# Exp-Graph: How Connections Learn Facial Attributes in Graph-based Expression Recognition

Nandani Sharma, Dinesh Singh
*Vision Intelligence and Machine Learning (VIML) Group*
*School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, India.*
d22180@students.iitmandi.ac.in, dineshsingh@iitmandi.ac.in

***Abstract*—Facial expression recognition is crucial for human-computer interaction applications such as face animation, video surveillance, affective computing, medical analysis, etc. Since the structure of facial attributes varies with facial expressions, incorporating structural information into facial attributes is essential for facial expression recognition. In this paper, we propose Exp-Graph, a novel framework designed to represent the structural relationships among facial attributes using graph-based modeling for facial expression recognition. For facial attributes graph representation, facial landmarks are used as the graph's vertices. At the same time, the edges are determined based on the proximity of the facial landmark and the similarity of the local appearance of the facial attributes encoded using the vision transformer. Additionally, graph convolutional networks are utilized to capture and integrate these structural dependencies into the encoding of facial attributes, thereby enhancing the accuracy of expression recognition. Thus, Exp-Graph learns from the facial attribute graphs highly expressive semantic representations. On the other hand, the vision transformer and graph convolutional blocks help the framework exploit the local and global dependencies among the facial attributes that are essential for the recognition of facial expressions. We conducted comprehensive evaluations of the proposed Exp-Graph model on three benchmark datasets: Oulu-CASIA, eNTERFACE05, and AFEW. The model achieved recognition accuracies of 98.09%, 79.01%, and 56.39%, respectively. These results indicate that Exp-Graph maintains strong generalization capabilities across both controlled laboratory settings and real-world, unconstrained environments, underscoring its effectiveness for practical facial expression recognition applications.**

***Index Terms*—Facial expression recognition, graph convolutional networks, and vision transformer.**

## I. Introduction

FACIAL expression recognition (FER) has garnered significant attention in computer vision research over the last few decades because of its critical role in enabling computers to comprehend human emotions and engage in human-to-human communication as shown in Fig. 1. However, their success lies in learning robust and discriminative representations of the expressions ( such as AN: anger, DI: disgust, FE: fear, HA: happy, SA: sad, SU: surprise, NE: neutral) from the facial images that are invariant to variations in the angle of viewpoint, lighting conditions, and head postures. Thus, devising a suitable feature representation is the primary objective of facial expression recognition. In the early research, successful hand-crafted features, such as Gabor wavelets [1], local binary pattern (LBP) [2], [3], histogram of oriented gradients (HoG) [4]–[7], etc., were used to represent the facial expressions. However, these methods fall short of adequate recognition in more complicated real-world situations because of their low semantic correlation with facial expressions.

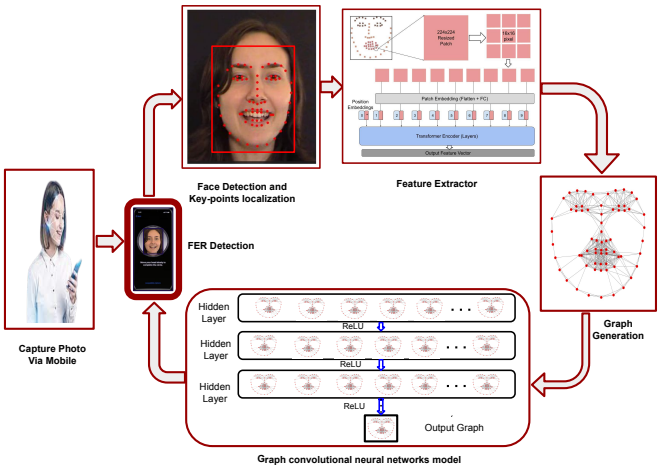



Fig. 1. System architecture of facial expression recognition using Exp-Graph framework. [Best shown in color]

Since deep learning techniques have rapidly advanced in the past decade, numerous attempts have been made to investigate discriminative representations for various recognition tasks. Deep models for visual representation, especially convolutional neural networks (CNNs) and vision transformers (ViTs), have shown promising results in various real-world applications because they can effectively learn discriminatory feature representations from visual observations [8]–[11]. Also, several studies have used CNNs to improve semantic representations of facial expressions and demonstrated good results in identifying human emotions [12]–[16]. However, they mostly depend on the appearance only and cannot exploit the deep structure for the problems where the training data is limited [17]–[19].

Vision transformers [20] capture global context by implementing self-attention processes; they are becoming a more popular choice for visual feature extraction compared to standard deep neural networks (DNNs) like CNNs. Vision transformers can analyze an image as a sequence of patches, allowing for greater input size flexibility and better scalability

This paper was produced by the IEEE Publication Technology Group.
Manuscript received April XX, 2025; revised August XX, 2025.

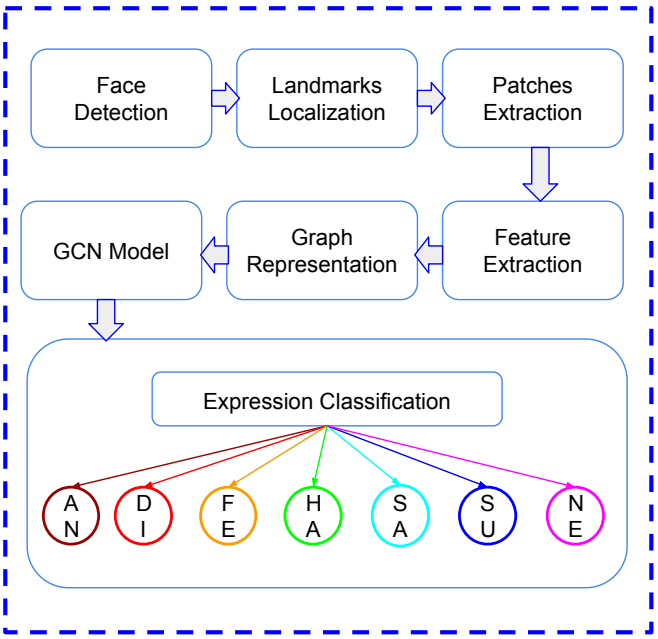



Fig. 2. General pipeline of our system.

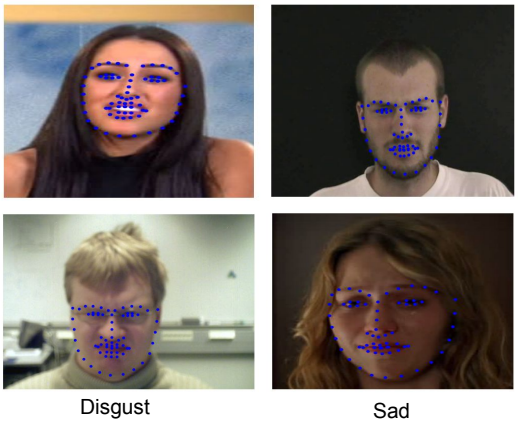



Fig. 3. Geometry alone is insufficient. [Best shown in color]

with large-scale datasets than CNNs, which concentrate on local patterns only. Additionally, ViTs are more generalizable because of their excellent transfer learning capabilities and less inductive bias. Vision transformers are increasingly favoured for visual feature extraction over traditional CNNs due to their ability to capture global context via self-attention, offering better scalability and input flexibility. Unlike CNNs, which focus on local patterns and are sensitive to variations like lighting and occlusions, ViTs can generalize better and enhance transfer learning.

This research explores using ViTs for facial expression recognition to address the limitations of CNNs, aiming to improve accuracy and reliability in recognizing complex facial expressions [21], [22]. Attention methods like graph attention (GAT) [23] and ViT [24] at both geometry and appearance levels are used to improve FER performance. Vision transformers excel in capturing global features for image recognition, including FER, but they struggle with local feature extraction and require large datasets [20], [25]. Graph convolutional networks (GCNs) address these limitations by enhancing local feature detection and improving data efficiency, making ViTs more effective for FER by capturing subtle facial details [9], [26], [27]. Integrating GCNs with ViTs offers a balanced solution, combining global and local feature learning for improved performance. Since facial expression highly depends on the relative change in the structure of facial attributes, incorporating graph structures of the facial attributes can significantly improve the facial expression representation, mainly when relying on transfer learning for visual encoding. However, traditional DNNs also struggle with non-Euclidean data, such as graphs, where relationships between data points are complex and irregular. Also, extensive feature engineering is required to capture complex relationships between data points. Scalability is another challenge for DNNs when dealing with large, highly connected datasets, and they need a large amount of labeled data to perform well. The lack of inductive bias in DNNs for relational data further hampers their ability to generalize across tasks where relational information is critical.

