Exploring New Frontiers in Facial Expression Recognition: Dual DenseNet-201 and Landmark Distance Analytics in the Wild

Conference Paper · July 2024

DOI: 10.1109/DDCLS61622.2024.10606687

3 authors, including:

La Sandhu
Central South University
22 PUBLICATIONS **42** CITATIONS

SEE PROFILE

Exploring New Frontiers in Facial Expression Recognition: Dual DenseNet-201 and Landmark Distance Analytics in the Wild

Abdullahi Mohamed Hassan
School of Automation
Central South University
Changsha, China
Gurey216@gmail.com

Xiaojun Zhou
School of Automation
Central South University
Changsha, China
michael.x.zhou@csu.edu.cn

Laeeq Aslam
School of Automation
Central South University
Changsha, China
204608004@csu.edu.cn

***Abstract*—This paper presents a novel approach to Facial Expression Recognition (FER), a critical component in human-computer interaction. Unlike traditional methods focusing solely on facial image analysis, our methodology intricately processes the distances between facial landmarks, employing dual DenseNet-201 models. This technique allows for comprehensive feature extraction from both the facial images and landmark distance data. The integration of these features using a multi-head attention mechanism within a transformer model marks a significant advancement in FER technology. Our method demonstrates improvements in performance, particularly in recognizing "Sad" expressions, as evidenced by extensive testing on the AffectNet dataset. This research not only sets a new benchmark in FER but also opens avenues for its application in areas such as psychology, surveillance, and interactive technologies.**

***Index Terms*—Facial Expression Recognition, Human-Computer Interaction, Deep Learning, Convolutional Neural Networks**

I. Introduction

As technology evolves and automation advances, the importance of human-computer interaction (HCI) intensifies. A key component in this domain is Facial Expression Recognition (FER), which enables machines to interpret human emotions through facial expressions. This capability positions FER as a pivotal element in HCI. Moreover, FER's sophisticated capability to comprehend expressions finds extensive applications in fields such as psychology, intelligent robotics, smart surveillance, virtual reality, and animated synthesis. Consequently, research in FER is not only beneficial but essential.

The field of Facial Expression Recognition (FER) has seen rapid advancements in recent years, garnering increasing attention. Initial FER research [1]–[4] relied on manual features [5]–[7] for analyzing facial expressions. These manually feature-based FER algorithms, however, were often limited to specific tasks and underperformed in real-world scenarios compared to controlled experiments. The advent of deep learning brought about a shift with the introduction of Convolutional Neural Networks (CNNs), enhancing the robustness in FER. Pioneers like Savchenko et al. [8] validated the effectiveness of CNNs, including MobileNet [9], EfficientNet [10], and RexNet [11] in FER. Zhao et al. introduced an innovative FER network, EfficientFace [12], optimizing facial expression analysis in natural settings. Despite their advancements, convolution-based FER algorithms have limitations in processing global image information due to their local receptive field. With the influence of vision transformers, Xue et al. [13] developed the first transformer-based FER network, enabling the modeling of long-range dependencies in FER. Further enhancements by Kim et al. [14] integrated global and local features in the Vision Transformer (ViT) to better adapt it for FER tasks.

Among the plethora of remarkable works in Facial Expression Recognition (FER), a simpler and stronger facial expression recognition network (POSTER) [15] emerges as a standout with its state-of-the-art performance. POSTER adeptly addresses three fundamental challenges in FER: inter-class similarity, intra-class discrepancy, and scale sensitivity. Its innovative approach intertwines facial landmark and image features using a dual-stream pyramidal cross-fusion transformer architecture. This design effectively mitigates the inter-class similarity and intra-class discrepancy by leveraging the distinctiveness and sparsity of landmarks. Additionally, POSTER's pyramid structure facilitates the integration of multi-scale features, adeptly handling the issue of scale sensitivity. The resolution of these core FER challenges underlines POSTER's exceptional capability in expression analysis.

The existing work has primarily focused on the analysis of facial images combined with their landmarks. These landmarks, which typically represent key facial features such as the eyes, nose, mouth, and jawline, provide crucial spatial information. By concatenating these landmarks with the facial images, researchers have been able to create a more enriched dataset. This enriched dataset, when analyzed using attention mechanisms, allows for a refined focus on specific areas of the face that are more expressive or relevant for emotion recognition. This approach enhances the FER process by effectively leveraging both the visual cues from the images and the spatial context provided by the landmarks, leading to potentially more accurate and nuanced emotion detection. The integration of landmarks with image data in FER represents

a significant stride in understanding the subtleties of human expressions. In this research, we diverge from the conventional approach of pairing landmark images with original facial images for Facial Expression Recognition (FER). Instead, we introduce a novel technique where the distances between facial landmarks are utilized as key features. This data is processed through two separate DenseNet models – one for the facial images and the other for the landmark distance features. The outputs of these models are then integrated using multi-head attention layers within a transformer model. This setup allows for cross-attention processing between the features of facial expressions and the landmark distances. This innovative methodology has demonstrated superior performance over existing approaches, showcasing marked improvements in average accuracy, thereby setting a new benchmark in the field of Facial Expression Recognition .

