep semantic features, then introduce a hashing encoding mechanism to localize key response areas, resulting in latent facial behavior features:

$$G_t = \text{FC}\left(\text{HashCoding}\left(\text{Pool}\left(\text{Conv}\left(E_t\right)\right)\right)\right), \tag{4}$$

here, G_t denotes the prototype vectors mapped from AU-related key cues in each channel of E_t. To improve prototype quality, we introduce a sparsity-driven important feature selection mechanism that extracts discriminative facial-behavior embeddings from the contextual set $\mathcal{G}$. We measure sparsity using the Kullback–Leibler divergence between a key-feature distribution $p_a(\mathcal{G}_i \mid G_t) = \dfrac{k(G_t, \mathcal{G}_i)}{\sum_l k(G_t, \mathcal{G}_l)}$ and a uniform distribution $p_u(\mathcal{G}_i \mid G_t) = \dfrac{1}{T}$. Let the sparse cue probability distribution computed as:

$$M\left(G_t, \mathcal{G}\right) = \ln \sum_{i=1}^{T} e^{\frac{G_t \mathcal{G}_i^\top}{\sqrt{d}}} - \frac{1}{T} \sum_{i=1}^{T} \frac{G_t \mathcal{G}_i^\top}{\sqrt{d}}, \tag{5}$$

here, the first term is the Log-Sum-Exp operation over all channels, while the second term is its arithmetic mean. For each semantic feature G_t in the contextual set $\mathcal{G}$, the range of the function $M(G_t, \mathcal{G})$ is defined as: $\ln T \leq M\left(G_t, \mathcal{G}\right) \leq \max_i \left\{G_t \mathcal{G}_i^\top / \sqrt{d}\right\} - \frac{1}{T} \sum_{i=1}^{T} \left\{G_t \mathcal{G}_i^\top / \sqrt{d}\right\} + \ln T$. Therefore, the maximum mean measurement $\bar{M}\left(G_t, \mathcal{G}\right)$ can be defined as follows:

$$\bar{M}\left(G_t, \mathcal{G}\right) = \max_i \left\{\frac{G_t \mathcal{G}_i^\top}{\sqrt{d}}\right\} - \frac{1}{T} \sum_{i=1}^{T} \frac{G_t \mathcal{G}_i^\top}{\sqrt{d}}. \tag{6}$$

Next, we perform the prototype update only on the top-u largest measures $\bar{M}\left(G_t, \mathcal{G}\right)$, where $u = c \cdot \ln T$ and c is

