Para-X: Graph-based Facial Paralysis Detection using Structural Deformations of Facial Expression

Nandani Sharma, Kajal Singh, Dinesh Singh
Vision Intelligence and Machine Learning (VIML) Group
School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, India.
{d22180, s23083}@students.iitmandi.ac.in, dineshsingh@iitmandi.ac.in

***Abstract*—Facial paralysis, marked by the inability to move specific facial muscles, often results from nerve damage, strokes, or neurological disorders. Prompt and accurate detection is essential for effective diagnosis and treatment, potentially improving recovery outcomes. Recent studies identified facial paralysis by analyzing alterations in facial expressions of affected individuals relative to those of unaffected individuals, utilizing facial attributes and landmark data; nevertheless, they overlooked the structural information among the diverse face features. This work utilizes structural information to offer a graphical depiction of facial features. In the graph representation of facial attributes, key-points serve as the vertices. At the same time, the edges are established based on the closeness of the key-points and the similarity of the local appearance of the facial attributes conveyed via the vision transformer. Advanced graph convolutional networks integrate structural information into face attribute encoding to enhance the identification of facial expressions. Consequently, Para-X acquires highly expressive semantic representations from facial attribute graphs. In contrast, the vision transformer and graph convolutional blocks enable the framework to leverage local and global dependencies among facial attributes, which are crucial for recognizing facial paralysis. Comprehensive studies demonstrate the robustness and generalizability of the proposed methodology to identify facial paralysis in various facial paralysis datasets such as AFLFP, YFPD, FDPDI, and our FPD dataset.**

***Index Terms*—Graph convolutional networks, facial paralysis detection, and vision transformer.**

I. Introduction

Facial paralysis is a medical condition often caused by infections, trauma, or neurological problems and is characterized by a lack of voluntary muscle activity on one side of the face [1]. The estimated annual rate of Bell's syndrome is 6.1 cases per 100,000 in children aged 1–15 years [2]. However, 40 cases per 100,000 patients per year have been reported in adults [3], [4]. The incidence of facial nerve palsy in the pediatric population is relatively low.

Effective treatment and rehabilitation depend on the prompt and precise diagnosis of facial paralysis. Clinical judgments and the subjective opinions of medical practitioners are the mainstays of traditional diagnostic techniques. Since the development of computer vision and deep learning, there has been increasing interest in creating automated methods for facial paralysis recognition. As deep learning techniques have advanced rapidly over the past ten years, significant attempts have been made to investigate discriminative representations of facial images for paralysis recognition using deep neural networks (DNNs). As research evolves, integrating innovative diagnostic tools, AI applications, and patient-centered approaches will likely play a significant role in shaping the future of facial nerve palsy management. Smartphone-based diagnostic systems have been proposed as accessible and practical tools to automate the assessment process [5]. Fig. 1 shows the framework for the proposed facial paralysis recognition (FPR) system using smartphones. Since people use smartphones everywhere. Thus, a mobile application is most desirable. Similarly, artificial intelligence (AI) technologies, particularly convolutional neural networks (CNNs), have demonstrated remarkable potential in providing accurate and efficient clinical outcomes by analyzing complex medical data [6]. In addition, the impact of global events, such as the COVID-19 pandemic, has drawn attention to the relationship between viral infections and facial nerve conditions. Research has revealed a possible association between COVID-19 and facial nerve palsy, emphasizing further investigation of its underlying mechanisms and long-term effects [7]. Innovative methods, such as facial emotion recognition systems, have been employed to infer stress levels and address psychological well-being in patients with facial nerve paralysis. This underscores the importance of comprehensive care that includes both physical and emotional health [8]. Despite significant progress, challenges remain in understanding these patients' long-term outcomes and rehabilitation needs. Recent studies have laid the groundwork for technical and interdisciplinary approaches,

