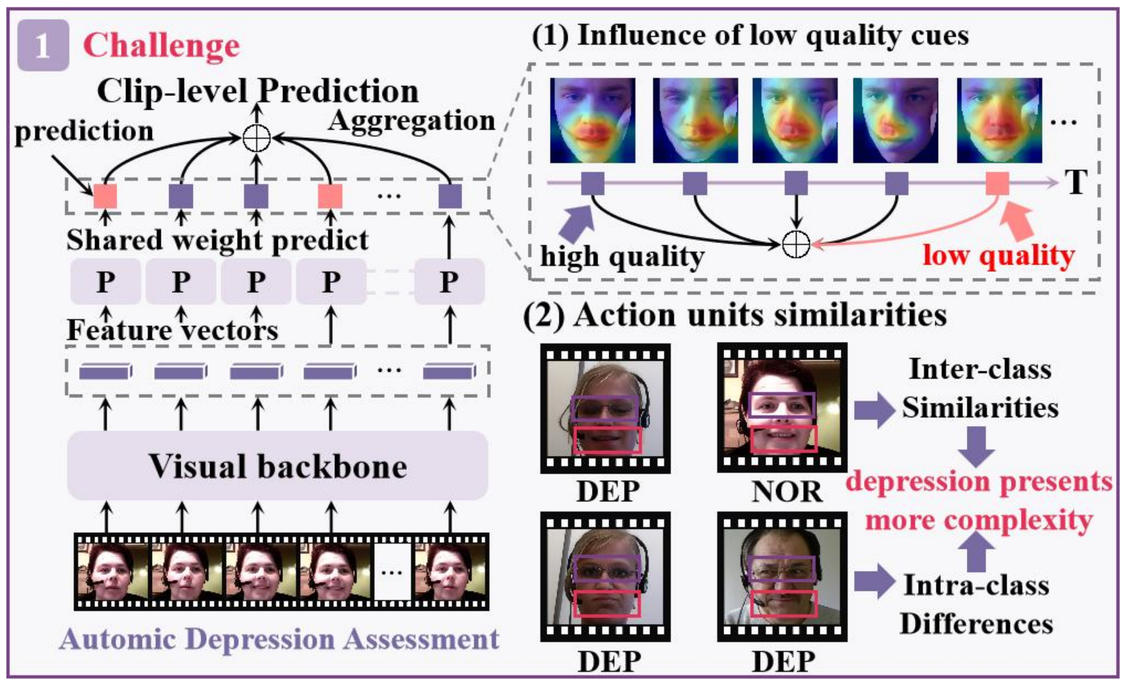
Dep-MAP: A Multi-level Alignment Framework with Semantic Prototypes for Video-based Automatic Depression Assessment

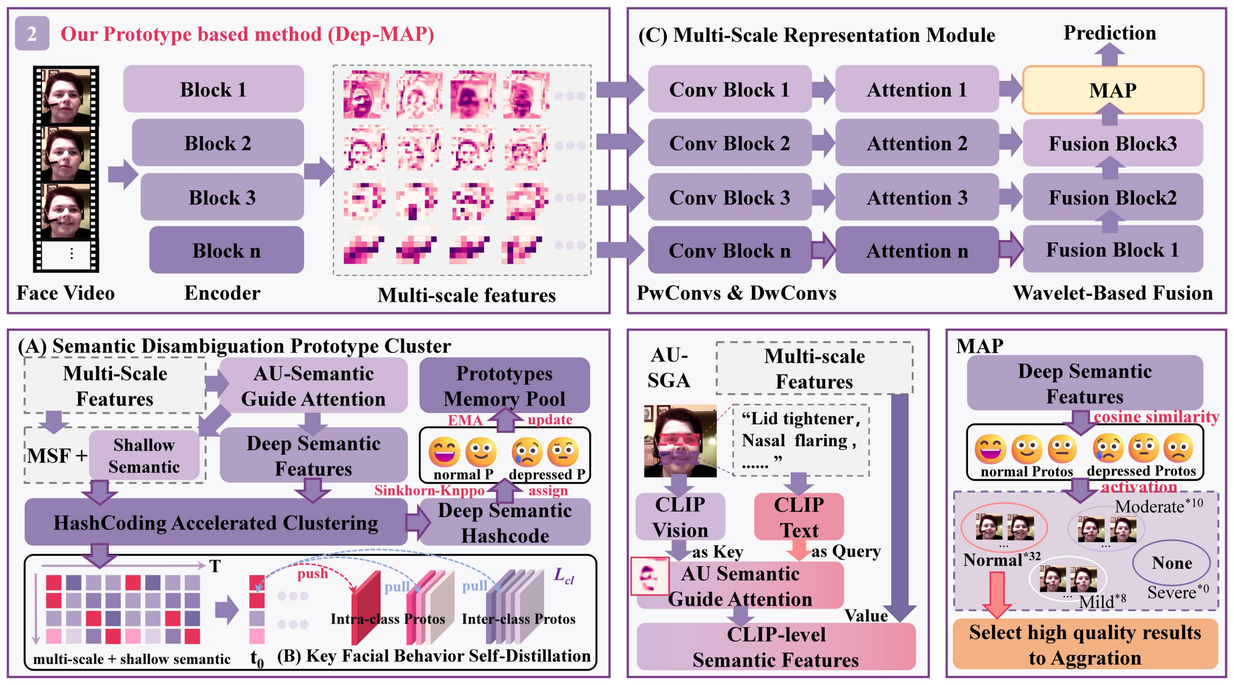
Hao Wang[1], Jiayu Ye[2] and Qingxiang Wang[1,*]
[1] Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2] Guangdong University of Technology, Guangzhou, China