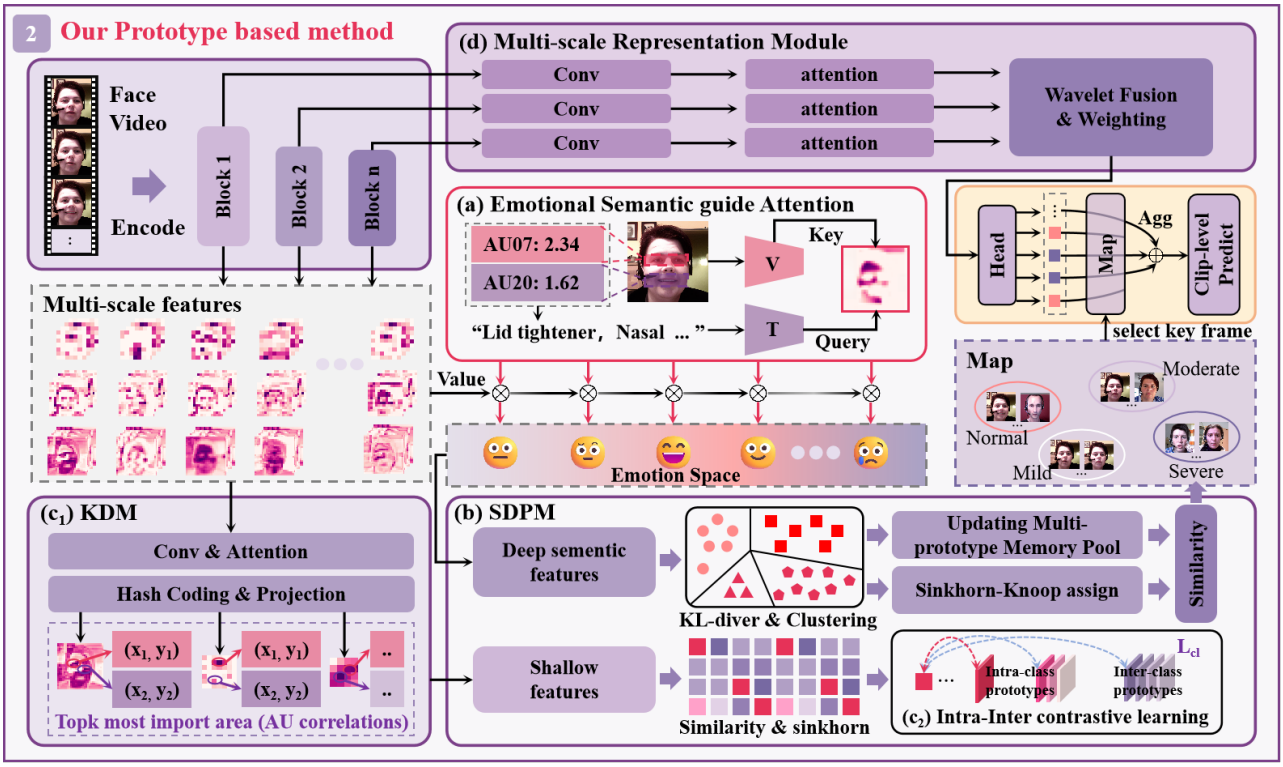

Figure 2: Overview of Dep-MAP: The backbone network extracts frame-level 2D facial feature maps across multi-scale; (1) Based on the emotion semantic embeddings captured by ESA (a), SDPM (b) clusters their deep semantic features, and uses sparsity to capture key emotion features to optimize prototype representations. (2) KDM (c) aligns latent semantics between shallow and deep features. (3) MSR (d) fuses multi-scale spatiotemporal features to achieve video-level depression assessment.

a constant sampling factor, to build the key facial-behavior feature set $\mathcal{F}_a \in \mathbb{R}^{N\times L}$. Features in $\mathcal{F}_a$ show high within-class consistency and between-class separation, so they are suitable for prototype updates. Specifically, for each depression level, we compute the cosine similarity between $\mathcal{F}_a$ and its prototype set $\{P_a^k\}_{k\in[1,K]}$ to obtain the similarity matrix $S_a \in \mathcal{R}^{N\times K}$, where K is the number of prototypes per class. Then we perform clustering assignment on the similarity matrix to obtain the assignment matrix $\mathcal{A} \in \{0,1\}^{N\times K}$, which indicates the prototype assigned to each feature vector. In the following, our goal is to optimize the assignment matrix A_a for class a to maximize the similarity between each key facial-behavior feature and its assigned prototype. Formally expressed as:

$$\max_{A_a\in\mathcal{A}_a} Tr((A_a)^\top S_a), \tag{7}$$

here, $Tr(\cdot)$ represents the trace of a matrix. To prevent degenerate case where all feature maps are assigned to the same prototype in A_a, and to ensure that each prototype covers semantically distinct AU response patterns, we impose unique and balanced assignment constraints on A_a, specified as follows:

$$(A_a)^\top \mathbf{1}_N = \frac{1}{K}\mathbf{1}_K, A_a\mathbf{1}_K = \frac{1}{N}\mathbf{1}_N. \tag{8}$$