In contrast, GCNs are designed to handle graph-structured data, allowing them to represent and process such information naturally. Graph convolutional networks effectively learn and represent node connections, requiring fewer labeled samples and improving their generalization capabilities because of their built-in structure and natural inductive bias for graphs [28]. Therefore, GCNs offer distinct advantages over DNNs in handling graph-structured data and learning complex relationships. However, facial expression recognition using GCNs facial challenges, such as constructing complex graphs, dealing with high-dimensional and irregular data, learning robust features, managing lighting, occlusions, head poses, scalability, and real-time performance. Integrating GCNs with ViTs in facial expression recognition can address challenges by combining GCNs' ability to capture structural relationships between facial landmarks with ViT's strength in modeling both local and global features. This hybrid approach requires advancement in GCNs' architecture and innovative training strategies to effectively merge the strengths of both models, leading to improved accuracy in recognizing complex facial expressions. Few works explore geometric knowledge of facial attributes for facial expression recognition [29]–[31]. Geometric information, such as relative location and self-deformation, can accurately describe emotional states based on facial observations [32]. Geometric face descriptions are more resistant to appearance changes, making them ideal for real-world facial expression recognition applications [9], [32], [33]. In the GCNs methods [9], [32], [32]–[46], landmarks, AUs (or nodes) and the connections (edges) between them are often predefined or fixed, meaning that the graph's structure remains constant during the learning process. In contrast, our model introduces a more flexible or dynamic approach. Instead of using a fixed graph structure, we are allowing the model to

learn the connections between nodes while using the threshold ($\tau$) hyperparameter, potentially evaluating the graph structure during training. Our approach could enable the model to learn more meaningful or relevant relationships between nodes rather than relying on static or predefined connections.

Our research presents an Exp-Graph framework as shown in Fig. 2 that uses GCNs [28], [47], [48] to learn geometric descriptions from facial landmarks. The system architecture seeks to increase emotional reasoning from facial images, as geometric information alone cannot distinguish geometrically identical expressions such as disgust and sadness, as shown in Fig 3. Consequently, GCNs offer a valuable substitute for incorporating the geometric information derived from facial landmarks into emotional representations. Local appearance representations are extracted from landmark positions and aggregated with geometric representations during graph learning. The following briefly describes the primary contributions of this paper:

- Achieves expressive representation of facial expressions by utilizing graph convolutional networks to model structural information extracted from features obtained via a pre-trained vision transformer.
- Captures local and global semantic relationships among facial attributes to learn meaningful expression representations effectively.
- Conducts comprehensive evaluations of the proposed Exp-Graph approach on publicly available datasets, including Oulu-CASIA [49], eNTERFACE05 [50], and AFEW [51] are publicly available datasets.

The rest of the paper is organized as follows: Section II reviews related work in facial expression recognition. Section III provides a comprehensive overview of the proposed Exp-Graph framework, detailing its design and methodology. Section IV describes the experimental setup, including dataset specifications and evaluation metrics, and presents a thorough analysis of the results. Finally, it concludes the paper and outlines potential directions in §V.

## II. RELATED WORK

Since structural information is crucial for facial expression recognition, as emphasized in the previous section, some works have incorporated geometric feature extraction [26], [31], [52] for facial expression recognition. Also, GCNs [28], [53], [54] and vision transformers [20] have already been explored in different works. Here, we present their limitations and the key differences with our proposed work.

### A. Geometry-based FER

Due to the high association between geometric knowledge and expression representations, several researchers propose using landmark geometries for face expression manipulation [52], [55], [56]. Furthermore, several studies were conducted that suggested using landmark placements as a guide to identify noteworthy local characteristics for representation learning [57]–[59]. Kotsia *et al.* [31] employed geometric information to identify informative frames from facial expression sequences, whereas Zhang *et al.* [60] incorporated fiducial points on face images to characterize emotions. Gaining momentum from the explosive growth of deep learning technology over the last decade, FER is paying more and more attention to the learning of geometry-associated representations of features. However, most of these techniques adopt multi-task learning approaches instead of directly learning from the geometric data. Devries *et al.* [61] introduced simultaneously learning facial landmark localization and facial expression recognition to enhance the geometric understanding of emotion-related features. Additionally, a multi-domain multi-task network with landmark detection for FER was presented by Gerard and Masip [62]. Zhang *et al.* [26] used generative adversarial networks in conjunction with face landmarks to train the pose-invariant features for FER. Investigates using landmark locations as feature descriptions for FER in geometric facial landmarks. However, in practical applications, it is challenging to generate discriminative features due to the poor semantic correlation of these locations. A multimodal auto-encoder was presented by Zhang *et al.* [63] to learn a combined representation from both geometric and visual modalities. Only landmark information is considered in the above research on graph-based representations for extracting geometric information from face images. However, this paper aims to look at a graph-based learning strategy proper for feature representations to provide reliable geometric knowledge.

### B. Transformer mechanism for FER

Attention mechanisms are explored in the highlight of the regions of interest (RoI) in the domain of facial expression recognition [64], [65]. Our studies investigate integrating attention mechanisms into GCNs to enable attentive graph-based representation learning. While transformers, initially proposed for natural language processing (NLP), have become a popular method for handling sequential data, their application in graph-based vision tasks has not been extensively explored. There has been significant research into using ViTs for image interpretation [25], [66], [67]. Dosovitskiy *et al.* [20] introduced a vision transformer for image classification, where an image is divided into patches that serve as tokens to learn non-linear mappings based on dense correlations among all tokens. Yuan *et al.* [68] further refined the ViT [20] approach by developing a more generalized transformer design. Our study aims to extend the vision transformer into the domain of graph-based learning to establish longer-range dependencies among vertices in a series of facial graphs. By doing so, it seeks to enhance the effectiveness of facial expression recognition.

### C. GCNs for Analyzing Facial Images

Graph convolutional networks effectively evaluate structured data across various domains, including skeleton-based action detection, NLP, and semi-supervised learning [28], [69], [70]. However, while GCNs have been extensively applied to text datasets, their application to image data presents unique challenges that previous studies have not fully addressed. In the domain of facial expression recognition, Zhou *et al.* [71] introduced a FER framework, which uses end-to-end feature learning based on facial topological structures to

automatically learn patterns over time and space. However, the method [71] relies on pre-established facial landmarks identified by HOG as nodes, limiting the flexibility of the graph learning approach. To address the method [71], Zhou *et al.* [72] later introduced an improved method, but it still faced similar limitations. Zhao *et al.* [32] proposed a geometry-aware FER framework that combines appearance and geometric data using GCNs. This method extensively evaluates the structural information of facial attributes across various expressions and uses CNNs to extract general expression characteristics.

Recently, Qu *et al.* [40] combined a spatio-temporal graph convolutional model with a self-attention mechanism, automatically adjusting attention distribution across peak frame images. Luo *et al.* [27] proposed NFER-Former, a hypergraph-guided feature embedding approach designed to model significant facial actions and capture their complex interrelationships, supported by the introduction of a large NIR-VIS facial expression dataset for validation. Dong *et al.* [37] developed an attention-based visual GCNs for FER, addressing data processing inflexibility by incorporating pixel-level composition strategies. Jin et al. [34] presented a region-of-interest (ROI) -based method, constructing facial graphs from cropped ROIs of action units (AUs) using a deep auto-encoder. Similarly, Chen *et al.* [34] proposed a node classification approach for FER based on dual subspace manifold learning. Despite these advancements, existing methods cannot dynamically allocate edges and nodes during GCNs training.

In summary, conventional facial feature extraction methods—such as HOG, LBP, and CNNs—struggle to capture subtle variations and interdependencies in facial expressions effectively. More advanced approaches, including hypergraph-based embeddings and vision transformers, are better equipped to handle these complexities, particularly when applied to large and diverse datasets. Vision transformers, in particular, excel at modeling global context and capturing nuanced feature interactions, making them highly suitable for facial expression recognition. Recent studies have focused on combining vision transformers with graph convolutional networks to improve FER performance. These efforts integrate global visual cues with structural facial information, leverage localized convolutional branches for detailed appearance features, and apply attention mechanisms within graph-based learning frameworks to better understand facial dynamics.

## III. Exp-Graph Framework

The proposed Exp-Graph framework integrates face detection, feature encoding using pre-trained ViT, as shown in Fig. 4, and recognition via GCNs, as demonstrated in Fig. 5. The main steps of the framework are as follows: (A) encoding facial attributes through graph-based representation, (B) facial expression recognition through graph convolutional networks. Additionally, we detail the encoding of landmark geometry, the extraction and alignment of local visual features with these geometric representations, and the network architectures involved.