Introduction & Motivation



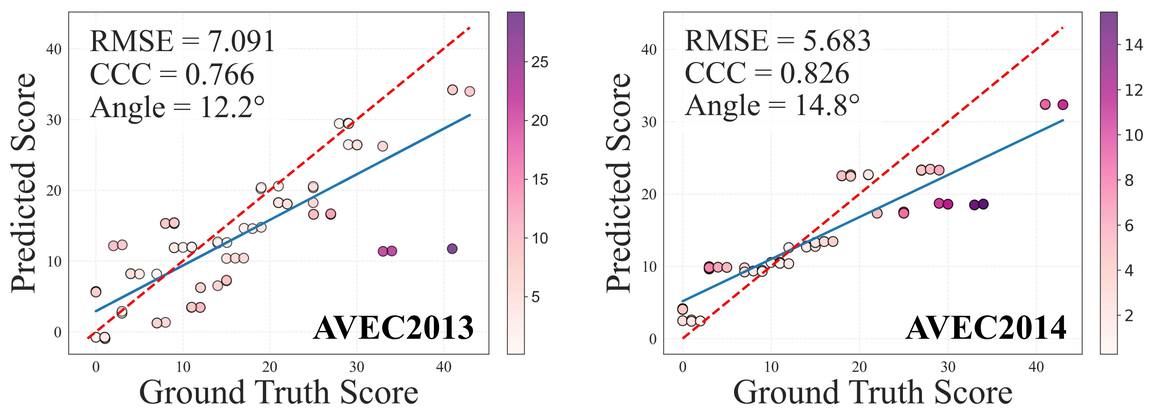
- **Background & Challenge:** Analyzing how facial expressions change over time is an important way to assess the mental state of depressed patients. However, in real life, many patients try to hide their symptoms and may look similar to healthy people, and those with different levels of severity can also show different facial behaviors, which makes depression assessment more difficult.
- This paper proposes **Dep-MAP**, a model that can automatically assess depression from videos by finding reliable key spatiotemporal patterns in complex facial behavior.