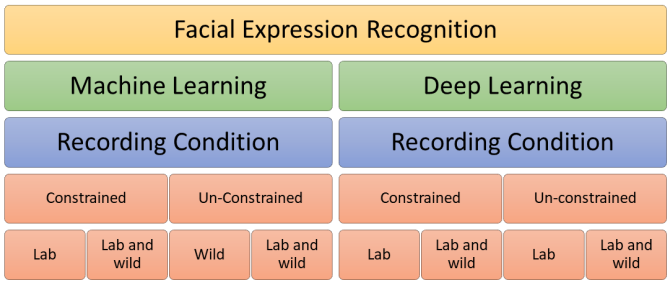

Fig. 1. Literature Review Break Down

II. Literature Review

This section reviews the literature on established traditional and advanced deep learning techniques for Facial Expression Recognition (FER). The structure of this review is illustrated in Fig. 1, where a classification based on recording conditions is presented. The data-sets used in these studies are categorized into two types: those recorded in controlled environments (labelled as lab recordings) and those captured in uncontrolled, natural settings (referred to as In-the-wild or IW). Studies that incorporate data-sets from both these environments are identified under both lab and IW recording conditions. This classification provides a comprehensive understanding of how different recording contexts influence FER research.

A. Traditional Machine Learning (ML) Approaches in Constrained and Unconstrained Environments

Traditional machine learning techniques often rely on handcrafted features for classifying expressions into respective emotion classes. For instance, Ghimire et al. [16] presented a method where geometric feature descriptor Normalized Central Moments (NCM) is combined with Local Binary Patterns (LBP) and processed using a Support Vector Machine (SVM) for classification. This approach particularly excels in recognizing expressions in constrained environments, demonstrating that domain-specific local region feature descriptors outperform holistic representations. Similarly, Zhong et al. [17] employed LBP and using a two-stage Multi-task Sparse Learning (MTSL). Niu et al. [18] enhanced the feature extraction process using LBP and an improved version of ORB, focusing on region-wise feature point extraction. Reference [19] explored the use of LBP-TOP and Bi-WOOF for capturing expression details in video apex frames. Liong et al. [20] employed Gabor filters and PCA, coupled with a genetic algorithm to optimize SVM for FER.

Guo et al. [21] proposed a scheme, capturing second-order discriminative information in local patches through Rotated Directional Local Binary Patterns and Adaptive Directional Local Binary Patterns - Three Orthogonal Planes. Rashmi and Annappa [22] developed an ensemble approach using DT and VD for geometric and texture feature combination, applying a stacking classifier and majority voting. Cruz et al. [23] utilized LBP in an unconstrained environment, while Chen et al. [24] combined visual and audio modalities with Histogram of Oriented Gradients - Three Orthogonal Planes by Pham et al. [25] introduced an MLP-based reliability assessment for facial expression classification, leveraging AUs in large, unlabeled datasets for more accurate micro-expression detection.

B. Deep learning (DL) approaches used in constrained and unconstrained environments

Deep learning techniques have demonstrated superior efficiency in extracting significant patterns from images, outperforming handcrafted features in classifying emotions. In constrained environments, Zhang et al. [26] utilized SIFT, feeding these features to a DNN for learning discriminative patterns. Lopes et al. [27] combined CNNs with image processing techniques for effective FER feature extraction. Barros et al. [28] focused on employing CNN convolution units to identify expression locations in cluttered scenes. Kim et al. [29] extracted spatial features using CNNs and trained LSTMs for generating spatio-temporal representations, enhancing FER with varied expression intensities. Pan et al. [30] aggregated spatial and temporal features, bridging the gap between visual features and emotions. Zhang et al. [31] applied PHRNN and MSCNN to capture dynamic and morphological facial expression variations in videos.

Sun et al. [32] employed CNNs, BLSTM-RNNs, and PCA for recognizing expressions from facial movements and gestures. Liang et al. [33] proposed a BiLSTM architecture, fusing spatial and temporal dynamics for FER. Sun et al. [34] used spatial features from gray-level images and optical flow from emotional and neutral faces. The features were fused using MDSTFN. Xie and Hu [35] introduced DCMA-CNN with ETI pooling to distinguish expression-sensitive elements. Ma et al. [36] conducted cross-modal noise modeling using 2D and 3D CNNs for audio and visual data, respectively. Majumder et al. [37] utilized autoencoders for non-linear data representation in lower dimensions. Tang et al. [38] demonstrated the superiority of automatically extracted features over handcrafted ones. Zhi et al. [39] employed an evolutionary DL approach for AU detection, showing high efficiency.