These constraints ensure that each key facial behavior feature is mapped to only one prototype of class a, while the number of features assigned to each prototype is approximately equal. This assignment mechanism can be formulated as an optimal transport problem, aiming to establish a cost-minimizing and balanced mapping between key AU response features and prototypes. To solve this problem, we construct a smoothed objective function to approximate the solution for $\mathcal{A}$. The resulting smoothed optimal transport formulation is as follows:

$$A_a^* = \max_{A_a\in\mathcal{A}_a} Tr((A_a)^\top S) + \kappa h(A_a), \tag{9}$$

here, $\kappa h(A_a)$ denotes the entropy regularization term that controls the smoothness of the assignment matrix $\mathcal{A}$. We set the regularization coefficient κ to 0.05. To efficiently solve this entropy-regularized optimal transport problem, we adopt the Sinkhorn-Knopp algorithm:

$$A_a^* = \text{diag}(\mathbf{u})\exp\left(\frac{S_a}{\kappa}\right)\text{diag}(\mathbf{v}). \tag{10}$$

After assignment, each prototype P_k^a is updated via momentum using the mean of the embeddings assigned to it:

$$P_a^k \leftarrow P_a^k + (1-\beta)\overline{\mathcal{F}}_a^k, \tag{11}$$

here, $\beta \in (0,1)$ is a coefficient that controls the update rate of the prototypes. This approach ensures that every latent facial behavior feature $\mathcal{F}_a$ contributes to its corresponding

prototype, ensuring that each class prototype representation covers a broad range of behavioral concepts. This learning scheme effectively captures the average embedding representations of salient AU-related cues in the training data. In Section V, we will provide a detailed analysis of key AU-related facial behaviors by mapping prototypes to the nearest AU cues in the training set.

Key Facial Behavior Self-Distillation

In the second stage, we propose Key Facial Behavior Self-Distillation Module. the core idea is prototype-guided cross-scale alignment, which ensures semantic consistency between shallow and deep features and assigns the appropriate class attributes to the shallow representations.

As shown in Figure 2(c), we learn a set of prototypes P_a^k to represent AU-related facial behavior features under class a. For the deep semantic feature of a single facial frame, the activation value of each prototype P_a^k is calculated as:

$$g_a^k = \max_{\widetilde{f} \in \mathcal{F}_a^k} \quad \mathrm{Sim}\left(\widetilde{f}, P_a^k\right) \in \mathbb{R}^{A \times K}. \tag{12}$$

Since not all prototypes contribute equally to the input image, we generate the logit for class a by weighting and averaging the activation values g_a of each prototype, according to the following formula:

$$logit_a = \sum_{k=1}^{K} w_a^k \cdot g_a^k, \tag{13}$$

here, w is a learnable weighting vector that controls how each prototype activation contributes to its class logit. Based on $logit_a$ at each time frame in the video sequence, we propose a MAP module to select high-confidence decision features from the multi-scale fusion results and aggregate them over time. Specifically, we select the time frames with the largest logit proportion from prototype activation to improve the stability of the final decision.

To reinforce local consistency across multi-scale features, we further introduce an intra- and inter-class prototype distance optimization, $\mathcal{L}_{(j,t)}^{cl}$ defined as follows:

$$-\log \frac{\exp\left(f_{j,t} \cdot P_a^k\right)}{\sum_{k' \in K \setminus k} \exp\left(f_{j,t} \cdot P_a^{k'}\right) + \sum_{a' \in A \setminus a} \exp\left(f_{j,t} \cdot P_{a'}^K\right)}, \tag{14}$$

here, $\sum_{k' \in K \setminus k}$ denotes all other prototypes in class a except the currently assigned prototype. This loss function "pulls" each feature map closer to its assigned prototype in the latent space, while "pushing" it away from other prototypes of the same class, and further increases the margin between it and prototypes from different classes. These distance measures complement each other, capturing the instance-specific features, while maintaining intra-class diversity, enhancing the discriminative ability of facial behavior representations. This contrastive loss is applied across multiscale features in Dep-MAP, with weighted summation of losses from each layer, to achieve cross-layer semantic alignment through joint optimization. Shallow visual features gradually approach deep semantic features under the supervision of semantic prototypes, realizing semantic-level key information self-distillation, enhancing the sensitivity of Dep-MAP to depression-related representations.