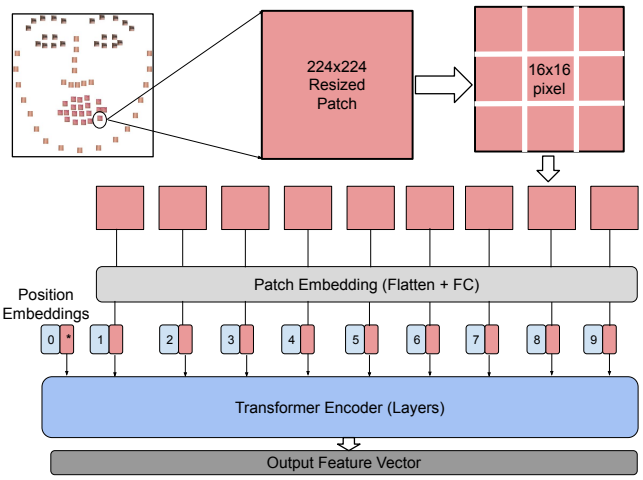



Fig. 4. An illustration of feature extraction of facial units using vision transformer

### A. Encoding Facial Attributes Through Graph-Based Representation

The process begins with image preprocessing to normalize dimensions and improve visual quality. Facial landmarks are detected using the Dlib [73], and the patch around the landmark points is encoded using a pre-trained ViT [20]. The facial attribute graph is built by generating an adjacency matrix $\boldsymbol{A}$, which captures relationships between facial landmarks based on their spatial proximity and feature similarity. This approach effectively captures the appearance of facial attributes and relates them to various expressions. The procedure starts by applying $L_2$ normalization to each feature vector, followed by computing a similarity measure $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ for pairs of feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. Simultaneously, a distance matrix is computed based on the squared Euclidean distances between the spatial coordinates of the landmarks. The initial adjacency matrix $\boldsymbol{A_{ij}}$ is derived by normalizing the similarity measure using an exponential function of the Euclidean distance, as illustrated in Eq. (1). This method effectively integrates both feature and spatial information into the graph structure.

$$\boldsymbol{A}_{ij} = \frac{\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)}{e^{\|\mathbf{p}_i - \mathbf{p}_j\|}}. \tag{1}$$

A thresholding step ($\boldsymbol{T}_s = \mu_K + \tau \cdot \sigma_K$) is applied to refine the adjacency matrix further. Specifically, a threshold parameter $\tau$ filters out weak connections, retaining only significant relationships between landmarks and $\mu_K$ and $\sigma_K$ are the mean and standard deviation of the $\boldsymbol{A}_{ij}$. As defined in Eq. (2), if $\boldsymbol{A}_{ij}$ surpasses the threshold $\boldsymbol{T}_s$, it is assigned a value of 1; otherwise, it is set to 0.

$$\boldsymbol{A} = \begin{cases} 1, & \text{if } \boldsymbol{A}_{ij} > \boldsymbol{T}_s \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This formulation ensures that landmarks are close in space and similar features have stronger connections in the matrix. The facial landmarks are nodes, and both feature similarity and spatial proximity weight the edges (connections) between

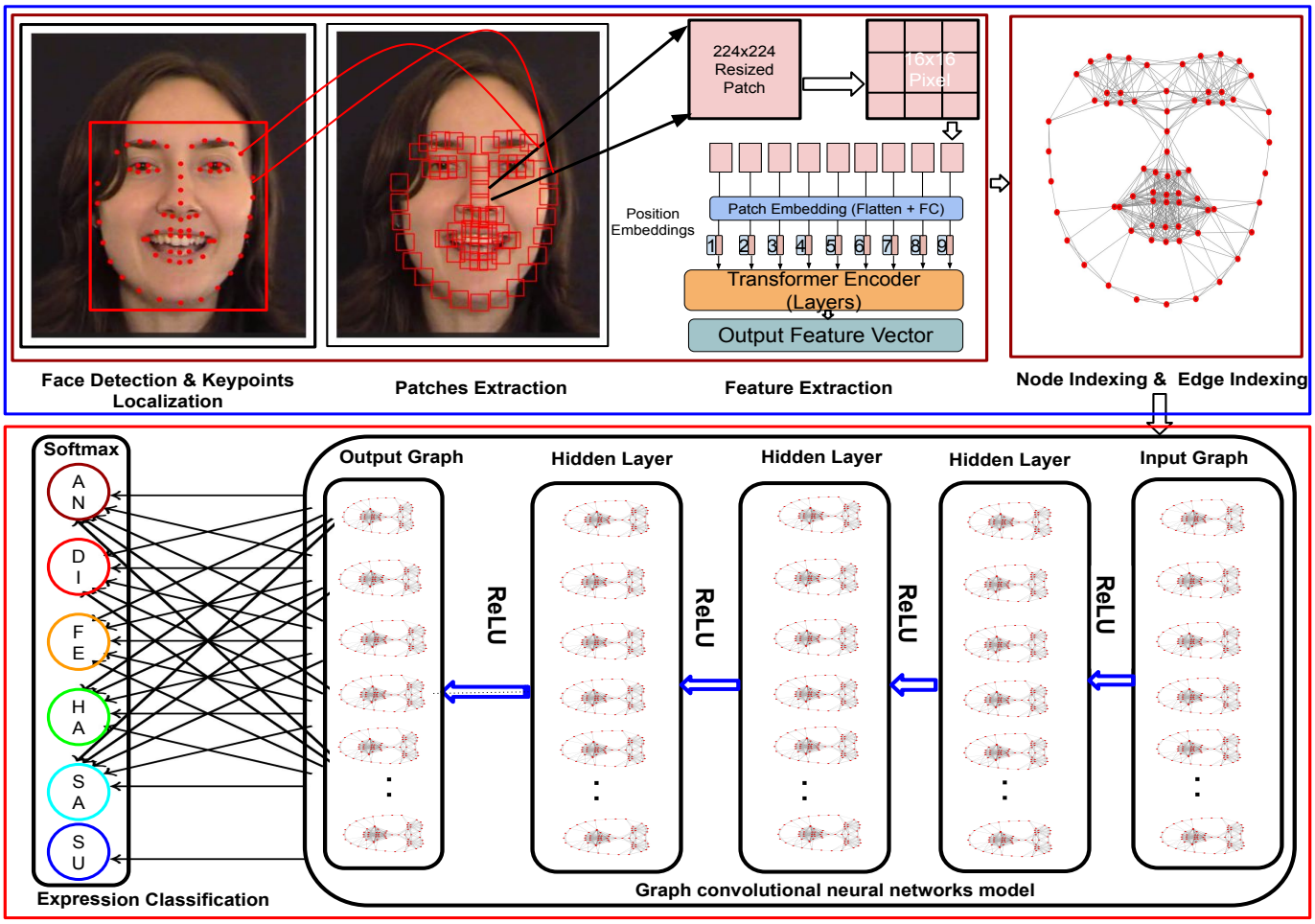



Fig. 5. An outline of the proposed framework for detecting and recognizing facial expressions. The exp-Graph framework is composed of two primary steps: (I) encoding facial attributes through graph-based representation and (II) facial expression recognition through graph convolutional networks. Part (I) is further divided into three subparts: (A) image pre-processing, (B) key points and feature extraction, and (C) graph representation. Part (II) is based on the GCNS model training and facial expression classification. [Best shown in color]

them. The final step applies a threshold to filter out weak connections, leaving only the most significant relationships between landmarks. This refined adjacency matrix can then be used for facial expression recognition for graph construction as shown in the Algorithm 1.

**Algorithm 1** Graph Representation

**Input:** Image $\mathbf{I}$
**Output:** Graph $\mathcal{G} = (\boldsymbol{P}, \boldsymbol{X}, \boldsymbol{A})$
Face $\leftarrow$ Dlib($\mathbf{I}$) {Detect face in the image}
$\boldsymbol{P} \leftarrow$ ExtractLandmarks(Face) {Landmark coordinates}
**for** each landmark $\mathbf{p}_i$ in $\boldsymbol{P}$ **do**
  $\mathbf{I}_i \leftarrow$ ExtractPatch($\mathbf{I}, \mathbf{p}_i$) {Extract patch around landmark}
  $\mathbf{x}_i \leftarrow$ ViT($\mathbf{I}_i$) {Compute feature vector}
  Append $\mathbf{x}_i$ to $\boldsymbol{X}$
**end for**
Compute the adjacency matrix $\boldsymbol{A}$ using Eq. (1) & (2).
**return** $\mathcal{G} = (\boldsymbol{P}, \boldsymbol{X}, \boldsymbol{A})$

### *B. Facial Expression Recognition Through Graph Convolutional Networks*

The matrix $\boldsymbol{A}$ defines the spatial relationships between facial landmarks, contributing to constructing a graph. We further learn feature embedding for the facial expression recognition using GCNs. The Algorithm 1 generates the set of facial landmarks $\boldsymbol{P}$, the set of graph-based features $\boldsymbol{X}$.