In uncontrolled environments, several deep learning approaches have demonstrated effectiveness in addressing challenges associated with in-the-wild databases, cross-cultural variability, computational complexity, overfitting, and small sample sizes. Georgescu et al. [40] utilized CNN architectures in conjunction with BOVW handcrafted features for facial expression recognition. They adopted both local and global learning strategies using SVM. Notably, the local learning SVM was particularly effective in FER prediction, overcoming issues like overfitting and vanishing gradients, thereby enhancing overall performance.

C. Advancements and Categorizations in Facial Expression Recognition Techniques

The field of Facial Expression Recognition (FER) has witnessed substantial advancements with the transition from traditional Machine Learning (ML) techniques to more advanced Deep Learning (DL) approaches. Traditional ML models are known for their effectiveness with limited data and fewer parameters, but they often struggle when dealing with larger datasets and ensuring generalizability [41]. On the other hand, DL methods, characterized by their multiple layers and the ability to autonomously learn high-level features, have shown remarkable capabilities in extracting nonlinear facial features and achieving higher accuracy in FER tasks [42]–[46]. Despite being more computationally intensive, the enhanced feature extraction and generalization abilities of DL models significantly surpass those of traditional ML models [41].

FER systems can be broadly categorized into two main approaches: static image-based and dynamic sequence-based methods [47]. The static approach focuses on analyzing current image features, but it lacks the ability to incorporate temporal information [27]. In contrast, the dynamic method utilizes the temporal dynamics between frames to recognize expressions, offering improved spatial and temporal feature extraction. However, this approach tends to increase computational complexity and the potential for noise in the data [44].

A key aspect in the effectiveness of FER algorithms, especially in AI-driven interactive systems, is the correlation between Action Units (AUs) and neural network features. The Facial Action Coding System (FACS), introduced by Ekman, plays a critical role in identifying facial regions and AUs associated with various muscle groups [48]. FER leveraging FACS generally involves the detection of AUs, followed by the categorization of emotions based on these units. Integrating AU information into neural network models significantly enhances the detection of subtle facial expressions [22], [49], although it requires meticulous engineering for optimal performance [50]. The strong correlation between the features learned by Convolutional Neural Networks (CNNs) and AUs underscores the effectiveness of CNNs in learning facial AUs, aligning with Ekman's FACS model [41], [51].

III. Preliminary Knowledge

This work intricately weaves together the concepts of landmarks and their inter-distances, Dense-Net for feature extraction, and attention mechanisms. In this section we be focusing on the preliminary knowledge of these components.

While machine learning has achieved remarkable success in various classification tasks, the challenge of facial expression recognition (FER) in "wild" or uncontrolled images remains significant. This difficulty arises due to the complexity and subtlety inherent in human facial expressions, which often involve nuanced, localized changes on the face. Attention models like the Vision Transformer (ViT) and Swin Transformer, although powerful in certain applications, tend to under perform in this context. The main issue is that the critical information for FER is often concentrated in small, specific areas of the face, which these models might not effectively capture or emphasize.

In contrast, Convolutional Neural Networks (CNNs) have shown better results in FER tasks. Architectures like ResNet, VGG16, and EfficientNet, which are adept at processing spatial hierarchies and local features, have been more successful. However, even these advanced CNN-based algorithms have limitations. The average accuracy rates for these models in FER tasks, particularly in "wild" or natural settings, are often reported to be less than 62% . This figure indicates that there's still significant room for improvement in accurately and reliably recognizing facial expressions from uncontrolled, real-world images.

The comparatively lower performance in FER tasks highlights the ongoing challenge of developing algorithms that can effectively interpret the complex, subtle variations in human facial expressions, especially when extracted from diverse and unstructured environments.

A. DenseNet Architecture

Dense Convolutional Network (DenseNet), a novel deep learning architecture, revolutionizes the way neural networks are structured. Distinct from traditional models, DenseNet introduces an innovative approach where each layer within a dense block concatenates the feature maps from all preceding layers, rather than adding them. This is described mathematically as $x_i = H([x_0, x_1, ..., x_{i-1}])$, where x_i represents the output of the i^{th} layer, and H symbolizes the composite function involving Batch Normalization, ReLU, and Convolution. This concatenation mechanism allows DenseNet to foster more complex and rich feature representations at each layer, significantly enhancing the model's learning and representational capabilities.

Unlike ResNet, which employs a summation of features, DenseNet's concatenation approach leads to a more effective gradient propagation and reduces the number of required parameters, making the network more efficient in terms of computational resources. The introduction of the 'growth rate' parameter in DenseNet is a crucial aspect, balancing the width of the network and controlling the number of feature maps produced by each layer, thereby ensuring an efficient usage of parameters.