Multi-Scale Representation Learning

Although deep features generally capture semantic information more relevant for automatic depression assessment, shallow features tend to be more generalizable across different tasks, thus features from various layers complement each other spatially and semantically. To this end, we propose a wavelet transform-based attention weighting module. Specifically, we apply pointwise convolution to unify channel dimensions, and use grouped convolution to capture prototype-level coupling relationships between channels. Then, introducing a channel attention mechanism to adaptively assess depression cues within features at various scales. Meanwhile, we apply discrete wavelet transform to the low-level feature map, decomposing it into four frequency sub-bands LL, LH, HL, HH, where LL denotes the low-frequency component. Next, we fuse the small-scale feature map with LL to obtain a structure-enhanced multi-scale representation, and using inverse discrete wavelet transform to merge it with the high-frequency components LH, HL, HH for generating a unified multi-scale feature.

Experiment

Experiment Setting

We evaluate our method on two visual depression assessment benchmarks: AVEC2013(Valstar et al. 2013) and AVEC2014(Valstar et al. 2014). In both datasets, depression levels are labeled by the Beck Depression Inventory(Jackson-Koku 2016) with scores from 0 to 63. AVEC2013 contains 150 videos. AVEC2014 includes two recording tasks: Northwind and Freeform, containing 300 videos, with each subset containing 150 clips lasting from 6 seconds to 4 minutes 8 seconds. Both datasets are evenly split into training, development, and test sets. Following prior work(Wu et al. 2025), we compare methods using mean absolute error, root mean square error, Pearson correlation, and concordance correlation between predictions and ground truth. All Dep-MAP experiments ran on a workstation with an NVIDIA RTX 4090 GPU. We set 4 classes with 8 prototypes each, use a sampling factor of 5, a contrastive weight of 0.1, and Haar wavelet–based fusion.

Comparison with SOTA

Table 1 shows that our method significantly outperforms all existing competitors and achieves new state-of-the-art performance on both datasets. Compared to the previous SOTA methods(Wu et al. 2025), our method achieves relative RMSE improvements of 2.34% on AVEC2013 and 9.55% on AVEC2014. Compared with the recently proposed prototype-based ADA method(Li et al. 2025), our method achieves over 8.86% and 7.47% relative improvements in RMSE on the two datasets, highlighting the advantages of the proposed semantic disambiguation prototype clustering module and key facial behavior self-distillation module in

Methods	AVEC2013		AVEC2014	
	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
(Li, Qu, and Zhou 2025)	8.64	6.82	8.11	6.29
(Liu et al. 2023)	7.59	6.08	7.98	6.04
(Xu et al. 2024)	7.57	5.95	7.65	6.24
(Niu et al. 2022a)	7.49	6.12	7.56	6.01
(Uddin, Joolee, and Sohn 2022)	7.32	5.90	6.98	5.75
(Pan et al. 2023)	7.26	5.97	7.30	5.99
(Li et al. 2025)	7.78	5.82	7.69	5.77
(Fu et al. 2025)	-	-	6.80	5.26
(Wu et al. 2025)	7.26	5.38	6.28	4.99
Ours	**7.09**	**5.19**	**5.68**	**4.43**

Table 1: Comparison of RMSE and MAE on AVEC2013 and AVEC2014 datasets. ↓ indicates lower is better.