Each layer performs two primary operations:

1) **Node Feature Intergration**: The node features $\boldsymbol{H}^{(l)}$ are integrated based on the graph structure encoded in $\hat{\boldsymbol{A}}$ as in Eq. (3).

$$\hat{\boldsymbol{A}} \leftarrow \tilde{\boldsymbol{D}}^{-1/2}(\boldsymbol{A} + \boldsymbol{I})\tilde{\boldsymbol{D}}^{-1/2}. \tag{3}$$

where $\tilde{\boldsymbol{D}}$ is the degree matrix of $\boldsymbol{A} + \boldsymbol{I}$.

2) **Node Latent Feature Projection**: The integrated features are then projected through a learnable weight matrix $\boldsymbol{W}^{(l)}$ and followed by a non-linear activation function $\sigma(\cdot)$ which essential for capturing complex patterns and interactions in the data, allowing the network

to learn more expressive and discriminative features in Eq. (4).

$$\boldsymbol{H}^{(l+1)} \leftarrow \sigma\left(\hat{\boldsymbol{A}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right). \tag{4}$$

where $\boldsymbol{W}^{(l)}$ is the weight matrix for layer $l$.

**Algorithm 2** Exp-Graph Training Model

**Input:** $\boldsymbol{X} \leftarrow$ feature matrix, $\boldsymbol{A} \leftarrow$ adjacency matrix, $L \leftarrow$ number of layers
**Output:** $\boldsymbol{H}^{(L)} \leftarrow$ node feature matrix
**Initialization:** Set the initial feature matrix $\boldsymbol{H}^{(0)} \leftarrow \boldsymbol{X}$
**Normalize Adjacency Matrix:** Compute the normalized adjacency matrix
$\hat{\boldsymbol{A}} \leftarrow \tilde{\boldsymbol{D}}^{-1/2}(\boldsymbol{A}+\boldsymbol{I})\tilde{\boldsymbol{D}}^{-1/2}$
where $\tilde{\boldsymbol{D}}$ is the degree matrix of $\boldsymbol{A}+\boldsymbol{I}$.
**for** $l = 0$ to $L-1$ **do**
**Feature Propagation:** Update the feature matrix
$\boldsymbol{H}^{(l+1)} \leftarrow \sigma\left(\hat{\boldsymbol{A}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right)$
where $\boldsymbol{W}^{(l)}$ is the weight matrix for layer $l$.
**end for**
**return** $\boldsymbol{H}^{(L)}$

The Exp-Graph is trained using the cross-entropy loss:

$$\mathcal{L}(Y,\hat{Y}) = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C} y_{ji}\log(\hat{y}_{ji}) \tag{5}$$

where $\mathcal{L}(Y,\hat{Y})$ denotes the loss function measuring the dissimilarity between the true labels $Y$ and the predicted probabilities $\hat{Y}$. $N$ is the total number of instances (samples) in the dataset. $C$ is the number of classes. $Y = [y_{j,i}] \in \{0,1\}^{N\times C}$ is the true label matrix, where $y_{j,i} = 1$ if the $j^{\text{th}}$ instance belongs to class $i$, and $y_{j,i} = 0$ otherwise. $\hat{Y} = [\hat{y}_{j,i}] \in [0,1]^{N\times C}$ is the predicted probability matrix, where $\hat{y}_{j,i}$ is the predicted probability that instance $j$ belongs to class $i$. These probabilities are typically obtained via a softmax output layer in multi-class classification.

## IV. Experimental Evaluation

This section presents the experimental setup used to evaluate the effectiveness of the proposed approach and compares its performance against existing state-of-the-art (SOTA) methods. We compared our Exp-Graph method with recent and SOTA methods such as DTAGN(Joint) [74], PPDN [75], GCNet [76], FN2EN [77], DeRL [78], SASE-FE [79], ArcFace + lmrk [22], ATT [45], STGCN+AM+PF [40], AT-ViG [37] on the Oulu-CASIA [49] dataset. For eNTERFACE05 [50] dataset, we compared with methods such as Mansoorizadeh*et al.* [80], Zhalehpour *et al.* [81], Vnet [82]. For the AFEW [51] dataset, we compared with methods such as HoloNet [83], Emotiw2018 [51], DSAN-VGG [84], DGNN [46]. To explore the generalizability of our method, we conducted extensive experiments with the standard benchmark datasets used to evaluate FER using the Exp-Graph method. The datasets and the details used in our experiments are as follows Oulu-CASIA, eNTERFACE05 and AFEW.

Face detection was performed using Dlib [73], while the deep learning models were implemented using PyTorch 2.1.2 with CUDA 12.8 support. All experiments were conducted on an NVIDIA RTX A6000 GPU equipped with 48 GB of memory. In our implementation, the hyperparameters are configured as follows: input images are resized to 224×224 pixels. The training begins with an initial learning rate of 0.001, which is gradually reduced to a minimum of 1e-4. A weight decay of 5e-4 is applied to prevent overfitting. The Adam optimizer is employed for optimization, and various activation functions—ReLU, GeLU, and ELU—are explored. The model architecture includes a hidden layer with 256 units and a dropout rate of 0.2 to enhance generalization. To ensure the reproducibility of results, a fixed random seed of 1000 is used.

TABLE I
Details of Facial Expression Datasets

| Dataset | Oulu | eNTERFACE05 | AFEW |
|---|---|---|---|
| **Resolution** | 320 X 240 | 224 X 224 | 224 X 224 |
| **# Expression** | 6 | 6 | 7 |
| **# Subjects** | 80 | 44 | in-the-wild |
| **Modality** | visual | visual-audio | visual |

Our study used datasets summarized in Table I. The Oulu-CASIA dataset [49], with a resolution of 320x240 pixels and a frame rate of 25 frames per second, captures expressions under three lighting conditions: normal, weak, and dark. The eNTERFACE05 dataset [50] is significantly larger, containing over 1,200 video sequences from 44 subjects, and is widely used for multi-modality (visual-audio) facial expression recognition and video-based technique evaluation, with each sequence lasting about four seconds and consisting of approximately 120 frames. The AFEW dataset [51], used in the EmotiW challenge, is a popular video-based FER dataset in the wild, sourced from various television shows and films, presenting challenges such as varying head poses, lighting, and occlusions.

TABLE II
Comparison of the Accuracy (%) of Exp-Graph with the SOTA Methods on Oulu-CASIA Dataset

| Method | Accuracy (%) | Info. |
|---|---|---|
| DTAGN(Joint) [74] | 81.46 | GA+GC |
| PPDN [75] | 84.59 | GA |
| GCNet [76] | 86.11 | GA |
| FN2EN [77] | 87.71 | GA |
| DeRL [78] | 88.0 | GA |
| SASE-FE [79] | 89.6 | GA+LA |
| ArcFace + lmrk [22] | 90.28 | GC+LA |
| ATT [45] | 89.03 | GC |
| STGCN+AM+PF [40] | 90.05 | GC |
| AT-ViG [37] | 92.03 | GC |
| **Exp-Graph (GeLU)** | **98.09** | GC+LA |
| **Exp-Graph (ELU)** | **98.09** | GC+LA |
| **Exp-Graph (ReLU)** | **98.09** | GC+LA |

TABLE III
COMPARISON OF THE ACCURACY (%) OF EXP-GRAPH WITH THE SOTA METHODS ON ENTERFACE05 DATASET

| Method | Accuracy (%) | Info. |
|---|---|---|
| Mansoorizadeh*et al.* [80] | 37.00 | GC |
| Zhalehpour *et al.* [81] | 42.12 | GA |
| Vnet [82] | 54.35 | GA |
| **Exp-Graph (ReLU)** | **79.01** | GC+LA |
| **Exp-Graph (GeLU)** | **79.01** | GC+LA |
| **Exp-Graph (ELU)** | **79.01** | GC+LA |