Moreover, DenseNet architecture demonstrates improved feature reuse, which allows the network to be deeper while

requiring fewer parameters than conventional architectures. This design also alleviates the vanishing gradient problem, making DenseNet particularly effective for deeper networks. The cumulative effect of these features sets DenseNet apart from other architectures like ResNet, positioning it as a highly efficient model in terms of computational resources and learning capacity, especially in tasks involving complex visual patterns.

B. Attention Mechanism

In neural networks, particularly within the realm of transformers, the attention mechanism plays a pivotal role. It enables the model to dynamically focus on specific segments of the input data, which is crucial for tasks that require contextual understanding. The core components of this mechanism are the query (Q), key (K), and value (V) matrices, derived from the input.

The process begins by computing the dot product of Q and K, which assesses how much each element in the key should be attended to for each element in the query. This dot product is scaled down by $\frac{1}{\sqrt{d_k}}$, where d_k is the dimension of the key vectors. This scaling is vital for preventing the softmax function from entering regions where it has extremely steep gradients, thereby stabilizing the training process.

The attention weights are then obtained through:

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

These weights represent the significance of each element in the input sequence relative to the current query.

Finally, the output of the attention mechanism is computed as a weighted sum of the values (V), with the weights specified by the attention scores. This results in the attention output, which is contextually enriched and focused on relevant parts of the input.

This mechanism is a cornerstone in the transformer's ability to handle sequences, providing a nuanced approach that emphasizes relevant information over uniform treatment of all input data.

C. Landmarks and Their Role in Facial Expression Recognition

Our methodology leverages the state-of-the-art Ensemble of Regression Trees (ERT) method for facial landmark detection, which is instrumental in accurately identifying facial features in real-time applications. Introduced by Kazemi and Sullivan [52], the ERT method employs a cascade of regression trees to iteratively refine the estimates of facial landmark positions. Each regression tree predicts adjustments to the landmark positions based on the current estimates and extracted image features, significantly reducing the prediction error through the formula:

$$L^{(t+1)} = L^{(t)} + \sum_{k=1}^{K} R_k(F(I, L^{(t)}); \theta_k) \tag{2}$$

where $L^{(t)}$ represents the estimated landmark positions at iteration t, $F(I, L^{(t)})$ denotes the extracted features from the image I, and R_k corresponds to the adjustments predicted by the k-th regression tree, with θ_k being its parameters.

Subsequent to the landmark detection, our methodology incorporates the calculation of Euclidean distances between pairs of landmarks to analyze facial expressions. This distance metric is vital for understanding the variations in facial expressions as it quantifies the spatial relationships between different facial features. The distances are computed as follows:

$$E_{L_1L_2} = \sqrt{(x_{L_1} - x_{L_2})^2 + (y_{L_1} - y_{L_2})^2} \tag{3}$$

where $E_{L_1L_2}$ denotes the Euclidean distance between two landmarks L_1 and L_2, with (x_{L_1}, y_{L_1}) and (x_{L_2}, y_{L_2}) indicating their respective coordinates. By leveraging the precision of ERT-based landmark detection and the analytical power of distance metrics, our approach provides a robust framework for facial expression recognition and analysis.

The significance of landmarks extends beyond mere detection. These dynamic points reflect a spectrum of emotional states, making landmark detection and analysis vital in applications ranging from interactive technologies to psychological studies. Understanding human emotions through facial expressions, facilitated by the mathematical and computational analysis of these landmarks, is not just a technical endeavor but a gateway to deeper comprehension of human emotional expression.

IV. Proposed Methodology

The proposed facial expression recognition scheme aims to address the identified research gap in accurately and reliably recognizing facial expressions, particularly in uncontrolled or "wild" environments. This sophisticated approach integrates cutting-edge techniques in image processing and neural network architecture to overcome the limitations of current models, especially in interpreting the nuanced and subtle variations of human facial expressions.

In the preprocessing stage, the scheme begins with the extraction of facial landmarks from input images, forming a detailed geometrical representation of facial features. Addressing the challenge of complexity in facial expressions, these landmarks are used to compute a distance matrix, capturing the spatial relationships between various facial points. This step is crucial in enhancing the scheme's ability to discern subtle facial changes, a key aspect where existing methods often fall short.

The computed distance matrix is then combined with the original image to create an augmented input, a novel approach that enriches the input data with both geometrical and visual information. This augmented input is processed through dual DenseNet-201 models, chosen for their exceptional feature propagation and reuse capabilities. The selection of DenseNet-201 models is a strategic response to the limitations of attention models like ViT and Swin Transformer, and traditional CNN architectures such as ResNet, VGG16, and EfficientNet,

which have shown less than optimal performance in complex, real-world scenarios.