S1	S2	S3	S4	RMSE↓	MAE↓	PCC↑	CCC↑
✓				8.11	6.25	0.84	0.60
	✓			6.55	5.22	0.92	0.74
		✓		7.13	5.51	0.90	0.68
			✓	6.10	4.71	0.91	0.81
		✓	✓	6.16	**4.39**	0.85	**0.84**
	✓	✓	✓	6.75	5.44	0.90	0.73
✓	✓	✓	✓	**5.68**	4.43	**0.92**	0.83

Table 2: Results achieved for different spatial scales.

fine-grained emotion modeling and alignment of key facial behavior signals. On the AVEC2014 dataset, applying our method to separately model task-specific behaviors (i.e., Freeform and Northwind expressions) and incorporating facial behavior prototypes of potential depressed individuals further boosts generalization. This suggests that our Dep-MAP can flexibly adapt to different depression representations across varying contexts and demonstrates strong task transfer ability.

Ablation Study

Impact of Different Dep-MAP Components on Performance: Table 3 demonstrates that the semantically distinct prototype representations learned by our Dep-MAP effectively capture depression-related facial cues in videos, even when such cues are temporally sparse, unevenly distributed, and affected by significant intra- and inter-class similarity, resulting in notable performance improvements. Additionally, Figure 3 highlight the advantages of Dep-MAP in addressing intra-class difference and inter-class similarity. Specifically, Dep-MAP: (1) avoids the need to compress all frame- or segment-level facial features into pooled vectors for prototype updates; (2) can process facial behaviors of varying severity levels and semantic contexts in batches, while traditional methods struggle to differentiate such behaviors at a fine-grained semantic level. These findings indicate that conducting semantic prototype clustering and cross-layer contrastive alignment directly on semantically guided deep feature maps preserves richer spatial structural information and enables the extraction of more depression-related spatiotemporal facial cues.

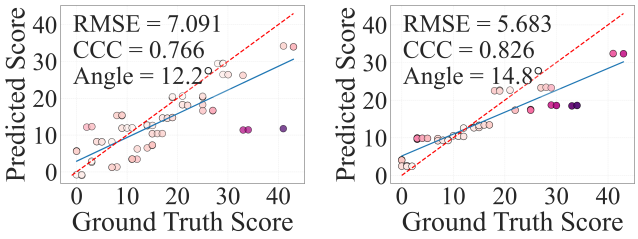

Figure 3: The prediction visualizations from (left) AVEC2013 and (right) AVEC2014.

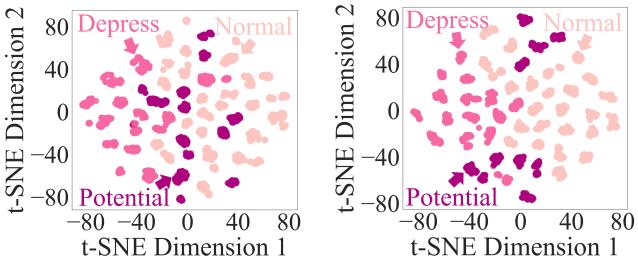

Figure 4: t-SNE visualizations of learned multi-scale fused features from (left) AVEC2013 and (right) AVEC2014.

Performance Comparison Under Different Parameter Configurations: Dep-MAP follows the ResNet architecture settings, specifically including ResNet-18, ResNet-34, and ResNet-50. As shown in Table 3, Dep-MAP-18 using ResNet-18 as the backbone achieves the best or second-best results on most tasks and metrics, demonstrating good computational efficiency, stability, and generalization ability. Dep-MAP-50 with ResNet-50 as the backbone performs better in the Freeform task, suggesting that model selection should be adjusted based on specific application scenarios.

Contribution of Different Spatial Scales: Table 2 compares the results obtained by applying Dep-MAP to process matrix maps at different spatial scales to encode depression-related facial behaviors. It is evident that considering more spatial scales of facial behavior is beneficial, as performance improves when more spatial scales are taken into account. Furthermore, we found that facial behaviors at each spatial scale can reflect depressive states to some extent (as evidenced by their high CCC performance), and they are complementary, as their combination outperforms the variants using only a single spatial scale. In addition, the configuration that fuses only the deepest two layers of visual features achieves the second-best performance, suggesting that the choice of fusion scales should be adjusted according to specific application scenarios.