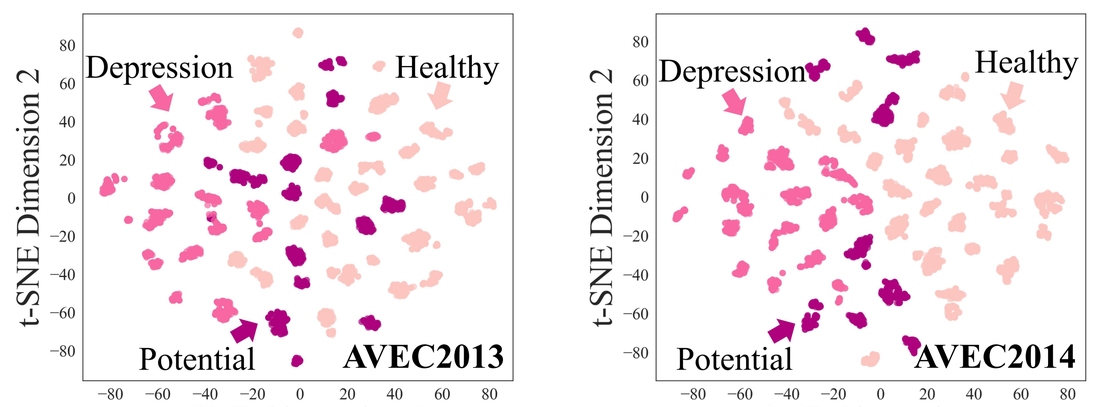
Structure of Dep-MAP

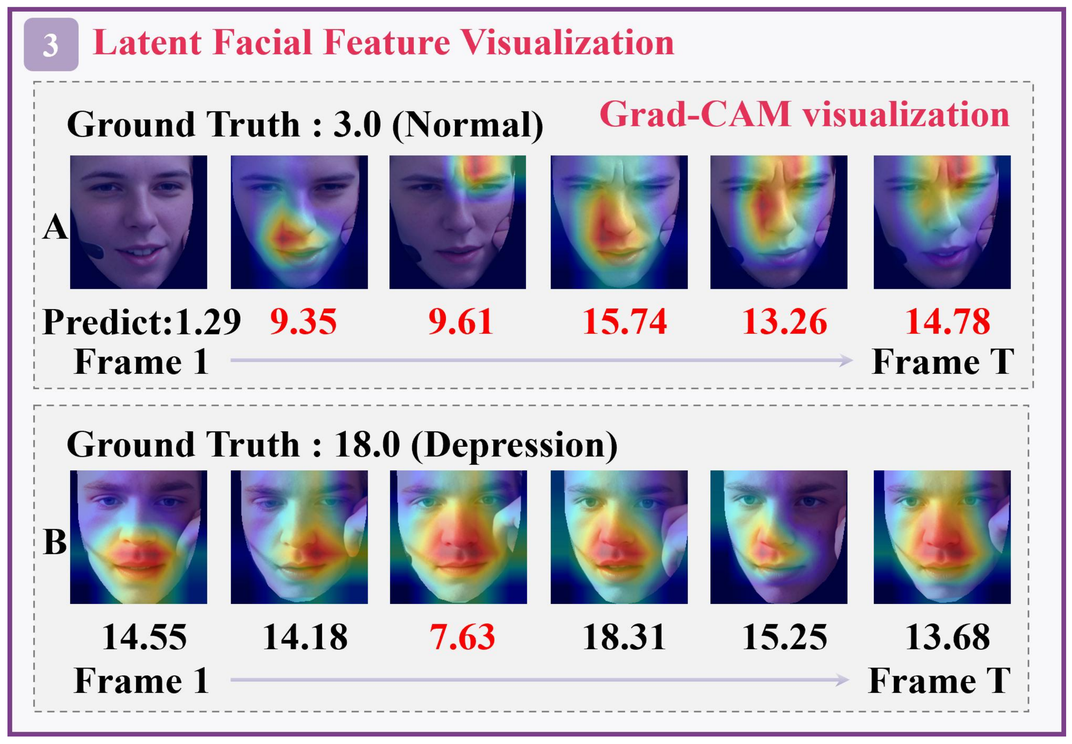


- We propose Dep-MAP, which has three main modules: (1) SDPM clusters deep emotion features into sparse, clear prototypes. (2) KBM aligns the semantics of shallow and deep features. (3) MSR fuses multi-scale spatiotemporal features to give a video-level depression score.
- First, Dep-MAP uses a **semantic prototype clustering module** to group facial features into clearly different pattern types, so it can describe how facial behavior changes with different levels of depression.
- Then, it adds a **cross-scale semantic alignment loss** that makes features from the same group closer and better controlled by semantics, which helps the Dep-MAP to detect very subtle facial changes in patients.

Visualizations







Experimental Results

- **Contributions:** Experiments show that **Dep-MAP can effectively detect important facial behaviors** that are hidden in the video context. By grouping and combining key frames with similar meaning, it achieves clearly better, state-of-the-art performance on the public AVEC2013 and AVEC2014 datasets.

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.

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.