The core of the scheme lies in its integration and classification stage. Here, the distinct sets of features extracted from both the image and the computed distances are merged using a multi-head attention mechanism and a series of neural network layers. This integration ensures a comprehensive analysis of the facial data, crucial for overcoming the challenge of accurately classifying emotions in diverse and unstructured environments. The scheme, with its methodical and nuanced approach, exemplifies the advancements in facial expression recognition technology, directly addressing the research gap in the field.

The overall block diagram of the proposed scheme is shown in fig. 2. The following subsections will discuss the proposed scheme in detail, elucidating how each component contributes to bridging the current research gap in FER technology.

1) Step 1: Preprocessing: The preprocessing stage involves facial landmark detection on the input face image. Landmarks L_k are identified, and the Euclidean distances between each pair of landmarks L_i and L_j are calculated using equation 2. The resulting distance matrix D is reshaped and zero-padded to align with the dimensions of the input image I, forming an augmented input I_{aug} by concatenating D and I.

2) Step 2: Dual DenseNet 201 Feature Extraction: The augmented image I_{aug} and its distance feature matrix are fed into two distinct DenseNet-201 models to extract features F_{image} and F_{dist}. This bifurcated approach allows for specialized processing of the image and landmark distance data, leveraging the efficient feature extraction capabilities of DenseNet-201.

3) Step 3: Feature Integration and Classification: The extracted features F_{image} and F_{dist} are then encoded with positional encoding PE and processed through a multi-head attention mechanism MHA:

$$F_{combined} = \text{MHA}(\text{PE}(F_{image}), \text{PE}(F_{dist})) \tag{4}$$

Following the attention mechanism, add-and-norm layers, a feed-forward network FF, and a final dense layer with a sigmoid activation function σ are used for classification:

$$\text{Output} = \sigma(\text{FF}(F_{combined})) \tag{5}$$

DenseNet-201 is selected for its effectiveness in feature propagation and reuse, surpassing other architectures like ResNet and RegNet in facial expression recognition tasks. Its dense connectivity pattern ensures a comprehensive analysis of both the image and landmark distance information, essential for accurate emotion classification. The pseudo code of proposed scheme is given below

Algorithm 1 Facial Expression Recognition Pipeline

Input: Image I
Output: Emotion class
procedure PREPROCESSING(I)
    $L_k \leftarrow \text{DetectLandmarks}(I)$
    $D_{ij} \leftarrow \sqrt{(L_{ix} - L_{jx})^2 + (L_{iy} - L_{jy})^2}$
    $I_{aug} \leftarrow \text{Concat}(\text{Reshape}(D), I)$
end procedure
procedure DUALDENSENET(I_{aug})
    $F_{image} \leftarrow \text{DenseNet201}(I)$
    $F_{dist} \leftarrow \text{DenseNet201}(D)$
end procedure
procedure INTEGRATION
    $F_{combined} \leftarrow \text{MHA}(\text{PE}(F_{image}), \text{PE}(F_{dist}))$
end procedure
procedure CLASSIFY
    return $\sigma(\text{FF}(F_{combined}))$
end procedure

V. Results and Comparison

A. Experiment Setup

Our evaluation of the Facial Expression Recognition (FER) performance in the Proposed Work leverages the AffectNet dataset, which exists in two distinct configurations:

- **AffectNet (7 cls)**: This configuration of the dataset encompasses a total of 280,401 images designated for training purposes. Additionally, it includes 3,500 images set aside for validation, distributed across seven fundamental categories of emotions.
- **AffectNet (8 cls)**: In this variant, the training set comprises 283,501 images, while the validation set contains 4,000 images. This version extends the emotional categories to eight, incorporating contempt as an additional category.

The extensive and diverse nature of the AffectNet dataset, recognized as one of the largest available datasets in the FER domain, provides a robust platform for the thorough evaluation of our Proposed Work. Its wide-ranging emotional representations, captured under various settings, are instrumental in facilitating a comprehensive assessment of FER performance. The selected hyperparatmers are show in table I

TABLE I
HYPERPARAMETERS OF THE PROPOSED FACIAL EXPRESSION RECOGNITION SCHEME

Parameter	Value
Preprocessing image size	128x128x3 pixels
Number of heads in MHA	4
Feed-forward network size in MHA	512
Activation function in final layer	Softmax
Learning rate	1e-3
Batch size	32
Number of training epochs	500

B. Comparison of Facial Expression Recognition Methods

This section presents a comprehensive comparison of various facial expression recognition methods, highlighting recent advancements in the field. The performance of these methods is evaluated using the AffectNet dataset in both 7-class and 8-class configurations.

