Hierarchical Granularity Alignment and State Space Modeling for Robust Multimodal AU Detection in the Wild

Jun Yu, Yunxiang Zhang, Naixiang Zheng, Lingsi Zhu, Guoyuan Wang
University of Science and Technology of China

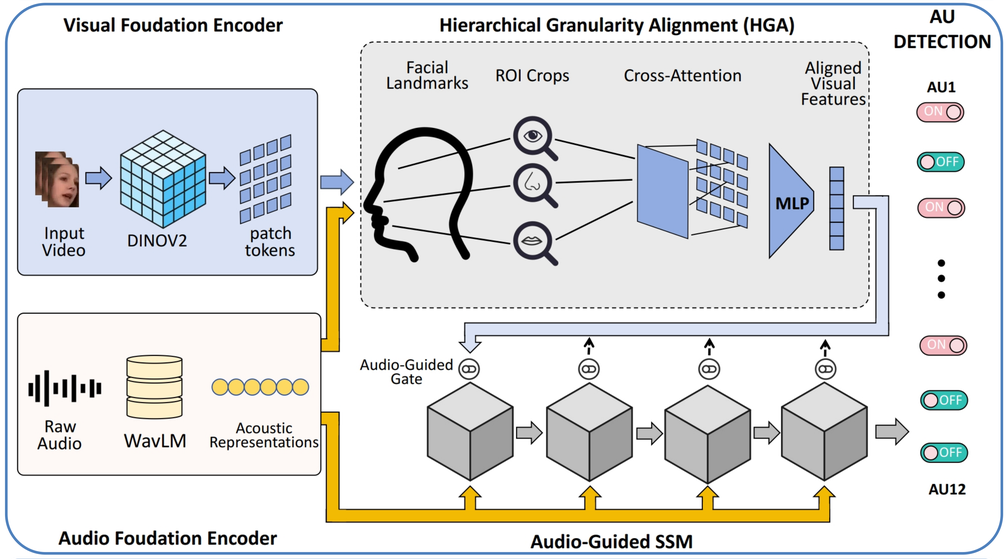
Problem & Motivation:

- Subtle AU activations are easily affected by pose, illumination, and occlusion.
- Local facial muscle movements require fine-grained spatial modeling.
- Long videos require efficient long-range temporal reasoning.
- Audio and visual cues are correlated but often fused shallowly.

Main Contributions:

- We use DINOv2 and WavLM to extract robust visual and audio representations.
- We propose Hierarchical Granularity Alignment (HGA) to align local AU regions with global facial context.
- We introduce Audio-Guided SSM (AG-SSM) for efficient long-range audio-visual temporal modeling.

Overall Framework



Method Details

Hierarchical Granularity Alignment:

- Facial landmarks define AU-related local regions.
- Local RoI features query global DINOv2 patch tokens.
- Cross-attention aligns local muscle activations with global facial semantics.

Hierarchical Granularity Alignment:

- Facial landmarks define AU-related local regions.
- Local RoI features query global DINOv2 patch tokens.
- Cross-attention aligns local muscle activations with global facial semantics.

Ablation Results on Validation Set:

Architecture Setup	FM	HGA	AG-SSM	ASL	F1 Score (%)
Baseline					36.50
+ Foundation Models	✓				52.41
+ HGA Module	✓	✓			55.18
+ AG-SSM Temporal Modeling	✓	✓	✓		57.82
Full Framework (+ ASL)	✓	✓	✓	✓	**59.45**

Key findings :

- Foundation models provide the largest performance gain.
- HGA improves robustness to local facial variations.
- AG-SSM improves temporal modeling for long in-the-wild videos.

Conclusion :

We present a multimodal AU detection framework combining foundation models, hierarchical spatial alignment, and audio-guided state space modeling. The method achieves 59.45% F1 on Aff-Wild2 validation and ranks 1st in the ABAW10 AU Detection Track.