Backbone	Branch		Module					Northwind				Freeform			
	Visual	Emotional	MSR	ESA	SDPM	KDM	MAP	RMSE ↓	MAE ↓	PCC ↑	CCC ↑	RMSE ↓	MAE ↓	PCC ↑	CCC ↑
ResNet-18	✓	-	-	-	-	-	-	10.32	7.77	0.60	0.50	9.96	7.44	0.63	0.52
ResNet-18	✓	-	✓	-	-	-	-	8.01	5.91	0.90	0.70	8.44	5.98	0.85	0.62
ResNet-18	✓	✓	✓	✓	-	-	-	7.79	5.82	0.89	0.72	8.47	6.06	0.83	0.61
ResNet-18	✓	✓	✓	✓	✓	-	✓	7.54	6.07	0.76	0.75	8.05	5.45	0.89	0.65
ResNet-18	✓	✓	✓	✓	✓	✓	-	6.62	5.80	0.92	0.79	7.09	5.35	0.87	0.77
ResNet-18	✓	✓	✓	✓	✓	✓	✓	6.25	5.25	0.95	0.79	7.42	5.60	0.91	0.69
ResNet-34	✓	✓	✓	✓	✓	✓	✓	6.84	5.66	0.95	0.70	8.73	6.71	0.90	0.47
ResNet-50	✓	✓	✓	✓	✓	✓	✓	8.15	6.55	0.91	0.55	7.08	5.52	0.91	0.69

Table 3: Ablation study results on Freeform and Northwind subsets.

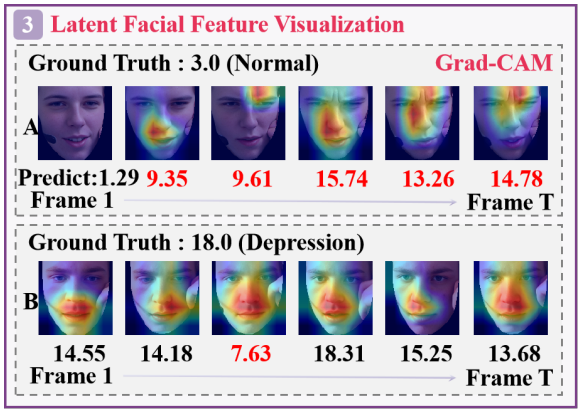

Figure 5: Latent feature visualization results.

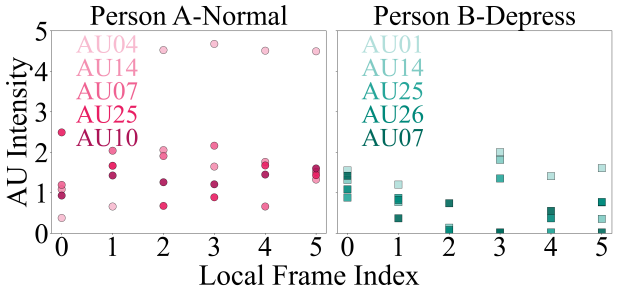

Figure 6: Trends in Action units.

Interpretability Analysis

We visualize the multi-scale spatiotemporal features after wavelet fusion using t-SNE, as shown in Figure 4. The main findings are as follows: (1) Compared to the AVEC2013 dataset, the multi-scale spatiotemporal features of AVEC2014 show more distinct separability in t-SNE space. This is because its specific emotion induction tasks can capture richer depression-related behavior patterns. (2) The feature distribution of potentially depressed patients is relatively sparse, whereas the features of healthy and depressed patients are more densely clustered. This suggests that the proposed SDPM and KDM modules effectively strengthen the capture and aggregation of highly discriminative behavioral patterns between depressive and normal states, enabling the identification of more critical facial expression features and thereby mitigating intra-class difference and inter-class similarity in depression assessment.