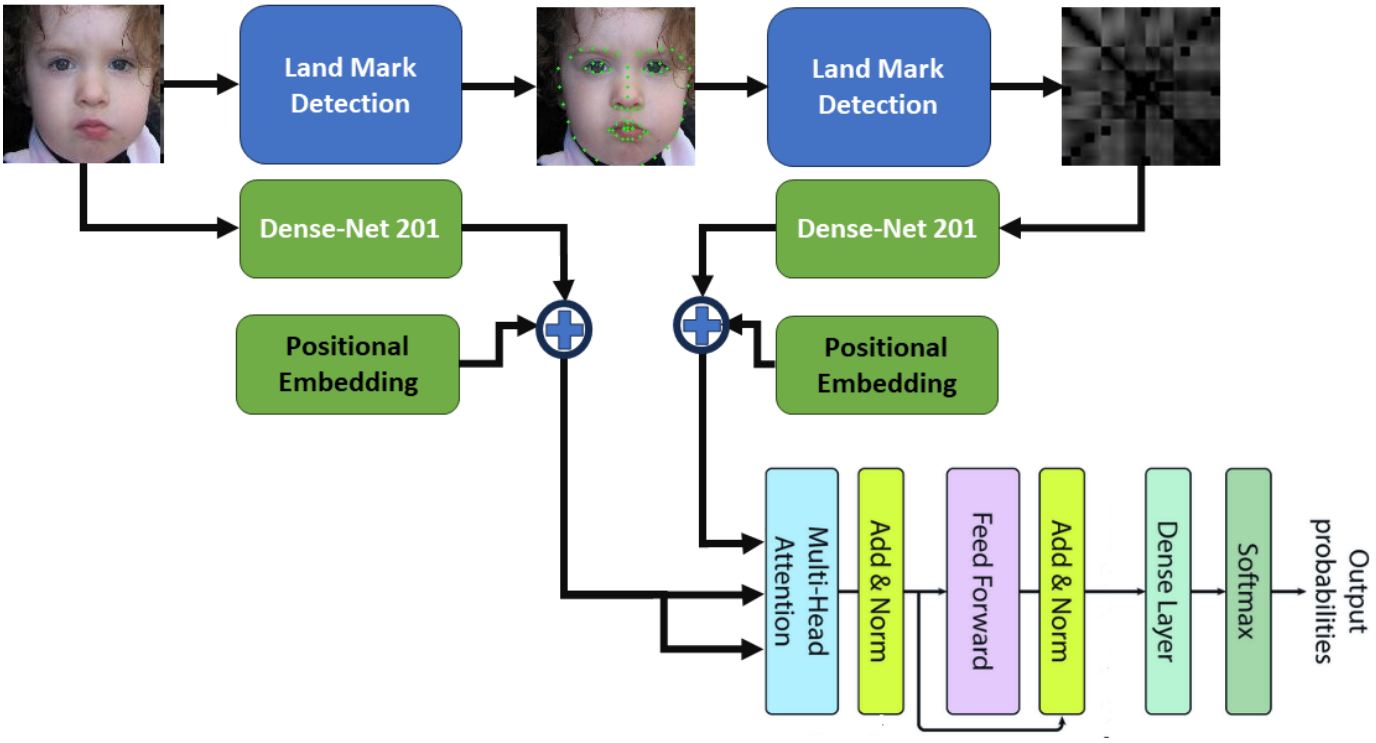

Fig. 2. Block diagram of Proposed scheme

TABLE II
RESULTS COMPARISON OF PROPOSED SCHEME WITH POSTER++

Dataset	Method	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt
AffectNet (7 cls)	Proposed Scheme	63.98%	63.12%	99.72%	64.35%	64.17%	63.89%	65.76%	-
AffectNet (7 cls)	POSTER++	65.40%	89.40%	68.00%	66.00%	64.20%	54.40%	65.00%	-
AffectNet (8 cls)	Proposed Scheme	60.88%	76.68%	67.08%	65.88%	63.28%	58.28%	60.48%	59.80%
AffectNet (8 cls)	POSTER++	60.60%	76.40%	66.80%	65.60%	63.00%	58.00%	60.20%	59.52%

C. Overview of Recent Methods

Table III summarizes the performance of different facial expression recognition methods developed in recent years. The methods are compared based on their accuracy on the AffectNet dataset, covering both 7-class and 8-class configurations. Notably, the approaches have evolved significantly over time, with newer methods like POSTER [53] and POSTER++ [54] showing substantial improvements in accuracy.It is observed that methods like PSR [55] and DA-N [56] have shown promising results, but the recent POSTER++ approach [54] and the proposed scheme seem to outperform others in overall accuracy.