TABLE IV
COMPARISON OF THE ACCURACY (%) OF EXP-GRAPH WITH THE SOTA METHODS ON AFEW DATASET

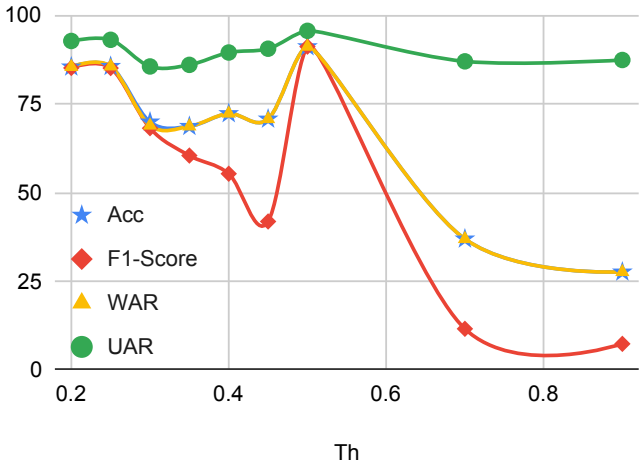
| Method | Accuracy (%) | Info. |
|---|---|---|
| HoloNet [83] | 38.81 | GA |
| Emotiw2018 [51] | 38.81 | GA |
| DSAN-VGG [84] | 52.74 | GA |
| DGNN [46] | 32.64 | GA+LA |
| **Exp-Graph (ReLU)** | **56.39** | GC+LA |
| **Exp-Graph (GeLU)** | **56.39** | GC+LA |
| **Exp-Graph (ELU)** | **56.39** | GC+LA |

Table II, III, IV compare the accuracy of various SOTA methods on Oulu-CASIA, eNTERFACE05 and AFEW datasets, respectively. In particular, our proposed Exp-Graph achieves the highest accuracy of 98.09%, 79.01%, and 56.39% on the Oulu-CASIA, eNTERFACE05, and AFEW datasets, respectively. A comparison of the Exp-Graph with the ensuing methods for facial expression recognition for accuracy, UAR (Unweighted Average Recall), WAR (Weighted Average Recall), cross-entropy loss, and F1-score are presented as metrics for the evaluation. Also, we present results using different thresholds (Th = 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.70, 0.90) and patch-size ( 10x10, 20x20, 30x30, 50x50, 70x70, 90x90). Also use the ResNet18 [85] and EfficientNetB0 [23] base model combined with the GAT [86] and GCNs for further exploration in our research with Exp-Graph.

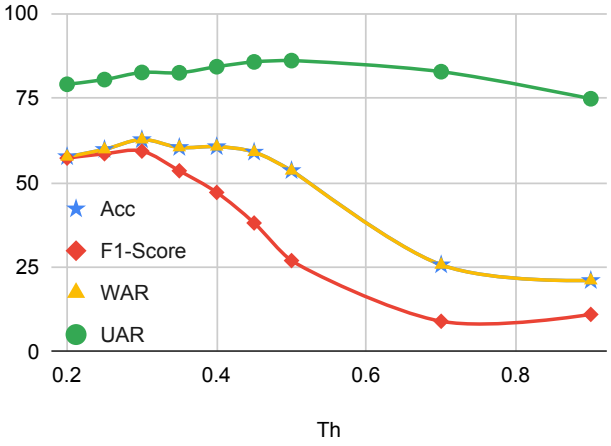
TABLE V
TEST PERFORMANCE METRICS FOR THE DATASETS

| Metrics | Th0.30 | Th0.50 | Th0.70 | Th0.90 | Dataset |
|---|---|---|---|---|---|
| Loss | 1.34 | 1.21 | 1.65 | 1.74 | Oulu |
| F1 (%) | 68.11 | 84.00 | 11.58 | 07.34 | Oulu |
| WAR (%) | 69.88 | 84.00 | 36.91 | 27.60 | Oulu |
| UAR (%) | 85.51 | 92.00 | 86.95 | 87.33 | Oulu |
| Loss | 1.34 | 1.21 | 1.65 | 1.74 | eNTER |
| F1 (%) | 68.11 | 84.00 | 11.58 | 07.34 | eNTER |
| WAR (%) | 69.88 | 84.00 | 36.91 | 27.60 | eNTER |
| UAR (%) | 85.51 | 92.00 | 86.95 | 87.33 | eNTER |
| Loss | 1.92 | 1.50 | 1.74 | 1.77 | AFEW |
| F1 (%) | 59.27 | 26.95 | 9.03 | 11.06 | AFEW |
| WAR (%) | 62.67 | 53.54 | 25.71 | 21.08 | AFEW |
| UAR (%) | 82.50 | 86.00 | 82.73 | 74.74 | AFEW |

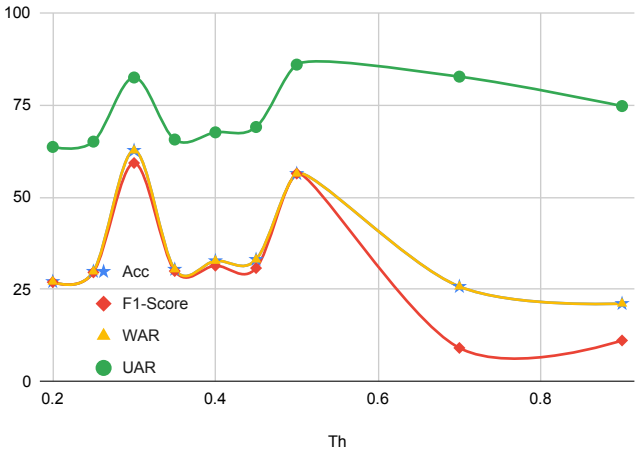
Table V presents the results of analyzing the effect of varying threshold values ($\tau$) on model performance, as illustrated in Figs. 6a, 6b, and 6c. The study evaluates the ViT+GCNs model across three benchmark datasets: Oulu-CASIA, eNTERFACE05, and AFEW. The results show that a threshold of 0.50 consistently yields the best overall performance, particularly in the Oulu-CASIA dataset, where the model achieves an F1 score (84%), a WAR (84%), and a UAR (92%). Performance degrades significantly at higher threshold




(a) Oulu-CASIA dataset




(b) eNTERFACE05 dataset




(c) AFEW dataset

Fig. 6. Test results across thresholds (Th = $\tau$) for three datasets. [Best shown in color]

values, with $\tau = 0.90$ resulting in the poorest results in all metrics on the Oulu-CASIA dataset. A similar trend is observed in the eNTERFACE05 and AFEW datasets, although the threshold of 0.30 yields the worst results in these cases. Notably, the threshold of 0.90 is consistently suboptimal across all datasets, indicating its tendency to introduce excessive sparsity in the relational graphs, which negatively impacts model learning and prediction stability. This configuration benefits from the synergistic use of global appearance (GA) and local appearance (LA) features, outperforming models that rely solely on geometric or individual feature types. The enhanced results can be attributed to the proposed model's ability to construct and utilize expressive graph representations that capture nuanced structural relationships between facial landmarks.

This inverse relationship between threshold value and performance can be attributed to the increased sparsity and potential noise in the constructed graph at higher thresholds, which results in less informative node connections. Lower thresholds preserve more connections, which are beneficial for learning discriminative patterns, particularly in imbalanced datasets. The accompanying figures— 6a, 6b, and 6c—visually reinforce these findings, showing a clear peak in accuracy and F1 Score around $\tau = 0.50$, followed by a noticeable decline as the threshold increases. While WAR and UAR display slightly more stable trends, they also show reduced performance at higher thresholds. These consistent patterns across all datasets highlight the ViT+GCNs model's robustness and adaptability, particularly its capacity to handle class imbalance effectively when configured with an appropriately tuned threshold.

Fig. 7 and Table VI present the results comparing the performance of various model architectures on the Oulu-CASIA dataset. Among the evaluated models, the proposed Exp-Graph (ViT+GCNs) configuration consistently achieves the highest performance across the evaluation metrics. These results indicate the model's superior capacity for accurate classification and its robustness in addressing class imbalance.

In contrast, models such as EfficientNet+GCNs, ResNet18+GAT, and EfficientNet+GAT demonstrate moderate performance, achieving results that are relatively close to each other but noticeably lower than those of ViT+GCNs. The ViT+GAT configuration yields the weakest performance across all metrics, suggesting that the combination of graph attention mechanisms with vision transformers may not be well-suited to this dataset or task without further architectural refinements. The consistently strong performance of ViT+GCNs highlights the effectiveness of integrating graph convolutional networks with vision transformers, leveraging both spatial relational structure and high-capacity feature extraction. This comprehensive approach to feature extraction and graph-based modeling significantly improves recognition performance, demonstrating the model's robustness and generalizability on complex facial expressions recognition tasks.

Fig. 8a and Fig. 8b present the t-SNE visualizations of learned feature representations on the Oulu-CASIA and eNTERFACE05 datasets, respectively. The plots were generated by evaluating the model's output at various graph threshold values, specifically $\tau \in 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.70, 0.90$.

