2nd INTERNATIONAL CONFERENCE ON
Artificial Intelligence
31 May - 1 June 2026

IICAI
Second International Interdisciplinary Conference on Artificial Intelligence

Enhancing Facial Emotion Recognition Using YOLO11 Classification on the AffectNet Dataset

Reza Nasiri*, Seyed Enayat Alavi, Mohammad Javad Rashti
Department of Computer Engineering, Shahid Chamran University of Ahvaz, Iran

*Reza.Nasiri@stu.scu.ac.ir

Abstract

Facial Expression Recognition (FER) is a key component of affective computing that enables intelligent systems to understand human emotions in real-world environments. Despite recent advances, FER performance on large-scale in-the-wild datasets such as AffectNet remains challenging due to pose variation, illumination changes, occlusion, and annotation noise. In this study, a high-capacity YOLO11X-CLS architecture is employed for FER without using handcrafted modules such as facial alignment or landmark extraction. The proposed framework achieves 79.50% Top-1 accuracy on AffectNet-7, outperforming the previous state-of-the-art by 10.81%. Experimental results demonstrate strong robustness against real-world challenges and significant improvements across all emotion categories. The findings highlight the effectiveness of capacity-driven deep architectures for large-scale FER tasks.

Introduction

Facial Expression Recognition (FER) plays an important role in affective computing and intelligent human–computer interaction systems. FER applications include healthcare monitoring, social robotics, driver assistance systems, and intelligent tutoring platforms. However, recognizing emotions in unconstrained environments remains difficult due to variations in lighting, pose, occlusion, and facial diversity.AffectNet is one of the most challenging FER datasets because it contains large-scale real-world facial images with significant intra-class variability and inter-class similarity. Previous FER methods relied heavily on attention mechanisms, transformer architectures, and landmark-guided preprocessing. However, most approaches achieved less than 70% accuracy on AffectNet-7.This study investigates whether increasing model capacity alone can improve FER performance without requiring handcrafted facial analysis modules.

Proposed Method

The proposed Facial Expression Recognition (FER) framework is based on the YOLO11X-CLS architecture, a high-capacity convolutional classification model containing 28.3 million parameters. Instead, the framework performs end-to-end feature learning directly from facial images.The model was trained and evaluated on the AffectNet-7 dataset, which contains seven emotion categories: Neutral, Happy, Sad, Surprise, Fear, Disgust, and Anger. The dataset was reorganized into 70% training, 20% validation, and 10% testing subsets to ensure balanced evaluation.Training was performed using the AdamW optimizer with a batch size of 32, an initial learning rate of 0.01, and 50 training epochs. The high representational capacity and enlarged receptive field of YOLO11X-CLS enable the model to capture both local facial details and global semantic structures, resulting in robust performance under pose variation, occlusion, illumination changes, and class imbalance.

Experimental Results

Fig. 1. Training loss and Top-1 validation accuracy over 50 epochs on AffectNet-7

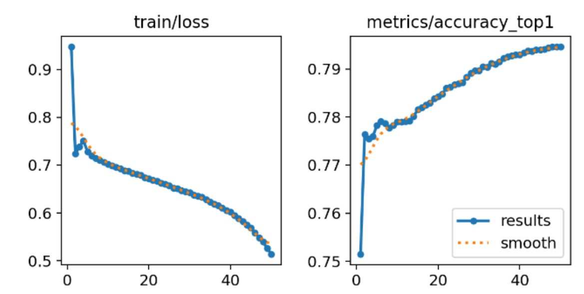

Table II. Comparison Of Accuracy With Sate-Of-The-ART FER Methods on AFFECTNET-7

Method	Accuracy (%)	Ref.
IPA2LT	57.31	[15]
RAN	59.50	[19]
SCN	60.23	[13]
RUL	61.43	[14]
DMUE	62.84	[16]
ARM	65.20	[17]
DAN	65.69	[18]
EAC	65.32	[20]
POSTER	66.42	[5]
S2D	66.42	[9]
TransFER	66.23	[7]
Poster++	67.49	[6]
NKF	68.17	[10]
NorFace	68.69	[2]
YOLO11X-CLS (Our)	**79.50**	—