D. Detailed Comparison of POSTER++ and Proposed Scheme

A closer inspection of the latest methods, POSTER++ and the proposed scheme, reveals interesting insights, particularly in their performance across different classes in the AffectNet dataset. In the 7-class configuration, while POSTER++ shows a higher accuracy in 'Neutral' and 'Happy' classes, the proposed scheme excels significantly in the 'Sad' class with an accuracy of 99.72%, far exceeding POSTER++'s 68.00%. This exceptional performance in detecting 'Sad' expressions contributes to the proposed scheme's higher overall mean accuracy of 69.28%, compared to POSTER++'s 67.45%. The proposed scheme excels in recognizing 'Sad', 'Disgust', and 'Anger' expressions by analyzing changes in distances between facial landmarks, which are prominent for these emotions, achieving high accuracy, notably 99.72% for sadness. However, its performance on 'Happy' expressions is slightly lower due to uniform facial movements that cause less distinct landmark distance changes. This highlights the scheme's strength in detecting expressions with clear geometric variations and suggests incorporating additional features for more uniformly expressed emotions like happiness. In the 8-class configuration, the proposed scheme continues to demonstrate its strength across most classes, slightly trailing in 'Neutral' and 'Happy'. However, it shows comparable or superior performance in 'Sad', 'Surprise', 'Fear', 'Disgust', 'Anger', and 'Contempt'. Notably, the performance in the 'Contempt' class stands out, with the proposed scheme achieving 59.80% accuracy, surpassing POSTER++'s 59.52%. Consequently, this results in an overall mean accuracy of 64.04% for the proposed scheme, which is marginally higher than POSTER++'s 63.76%. These

TABLE III
COMPARISON OF DIFFERENT METHODS ON AFFECTNET (7 CLS) AND AFFECTNET (8 CLS)

Methods	**Year**	**AffectNet (7 cls)**	**AffectNet (8 cls)**
SCN [57]	CVPR 2020	-	60.23
PSR [55]	CVPR 2020	63.77	60.68
LDL-ALSG [58]	CVPR 2020	59.35	-
RAN [59]	TIP 2020	-	-
DACL [60]	WACV 2020	65.20	-
KTN [61]	TIP 2021	63.97	-
DMUE [62]	CVPR 2021	63.11	-
FDRL [63]	CVPR 2021	-	-
VTFF [64]	TAC 2021	61.85	-
ARM [65]	2021	65.20	61.33
TransFER [66]	ICCV 2021	66.23	-
DAN [56]	2021	65.69	62.09
EfficientFace [67]	AAAI 2021	63.70	60.23
MA-Net [68]	TIP 2021	64.53	60.29
Meta-Face2Exp [69]	CVPR 2022	64.23	-
EAC [70]	ECCV 2022	65.32	-
POSTER [53]	2022	67.31	63.34
POSTER++ [54]	2023	67.49	63.77
Proposed Scheme	-	**69.28**	**64.04**

results indicate that the proposed scheme, while slightly underperforming in certain expressions, offers a more balanced and consistently high accuracy across a broader range of emotional expressions compared to POSTER++.

VI. CONCLUSION

This study marks a significant stride in FER, employing a novel approach centered on analyzing distances between facial landmarks using dual DenseNet-201 models. Our method has demonstrated impressive accuracy, particularly in detecting "Sad" expressions, as evidenced by tests on the AffectNet dataset. However, it is important to recognize certain limitations. The current approach may not fully account for the variability in facial expressions across different demographic groups, which could affect its applicability in global, diverse settings. Additionally, the computational complexity of our method may pose challenges in real-time applications. Future work could focus on optimizing the model for faster processing and enhancing its adaptability to diverse demographic characteristics. By addressing these limitations, we aim to further refine FER technology, broadening its potential in enhancing human-computer interaction and other applications.

REFERENCES

[1] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[2] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2562–2569.

[3] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Computer vision and image understanding*, vol. 115, no. 4, pp. 541–558, 2011.

[4] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*. IEEE, 2008, pp. 1–6.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[7] T. Jabid, M. H. Kabir, and O. Chae, "Local directional pattern (ldp) for face recognition," in *2010 digest of technical papers international conference on consumer electronics (ICCE)*. IEEE, 2010, pp. 329–330.

[8] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 2021, pp. 119–124.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[10] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[11] D. Han, S. Yun, B. Heo, and Y. Yoo, "Rexnet: Diminishing representational bottleneck on convolutional neural network," *arXiv preprint arXiv:2007.00992*, 2020.

[12] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 3510–3519.

[13] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3601–3610.

[14] S. Kim, J. Nam, and B. C. Ko, "Facial expression recognition based on squeeze vision transformer," *Sensors*, vol. 22, no. 10, p. 3729, 2022.

[15] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," *arXiv preprint arXiv:2204.04083*, 2022.

[16] D. Ghimire, S. Jeong, J. Lee, and S. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7803–7821, 2017.

[17] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1499–1510, 2014.

[18] B. Niu, Z. Gao, and B. Guo, "Facial expression recognition with lbp and orb features," *Computational Intelligence and Neuroscience*, 2021.