We further performed a visual analysis of the deep features guided by semantic prototypes, and the results are shown in Figure 5 and 6. Compared to normal facial behaviors, depression-related facial cues exhibit significantly higher activation in key AU regions and more spatially concentrated distributions. This indicates that the discriminative semantic prototypes learned through the SDPM and KDM effectively guide deep features to focus on depression-related facial regions, significantly enhancing the ability of Dep-MAP to capture subtle depression-related cues. Dep-MAP also shows sensitivity to transient depression-related micro-expressions in healthy individuals (e.g., occasional frowning, periocular muscle tension), which reflect the high correlation between such negative emotional expressions and depressive symptoms. Meanwhile, in depressed group, Dep-MAP still responds to relatively normal facial behavior segments. This suggests that the multi-frame logits aggregation based on prototype similarity can effectively suppress the interference from local atypical behavioral segments, dynamically identify key depressive behavior patterns that are highly consistent and discriminative in the semantic prototype space, thereby significantly improving the robustness of video-level depression assessment.

Conclusion

This study introduces a video-based automatic depression assessment method tailored to the complex facial behaviors of depressed patients. The proposed Dep-MAP outperforms traditional approaches by suppressing inter-class similarity and intra-class variation in the semantic space. This effectively narrows the gap between visual features and the emotional semantics of facial expressions, establishing Dep-MAP as a state-of-the-art solution for video-based depression assessment. However, this performance gain comes at the cost of losing some local information. Future work will explore advanced interpretability techniques to uncover the underlying contextual cues.

Acknowledgements

This work was funded by project ZR2025MS1079 supported by Shandong Provincial Natural Science Foundation; and in part by the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, under Grant 2024PY035.

References

Di, Y.; Wang, J.; Li, W.; and Zhu, T. 2021. Using i-vectors from voice features to identify major depressive disorder. *Journal of Affective Disorders*, 288: 161–166.

Fu, C.; Qian, F.; Su, Y.; Su, K.; Song, S.; Niu, M.; Shi, J.; Liu, Z.; Liu, C.; Ishi, C. T.; et al. 2025. Facial action units guided graph representation learning for multimodal depression detection. *Neurocomputing*, 619: 129106.

He, L.; Tiwari, P.; Lv, C.; Wu, W.; and Guo, L. 2022. Reducing noisy annotations for depression estimation from facial images. *Neural Networks*, 153: 120–129.

Hook, A.; and Andrews, B. 2005. The relationship of nondisclosure in therapy to shame and depression. *British Journal of Clinical Psychology*, 44(3): 425–438.

Jackson-Koku, G. 2016. Beck depression inventory. *Occupational medicine*, 66(2): 174–175.

Kupfer, D. J.; Frank, E.; and Phillips, M. L. 2012. Major depressive disorder: new clinical, neurobiological, and treatment perspectives. *The Lancet*, 379(9820): 1045–1055.

Li, G.; Zhu, X.; Zeng, Y.; Wang, Q.; and Lin, L. 2019. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8594–8601.

Li, Y.; Qu, S.; and Zhou, X. 2025. Conformal depression prediction. *IEEE Transactions on Affective Computing*.

Li, Y.; Wei, Z.; Guo, G.; and Zhou, X. 2025. Memrank: Memory-augmented similarity ranking for video-based depression severity estimation. *IEEE Transactions on Affective Computing*.

Liu, Z.; Yuan, X.; Li, Y.; Shangguan, Z.; Zhou, L.; and Hu, B. 2023. PRA-Net: Part-and-Relation Attention Network for depression recognition from facial expression. *Computers in biology and medicine*, 157: 106589.

Niu, M.; He, L.; Li, Y.; and Liu, B. 2022a. Depressioner: Facial dynamic representation for automatic depression level prediction. *Expert Systems with Applications*, 204: 117512.

Niu, M.; Li, Y.; Tao, J.; Zhou, X.; and Schuller, B. W. 2024. Depressionmlp: A multi-layer perceptron architecture for automatic depression level prediction via facial keypoints and action units. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9): 8924–8938.