Discussion

The experimental results demonstrate that the proposed YOLO11X-CLS framework significantly improves Facial Expression Recognition performance on the AffectNet-7 benchmark. The model achieved 79.50% Top-1 accuracy and surpassed the strongest previously reported method by 10.81%. The deep convolutional hierarchy captures both local facial micro-patterns and global semantic facial structures without relying on handcrafted preprocessing modules such as facial alignment or landmark-guided attention. The normalized confusion matrix confirmed strong performance across all emotion categories, especially for Happy and Neutral classes, while also improving difficult categories such as Fear and Disgust. The results indicate that architectural depth, expanded receptive fields, and representational capacity play a major role in improving class separability and robustness under pose variation, illumination inconsistency, occlusion, and class imbalance in large-scale in-the-wild FER datasets.

References

[1]A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal in the Wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2019.

[2]H. Liu et al.,"NorFace: Improving Facial Expression Analysis by Identity Normalization,"Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol. 15113, 2025.

[3]R. Khanam and M. Hussain,"YOLOv11: An Overview of the Key Architectural Enhancements,"arXiv preprint arXiv:2410.17725, Oct. 2024.

[4]B. Hasani and M. H. Mahoor,"Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields,"roc. IEEE FG, 2017, pp. 790–795.

[5]C. Zheng, M. Mendieta, and C. Chen,"POSTER: Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition,"
Proc. IEEE/CVF ICCV Workshops, 2023, pp. 3146–3155.

[6]J. Mao et al.,"Poster++: A Simpler and Stronger Facial Expression Recognition Network,"Pattern Recognition, vol. 148, 110951, 2024.

[7]A. Farzaneh and X. Deng,"Facial Expression Recognition in the Wild via Deep Attentive Center Loss,"IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 1082–1093, 2021.

[8]F. Xue et al.,"Vision Transformer With Attentive Pooling for Robust Facial Expression Recognition,"
IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3244–3256, 2023.

[9]Y. Chen et al.,"From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos,"
IEEE Transactions on Affective Computing, vol. 16, no. 2, pp. 624–638, 2025.

[10]J. So and Y. Han,"Facial Landmark-Driven Keypoint Feature Extraction for Robust Facial Expression Recognition,"
Sensors, vol. 25, no. 12, p. 3762, 2025.

[11]H. Tao and Q. Duan,"Hierarchical Attention Network with Progressive Feature Fusion for Facial Expression Recognition,"
Neural Networks, vol. 170, pp. 337–348, 2024.

[12]Y. Zhang et al.,"Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition,"arXiv:2207.10299, 2022.

[13]K. Wang et al.,"Suppressing Uncertainties for Large-Scale Facial Expression Recognition,"roc. IEEE/CVF CVPR, 2020, pp. 6897–6906.

[14]Y. Zhang, C. Wang, and W. Deng,"Relative Uncertainty Learning for Facial Expression Recognition,"Proc. NeurIPS, 2021.

[15]J. Zeng, S. Shan, and X. Chen,"Facial Expression Recognition With Inconsistently Annotated Datasets,"Proc. ECCV, 2018, pp. 222–237.

[16]J. She et al.,"Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition,"
Proc. IEEE/CVF CVPR, 2021, pp. 6248–6257.

[17]J. Shi et al.,"ARM: A Lightweight Module to Amend Facial Expression Representation,"Signal, Image and Video Processing, vol. 17, pp. 1315–1323, 2023.

[18]Z. Wen et al.,"Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition,"Biomimetics, vol. 8, no. 2, p. 199, 2023.

[19]K. Wang et al.,"Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition,"IEEE Transactions on Image Processing, vol. 29, pp. 4057–4069, 2020.

[20]Y. Zhang et al.,"EAC: Efficient Attention Calibration for Facial Expression Recognition,"Proc. ECCV, 2020.