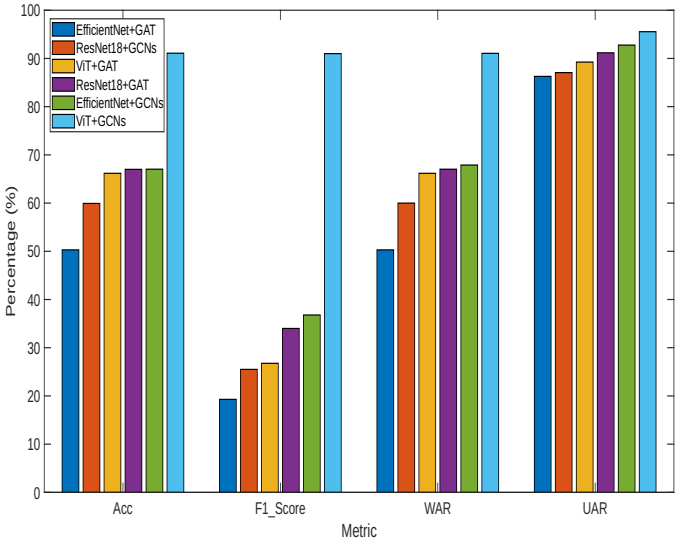


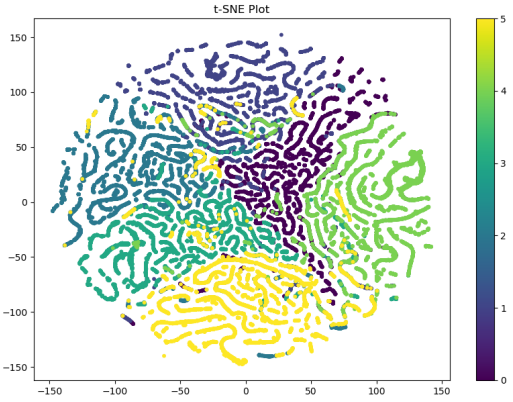

Fig. 7. Test result on the Oulu-CASIA dataset across models. [Best shown in color]

TABLE VI
TEST RESULTS ON THE DIFFERENT MODELS' PERFORMANCE METRICS ON OULU-CASIA

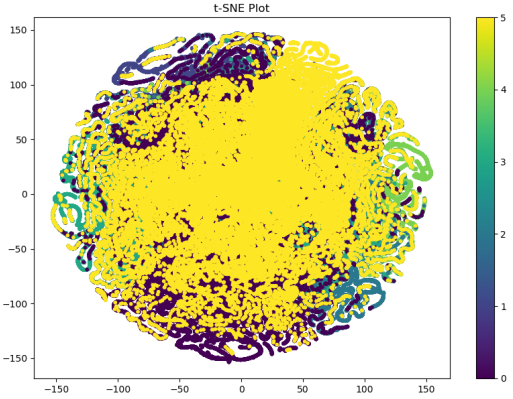
| Model | Acc. | F1-Score | WAR | UAR |
|---|---|---|---|---|
| ResNet18+GCNs | 59.93 | 34.00 | 60 | 87.06 |
| EfficientNet+GCNs | 67.03 | 36.79 | 67.03 | 89.25 |
| ViT+GCNs | 91.09 | 91.00 | 91.09 | 95.55 |
| ResNet18+GAT | 67.00 | 26.77 | 67.89 | 92.77 |
| EfficientNet+GAT | 50.30 | 19.29 | 50.30 | 86.28 |
| ViT+GAT | 66.18 | 25.51 | 66.18 | 91.17 |

Among these, thresholds of 0.20 and 0.30 consistently produced the most distinct and well-separated clusters, indicating effective discrimination between different facial expression classes. Interestingly, the threshold of 0.30 offered good separation, it also resulted in overly dispersed clusters, potentially obscuring the underlying relational structure among expressions. The most balanced and interpretable visualizations were obtained at a threshold of 0.50, where the t-SNE plots showed both clear inter-class separation and coherent intra-class clustering. This suggests that the threshold of 0.20 in the Oulu-CASIA effectively preserves both local and global topological structures within the learned graph representations. The clarity of the clustering at this threshold, particularly when using features extracted from the final layer of the GCNs, underscores its ability to capture complex patterns in facial expression data, thus supporting more accurate and meaningful interpretation of the learned embeddings.

Fig. 9 illustrates the Visualization of the learned graph for the Oulu-CASIA dataset sample images at a threshold value of $\tau = 0.30$. In this visualization, the first column displays samples of different facial expressions (e.g., anger, disgust, fear), while the second column overlays the corresponding connection graph on the original image, highlighting the relational structure among key facial landmarks. The performance metrics of the proposed model across varying threshold values




(a) Oulu-CASIA Dataset (Threshold = 0.20)




(b) eNTERFACE05 Dataset (Threshold = 0.30)

Fig. 8. t-SNE test results on two FER datasets. [Best shown in color]


Fig. 9. Visualization of learned graphs for the Oulu-CASIA dataset sample images for $\tau$ = 0.30. [Best shown in color]

TABLE VII
PERFORMANCE METRICS FOR DIFFERENT THRESHOLDS ON THE OULU-CASIA

| Th = $\tau$ | Acc | F1-Score | WAR | UAR |
|---|---|---|---|---|
| 0.20 | 85.38 | 85.09 | 85.4 | 92.69 |
| 0.25 | 85.54 | 85.06 | 85.54 | 93.02 |
| 0.30 | 69.88 | 68.11 | 68.88 | 85.51 |
| 0.35 | 68.64 | 60.4 | 68.64 | 85.96 |
| 0.40 | 72.24 | 55.28 | 72.25 | 89.44 |
| 0.45 | 70.7 | 41.82 | 70.7 | 90.51 |
| 0.50 | 91.09 | 91.1 | 91.1 | 95.55 |
| 0.70 | 36.91 | 11.58 | 36.9 | 86.95 |
| 0.90 | 27.6 | 7.34 | 27.6 | 87.33 |

TABLE VIII
PERFORMANCE METRICS FOR DIFFERENT THRESHOLDS ON THE ENTERFACE05

| Th = $\tau$ | Acc | F1-Score | WAR | UAR |
|---|---|---|---|---|
| 0.20 | 57.65 | 57.2 | 57.62 | 79.01 |
| 0.25 | 59.77 | 58.5 | 59.77 | 80.4 |
| 0.30 | 62.67 | 59.27 | 62.7 | 82.5 |
| 0.35 | 60.37 | 53.47 | 60.37 | 82.41 |
| 0.40 | 60.6 | 47.07 | 60.6 | 84.2 |
| 0.45 | 58.98 | 38.05 | 58.98 | 85.62 |
| 0.50 | 53.54 | 26.95 | 53.5 | 86 |
| 0.70 | 25.71 | 9.03 | 25.7 | 82.73 |
| 0.90 | 21.08 | 11.06 | 21.08 | 74.74 |

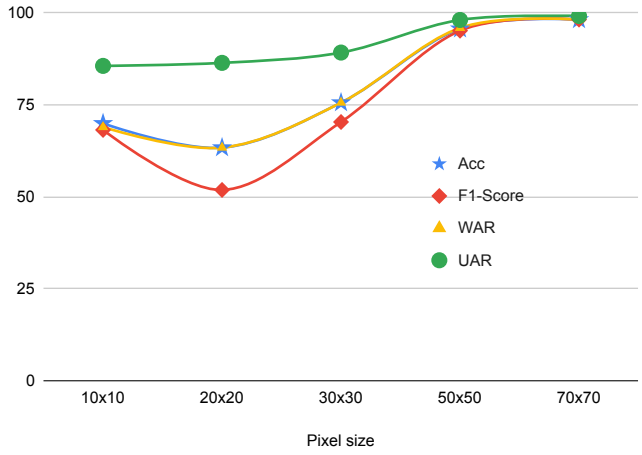
are reported in Tables VII, VIII, and IX, offering insight into the threshold's influence on model effectiveness across different datasets.