[19] S. Liong, J. See, K. Wong, and R. Phan, "Less is more: micro-expression recognition from video using apex frame," *Signal Processing*, vol. 62, pp. 82–92, 2018.

[20] A. Boughida, M. Kouahla, and Y. Lafifi, "A novel approach for facial expression recognition based on gabor filters and genetic algorithm," *Evolutionary Systems*, vol. 13, no. 2, pp. 331–345, 2022.

[21] C. Guo, J. Liang, G. Zhan, Z. Liu, M. Pietikäinen, and L. Liu, "Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition," *IEEE Access*, vol. 7, pp. 174 517–174 530, 2019.

[22] R. Rashmi and B. Annappa, "Micro expression recognition using delaunay triangulation and voronoi tessellation," *IETE Journal of Research*, pp. 1–17, 2022.

[23] A. Cruz, B. Bhanu, and N. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 418–431, 2014.

[24] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38–50, 2016.

[25] T. Pham, S. Kim, Y. Lu, S. Jung, and C. Won, "Facial action units-based image retrieval for facial expression recognition," *IEEE Access*, vol. 7, pp. 5200–5207, 2019.

[26] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[27] A. Lopes, E. de Aguiar, A. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with

few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[28] P. Barros, G. Parisi, C. Weber, and S. Wermter, "Emotion-modulated attention improves expression recognition: a deep learning model," *Neurocomputing*, vol. 253, pp. 104–114, 2017.

[29] D. Kim, W. Baddar, J. Jang, and Y. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, 2017.

[30] X. Pan, G. Ying, G. Chen, H. Li, and W. Li, "A deep spatial and temporal aggregation framework for video-based facial expression recognition," *IEEE Access*, vol. 7, pp. 48 807–48 815, 2019.

[31] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[32] B. Sun, S. Cao, J. He, and L. Yu, "Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy," *Neural Networks*, vol. 105, pp. 36–51, 2018.

[33] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional bilstm fusion network for facial expression recognition," *Visual Computer*, vol. 36, no. 3, pp. 499–508, 2020.

[34] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, vol. 119, pp. 49–61, 2019.

[35] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2018.

[36] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[37] A. Majumder, L. Behera, and V. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 103–114, 2016.

[38] Y. Tang, X. Zhang, and H. Wang, "Geometric-convolutional feature fusion based on learning propagation for facial expression recognition," *IEEE Access*, vol. 6, pp. 42 532–42 540, 2018.

[39] R. Zhi, C. Zhou, T. Li, S. Liu, and Y. Jin, "Action unit analysis enhanced facial expression recognition by deep neural network evolution," *Neurocomputing*, vol. 425, pp. 135–148, 2021.

[40] M. Georgescu, R. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64 827–64 836, 2019.

[41] J. Hung, K. Lin, and N. Lai, "Recognizing learning emotion based on convolutional neural networks and transfer learning," *Applied Soft Computing*, vol. 84, p. 105724, 2019.

[42] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, and Z. Luo, "Conditional convolution neural network enhanced random forest for facial expression recognition," *Pattern Recognition*, vol. 84, pp. 251–261, 2018.

[43] A. Mollahosseini, D. Chan, and M. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[44] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, vol. 119, pp. 49–61, 2019.

[45] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Computer Methods and Programs in Biomedicine*, p. 106621, 2022.

[46] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *arXiv preprint arXiv:1912.00271*, 2019.

[47] D. Kim, W. Baddar, J. Jang, and Y. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, 2017.

[48] P. Ekman and W. Friesen, *Facial Action Coding System: Investigator's Guide*. Washington, DC: Consulting Psychologists Press, 1978.

[49] C. Guo, J. Liang, G. Zhan, Z. Liu, M. Pietikäinen, and L. Liu, "Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition," *IEEE Access*, vol. 7, pp. 174 517–174 530, 2019.

[50] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 19–27.

[51] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[52] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[53] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," *arXiv preprint arXiv:2204.04083*, 2022.

[54] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, "Poster v2: A simpler and stronger facial expression recognition network," *arXiv preprint arXiv:2301.12149*, 2023.

[55] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131 988–132 001, 2020.

[56] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *arXiv preprint arXiv:2109.07270*, 2021.

[57] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906.

[58] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 984–13 993.

[59] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[60] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2402–2411.

[61] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via cf labels and distillation," *IEEE Transactions on Image Processing*, vol. 30, pp. 2016–2028, 2021.

[62] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.

[63] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7660–7669.

[64] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, 2021.

[65] J. Shi, S. Zhu, and Z. Liang, "Learning to amend facial expression representation via de-albino and affinity," *arXiv preprint arXiv:2103.10189*, 2021.

[66] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3601–3610.

[67] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 3510–3519.

[68] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021.

[69] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2exp: Combating data biases for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 291–20 300.

[70] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 418–434.