Niu, M.; Zhao, Z.; Tao, J.; Li, Y.; and Schuller, B. W. 2022b. Selective element and two orders vectorization networks for automatic depression severity diagnosis via facial changes. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 8065–8077.

Nowicki Jr, S.; and Duke, M. P. 1994. Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal behavior*, 18(1): 9–35.

Pampouchidou, A.; Simantiraki, O.; Fazlollahi, A.; Pediaditis, M.; Manousos, D.; Roniotis, A.; Giannakakis, G.; Meriaudeau, F.; Simos, P.; Marias, K.; et al. 2016. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 27–34.

Pan, Y.; Shang, Y.; Shao, Z.; Liu, T.; Guo, G.; and Ding, H. 2023. Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. *IEEE Transactions on Affective Computing*.

Parikh, A.; Sadeghi, M.; and Eskofier, B. 2024. Exploring facial biomarkers for depression through temporal analysis of action units. *arXiv preprint arXiv:2407.13753*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ray, A.; Kumar, S.; Reddy, R.; Mukherjee, P.; and Garg, R. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, 81–88.

Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; and Wang, H. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7660–7669.

Shang, Y.; Pan, Y.; Jiang, X.; Shao, Z.; Guo, G.; Liu, T.; and Ding, H. 2021. LQGDNet: A local quaternion and global deep network for facial depression recognition. *IEEE transactions on affective computing*, 14(3): 2557–2563.

Shangguan, Z.; Liu, Z.; Li, G.; Chen, Q.; Ding, Z.; and Hu, B. 2022. Dual-stream multiple instance learning for depression detection with facial expression videos. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 554–563.

Uddin, M. A.; Joolee, J. B.; and Sohn, K.-A. 2022. Deep multi-modal network based automated depression severity estimation. *IEEE transactions on affective computing*, 14(3): 2153–2167.

Valstar, M.; Schuller, B.; Smith, K.; Almaev, T.; Eyben, F.; Krajewski, J.; Cowie, R.; and Pantic, M. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 3–10.

Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; and Pantic, M. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 3–10.

Vemulapalli, R.; and Agarwala, A. 2019. A compact embedding for facial expression similarity. In *proceedings of the*

IEEE/cvf conference on computer vision and pattern recognition, 5683–5692.

Wen, L.; Li, X.; Guo, G.; and Zhu, Y. 2015. Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Transactions on Information Forensics and Security*, 10(7): 1432–1441.

Wu, Z.; Zhou, L.; Li, S.; Fu, C.; Lu, J.; Han, J.; Zhang, Y.; Zhao, Z.; and Song, S. 2025. DepMGNN: Matrixial Graph Neural Network for Video-based Automatic Depression Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1610–1619.

Xu, J.; Gunes, H.; Kusumam, K.; Valstar, M.; and Song, S. 2024. Two-stage temporal modelling framework for video-based depression recognition using graph representation. *IEEE Transactions on Affective Computing*, 16(1): 161–178.

Yang, H.; Yin, L.; Zhou, Y.; and Gu, J. 2021. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10482–10491.

Ye, J.; Yu, Y.; Fu, G.; Zheng, Y.; Liu, Y.; Zhu, Y.; and Wang, Q. 2023. Analysis and recognition of voluntary facial expression mimicry based on depressed patients. *IEEE Journal of Biomedical and Health Informatics*, 27(8): 3698–3709.

Ye, J.; Yu, Y.; Lu, L.; Wang, H.; Zheng, Y.; Liu, Y.; and Wang, Q. 2024. DEP-former: Multimodal depression recognition based on facial expressions and audio features via emotional changes. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhou, X.; Wei, Z.; Xu, M.; Qu, S.; and Guo, G. 2020. Facial depression recognition by deep joint label distribution and metric learning. *IEEE Transactions on Affective Computing*, 13(3): 1605–1618.