Table VII, which presents the performance without data augmentation on the Oulu-CASIA dataset, the model achieves its best performance at $\tau = 0.50$, with an accuracy of 91.09%, F1-Score of 91.1%. As the threshold increases (e.g., $\tau = 0.70$ or $\tau = 0.90$), both accuracy and F1-Score degrade noticeably, indicating that excessive sparsity in the graph structure adversely affects learning and classification performance. Similarly, Table VIII presents evaluation results on the eNTERFACE05 dataset, demonstrates a similar trend. The highest metrics are recorded around $\tau = 0.30$, with an accuracy of 62.67% and an F1-Score of 59.27%, reaffirming the effectiveness of lower thresholds for preserving discriminative information in the relational graph. Finally, Table IX summarizes the model's performance

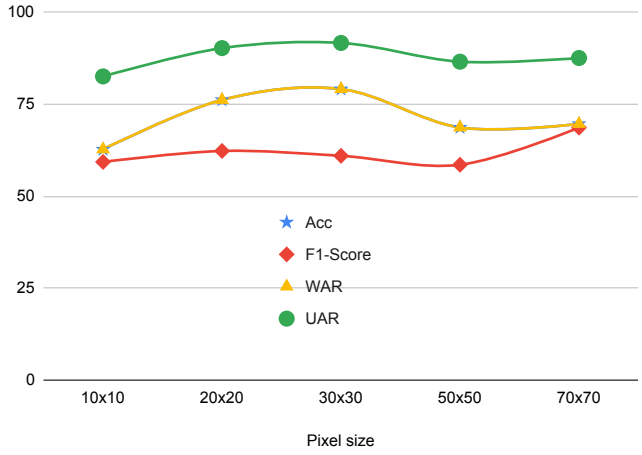
on the AFEW dataset, where the optimal results are also observed at $\tau = 0.30$. These findings collectively highlight the importance of selecting an appropriate threshold value, as

TABLE IX
PERFORMANCE METRICS FOR DIFFERENT THRESHOLDS ON THE AFEW

| Th = $\tau$ | Acc | F1-Score | WAR | UAR |
|---|---|---|---|---|
| 0.20 | 27.04 | 26.9 | 27.04 | 63.64 |
| 0.25 | 29.81 | 29.56 | 29.81 | 65.08 |
| 0.30 | 62.67 | 59.27 | 62.67 | 82.5 |
| 0.35 | 30.37 | 29.92 | 30.37 | 65.65 |
| 0.40 | 32.69 | 31.38 | 32.69 | 67.6 |
| 0.45 | 33.05 | 30.72 | 33.05 | 69.05 |
| 0.50 | 56.39 | 56.39 | 56.39 | 86 |
| 0.70 | 25.71 | 9.03 | 25.71 | 82.73 |
| 0.90 | 21.08 | 11.06 | 21.08 | 74.74 |




(a) Oulu-CASIA dataset (Threshold = 0.3)




(b) eNTERFACE05 dataset (Threshold = 0.3)

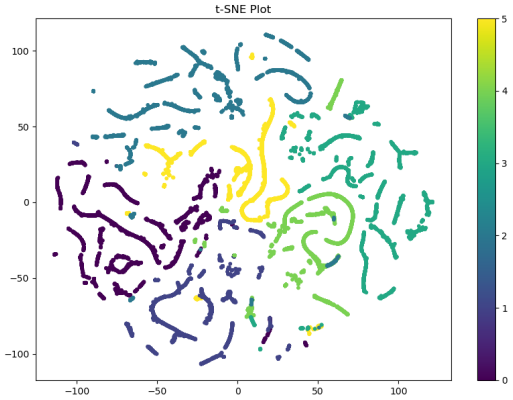
Fig. 10. Test results across patch sizes for two FER datasets at a threshold of 0.3. [Best shown in color]

TABLE X
PERFORMANCE METRICS FOR DIFFERENT PATCH SIZES ON THE OULU

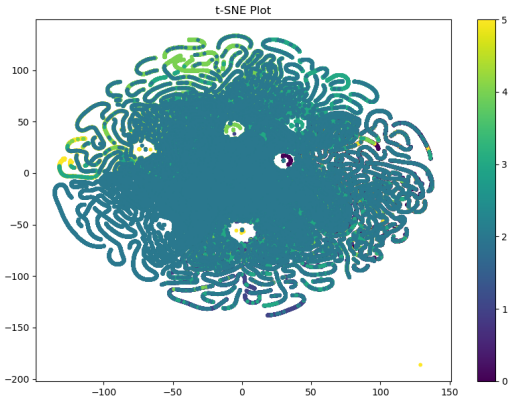
| HxW | Acc | F1-Score | WAR | UAR |
|---|---|---|---|---|
| 10x10 | 69.88 | 68.11 | 68.88 | 85.51 |
| 20x20 | 63.30 | 51.81 | 63.30 | 86.33 |
| 30x30 | 75.52 | 70.27 | 75.52 | 89.13 |
| 50x50 | 95.58 | 95.07 | 95.58 | 97.99 |
| **70x70** | **98.09** | **98.09** | **98.09** | **99.06** |

TABLE XI
PERFORMANCE METRICS FOR DIFFERENT PATCH SIZES ON THE ENTERFACE05

| HxW | Acc | F1-Score | WAR | UAR |
|---|---|---|---|---|
| 10x10 | 62.67 | 59.27 | 62.7 | 82.5 |
| 20x20 | 76.12 | 62.25 | 76.12 | 90.19 |
| **30x30** | **79.06** | **60.93** | **79.06** | **91.59** |
| 50x50 | 68.63 | 58.50 | 68.63 | 86.50 |
| 70x70 | 69.60 | 68.59 | 69.60 | 87.50 |




(a) t-SNE Test Results on Oulu-CASIA Dataset on the best Patch size.




(b) t-SNE Test Results on eNTERFACE05 Dataset on the best Patch size.

Fig. 11. t-SNE test results on the best patch size for two datasets. [Best shown in color]

it plays a critical role in balancing graph connectivity and expressiveness, ultimately influencing the model's ability to accurately capture and classify subtle facial expressions across different datasets.

As shown in Table X and Fig. 10a, the performance on the Oulu-CASIA dataset exhibits a positive correlation with increasing patch size. The model achieves its best performance at a patch size of 70×70, obtaining the highest accuracy (98.09%), F1-score (98.09%), WAR (98.09%), and UAR (99.06%). A noticeable performance improvement begins at the 30×30 patch size, continuing to rise steadily with larger patches, and peaking at 70×70. In contrast, smaller patches

such as 10×10 and 20×20 result in significantly lower performance across all metrics, with 20×20 yielding the weakest results, suggesting that smaller patches may fail to capture sufficient contextual and spatial information.

For the eNTERFACE05 dataset, results reported in Table XI and visualized in Fig. 10b, the optimal performance is observed at a patch size of 30×30, where the model achieves its highest values for accuracy (79.06%), F1-score (60.93%), WAR (79.06%), and UAR (91.59%). Performance declines for both smaller and larger patch sizes. While 20×20 still performs reasonably well, 10×10 and 50×50 show a marked drop in effectiveness, and the 70×70 patch—which was optimal for the Oulu dataset—fails to deliver comparable results on eNTERFACE05. These differences highlight the dataset-specific sensitivity to patch size and the importance of tailoring patch-based feature extraction to the characteristics of each dataset. Additionally, 10b presents t-SNE visualizations of the test samples using the optimal patch sizes for each dataset. These plots demonstrate clear and distinct class separations, further validating the effectiveness of the selected patch sizes in preserving discriminative features and supporting accurate classification.

Overall, these results highlight that the optimal patch size varies by dataset, with larger patches proving more effective on Oulu, while a mid-sized patch (30×30) yields the best performance on eNTERFACE05. We also demonstrate how varying the threshold impacts the model's training and validation performance, recall metrics, and feature detection capabilities. Lower thresholds generally lead to better accuracy, reduced loss, and improved recall, while enabling more comprehensive facial feature analysis.

## V. Conclusion

In conclusion, this work introduced Exp-Graph, a novel framework for facial expression recognition that utilizes a graph-based representation of facial attribute structures. By modeling facial landmarks as graph nodes and defining edges through spatial and appearance-based relationships, Exp-Graph captures intricate dependencies among facial features. Selecting an appropriate patch size and threshold ($\tau$) is crucial for the optimal performance of the Exp-Graph. An appropriate patch size with a suitable threshold can reduce information loss and result in a more relevant graph representation. In contrast, too large or too small patch sizes and thresholds may result in either similar graph structures or disconnected graphs due to the loss of important information. Therefore, determining the optimal threshold and patch size is essential for effectively preserving facial features and ensuring robust performance. However, the optimal size may vary depending on the characteristics of the dataset. The integration of vision transformers and graph convolutional networks enables the framework to effectively encode global context and structural information. The experimental results confirm the robustness and strong generalization of the method across the datasets.

## References


[1] X. Xie and K.-M. Lam, "Gabor-based kernel pca with doubly nonlinear mapping for face recognition with a single face image," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2481–2492, 2006.

[2] 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.

[3] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[4] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 878–883.

[5] C. F. Liew and T. Yairi, "Facial expression recognition and analysis: a comparison study of feature descriptors," *IPSJ transactions on computer vision and applications*, vol. 7, pp. 104–120, 2015.

[6] 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.

[7] H. Wang, S. Wei, and B. Fang, "Facial expression recognition using iterative fusion of mo-hog and deep features," *The Journal of Supercomputing*, vol. 76, no. 5, pp. 3211–3221, 2020.

[8] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 143–10 152.

[9] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam, "Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2751–2767, 2022.

[10] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 234–14 243.

[11] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[12] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, proceedings, part VII 14*. Springer, 2016, pp. 499–515.

[13] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2852–2861.

[14] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 302–309.

[15] Y. Li, G. Lu, J. Li, Z. Zhang, and D. Zhang, "Facial expression recognition in the wild using multi-level features and attention mechanisms," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 451–462, 2020.

[16] R. Zhao, T. Liu, J. Xiao, D. P. Lun, and K.-M. Lam, "Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing," in *2020 25th Proceedings of the International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4412–4419.

[17] D. Singh and C. K. Mohan, "Graph formulation of video activities for abnormal activity recognition," *Pattern Recognition*, vol. 65, pp. 265–272, 2017.

[18] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "Graphlime: Local interpretable model explanations for graph neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6968–6972, 2022.

[19] D. Singh, "Graph representation for weakly-supervised spatio-temporal action detection," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–9.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] M. Li, S. Tu, and S. u. Rehman, "Facial expression recognition from occluded images using deep convolution neural network with vision transformer," in *International Conference on Image and Graphics*. Springer, 2023, pp. 289–299.

[22] T. Liu, J. Li, J. Wu, B. Du, J. Chang, and Y. Liu, "Facial expression recognition on the high aggregation subgraphs," *IEEE Transactions on Image Processing*, 2023.

[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[24] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020.

[25] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, 2021.

[26] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4445–4460, 2020.

[27] B. Luo, H. Wang, J. Wang, J. Zhu, X. Zhao, and Y. Gao, "Hypergraph-guided disentangled spectrum transformer networks for near-infrared facial expression recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 9, 2024, pp. 10 101–10 109.

[28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[29] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[30] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2014.

[31] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE transactions on image processing*, vol. 16, no. 1, pp. 172–187, 2006.

[32] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam, "Geometry-aware facial expression recognition via attentive graph convolutional networks," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1159–1174, 2021.

[33] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12,Proceedings, Part VI 13*. Springer, 2014, pp. 94–108.

[34] X. Jin, X. Song, X. Wu, and W. Yan, "Transformer embedded spectral-based graph network for facial expression recognition," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 6, pp. 2063–2077, 2024.

[35] C. Xu, Y. Du, J. Wang, W. Zheng, T. Li, and Z. Yuan, "A joint hierarchical cross-attention graph convolutional network for multi-modal facial expression recognition," *Computational Intelligence*, vol. 40, no. 1, p. e12607, 2024.

[36] S. Liu, S. Huang, W. Fu, and J. C.-W. Lin, "A descriptive human visual cognitive strategy using graph neural network for facial expression recognition," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 1, pp. 19–35, 2024.

[37] W. Dong, X. Zheng, L. Zhang, and Y. Zhang, "Attentional visual graph neural network based facial expression recognition method," *Signal, Image and Video Processing*, vol. 18, no. 12, pp. 8693–8705, 2024.

[38] H. F. T. Al-Saadawi and R. Das, "Ter-ca-wgnn: trimodel emotion recognition using cumulative attribute-weighted graph neural network," *Applied Sciences*, vol. 14, no. 6, p. 2252, 2024.

[39] S. Mao, X. Li, F. Zhang, X. Peng, and Y. Yang, "Facial action units as a joint dataset training bridge for facial expression recognition," *IEEE Transactions on Multimedia*, 2025.

[40] Y. Qu and Y. Liu, "Design and research of facial expression recognition system based on key point extraction." *KSII Transactions on Internet & Information Systems*, vol. 19, no. 1, 2025.

[41] C. Huang, F. Jiang, Z. Han, X. Huang, S. Wang, Y. Zhu, Y. Jiang, and B. Hu, "Modeling fine-grained relations in dynamic space-time graphs for video-based facial expression recognition," *IEEE Transactions on Affective Computing*, 2025.

[42] C. Tanchotsrinon, S. Phimoltares, and S. Maneeroj, "Facial expression recognition using graph-based features and artificial neural networks," in *2011 IEEE International Conference on Imaging Systems and Techniques*. IEEE, 2011, pp. 331–334.

[43] H. Kassab, M. Bahaa, and A. Hamdi, "Gcf: Graph convolutional networks for facial expression recognition," in *2024 Intelligent Methods, Systems, and Applications (IMSA)*. IEEE, 2024, pp. 166–171.

[44] X. Xu, Z. Ruan, and L. Yang, "Facial expression recognition based on graph neural network," in *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2020, pp. 211–214.

[45] J. Chen, J. Shi, and R. Xu, "Dual subspace manifold learning based on gcn for intensity-invariant facial expression recognition," *Pattern Recognition*, vol. 148, p. 110157, 2024.

[46] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial landmark-based emotion recognition via directed graph neural network," *Electronics*, vol. 9, no. 5, p. 764, 2020.

[47] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7,proceedings 15*. Springer, 2018, pp. 593–607.

[48] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7370–7377.

[49] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and vision computing*, vol. 29, no. 9, pp. 607–619, 2011.

[50] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proceedings of the 22nd IEEE International Conference on Data Engineering Workshops (ICDEW)*. IEEE, 2006, pp. 8–8.

[51] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 653–656.

[52] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 627–635.

[53] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, "Mer-gcn: Micro-expression recognition based on relation modeling with graph convolutional networks," in *Proceedings of the IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2020, pp. 79–84.

[54] Y. Liu, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via deep action units graph network based on psychological mechanism," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 311–322, 2019.

[55] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive gan for facial expression transfer," *arXiv preprint arXiv 1802.01822*, 2018.

[56] R. Bodur, B. Bhattarai, and T.-K. Kim, "3d dense geometry-guided facial expression synthesis by adversarial learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Winter*, 2021, pp. 2392–2401.

[57] S. Mo, W. Yang, G. Wang, and Q. Liao, "Emotion recognition with facial landmark heatmaps," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, Proceedings, Part I 26*. Springer, 2020, pp. 278–289.

[58] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Proceedings of the IEEE CVPRW*, 2017, pp. 30–40.

[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] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*. IEEE, 1998, pp. 454–459.

[61] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *2014 Canadian conference on computer and robot vision*. IEEE, 2014, pp. 98–103.

[62] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," *arXiv preprint arXiv:1802.06664*, 2018.

[63] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognition*, vol. 48, no. 10, pp. 3191–3202, 2015.

[64] W. Sun, H. Zhao, and Z. Jin, "A visual attention based roi detection method for facial expression recognition," *Neurocomputing*, vol. 296, pp. 12–22, 2018.

[65] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th Proceedings of the International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2209–2214.

[66] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3146–3155.

[67] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[68] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 558–567.

[69] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[70] C. Zhuang and Q. Ma, "Dual graph convolutional networks for graph-based semi-supervised classification," in *Proceedings of the ACM Web Conference*, 2018, pp. 499–508.

[71] J. Zhou, X. Zhang, Y. Liu, and X. Lan, "Facial expression recognition using spatial-temporal semantic graph network," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1961–1965.

[72] J. Zhou, X. Zhang, and Y. Liu, "Learning the connectivity: Situational graph convolution network for facial expression recognition," in *Proceedings of the IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 230–234.

[73] D. E. King, "dlib c++ library," http://dlib.net/, Accessed: ¡Insert Access Date¿.

[74] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 2983–2991.

[75] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, Proceedings, Part II 14*. Springer, 2016, pp. 425–442.

[76] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," *arXiv preprint arXiv:1703.07140*, 2017.

[77] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 118–126.

[78] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE conference on CVPR*, 2018, pp. 2168–2177.

[79] K. Kulkarni, C. A. Corneanu, I. Ofodile, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of facial displays of unfelt emotions," *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 377–390, 2018.

[80] M. Mansoorizadeh and N. Moghaddam Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools and Applications*, vol. 49, no. 2, pp. 277–297, 2010.

[81] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "Baum-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2016.

[82] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE transactions on circuits and systems for video technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[83] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "Holonet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 472–478.

[84] Y. Fan, V. O. Li, and J. C. Lam, "Facial expression recognition with deeply-supervised attention network," *IEEE transactions on affective computing*, vol. 13, no. 2, pp. 1057–1071, 2020.

[85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

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