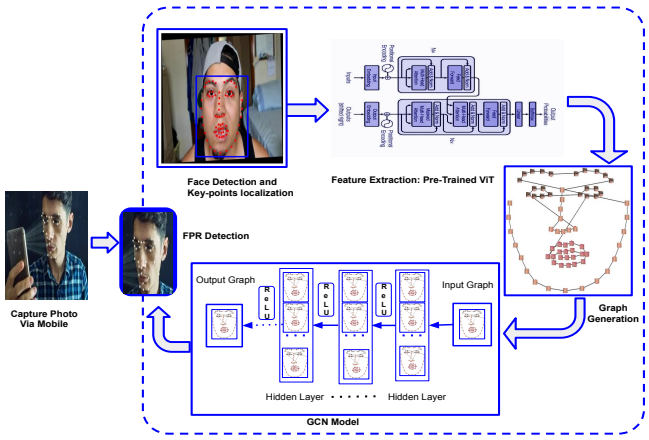

Fig. 1: Framework for proposed FPR systems.

yet shortcomings remain in developing individualized care and rehabilitation strategies [9].

Previous methods to quantitatively analyze facial paralysis include HDN [10], an incremental face alignment framework [11], and Gaussian mixture model (GMM) with a dynamic kernel [12]. However, some methods for facial paralysis detection are not based on facial key-points, and some are based on private datasets. Major anatomical locations on the face, including the mouth, nose, and eyes, are represented by facial key-points, and understanding their spatial relationships is essential to capturing the small asymmetries indicative of facial paralysis [6]. Graph neural networks (GCNs) present a viable path toward creating an accurate and understandable facial paralysis detection system. Zhao *et al.* [13] Spatio-temporal graph networks integrated with a transformer to recognize facial expressions using geometric guidance. Guo *et al.* [14] developed a technique to assess unilateral facial paralysis through facial imaging and historical identification. Valter *et al.* [15] explored machine learning-based methods for diagnosing peripheral and central facial paralysis, taking advantage of facial features. To encourage facial paralysis recognition in real-world tasks, it is highly beneficial to investigate a geometry-guided approachs [10]–[17]. Our work uses geometric information, structural deformities, and comparative alterations of essential parts of the face that are equally susceptible to facial paralysis. Consequently, exploring a geometry-guided approach to encourage FPR in real-world applications is highly beneficial. Graph neural networks [13], [18]–[21] are a potent tool for structured data analysis and are a good fit for applications utilizing linked face key-points [13], [22]–[26]. The capacity of the model to generalize across many datasets such as YFP [10], FDFPI [16], and we also introduce our new dataset facial paralysis detection (FPD), AFLFP [27] for the facial paralysis class. Our new dataset addresses an important limitation observed in previous datasets, which focused primarily on young individuals. A significant drawback of those datasets is the lack of data on children and elderly individuals, making it challenging to analyze facial deformities across all age groups. Our dataset includes individuals of all age groups – children, young adults, and the elderly – making it more comprehensive and challenging. This diverse age representation increases its effectiveness, especially for detecting facial paralysis in children and younger and older populations. For the non-paralysis faces used from the facial expressions recognition (FER) datasets such as DISFA [28], DISFA+ [29], OuluCASIA [30], Caltech face [31], and RML [32] circumstances, while simultaneously increasing the accuracy of automatic paralysis recognition.

In this paper, we present a Para-X framework shown in Fig. 2 that uses GCNs with ViT to learn geometric descriptions from facial key-points. Consequently, GCNs offer a valuable substitute for incorporating the geometric information derived from facial key-points into paralysis and normal facial representations. Local appearance representations are extracted from landmark positions and aggregated with geometric representations during graph learning—an innovative and efficient method that provides several important advances in facial paralysis recognition. In the study [13], [22]–[26], the number of nodes and connections between them are fixed during GCNs training. However, using hyperparameter thresholds τ, our proposed model dynamically learns the connectivity structure based on a weighted adjacency matrix $\boldsymbol{W}$. We conduct extensive experiments using different threshold values to analyze how different connection patterns affect the recognition of paralyzed and non-paralyzed faces. Our study's results have significantly impacted the development of computer-assisted medical diagnosis, providing medical practitioners with a reliable and unbiased tool to evaluate facial paralysis. The following briefly describes the primary contributions of this paper:

- Effective facial representation by exploring the structural information through GCNs on the applied information needed from the pre-trained ViT.
- Learns semantic representations of facial by leveraging local and global dependencies among facial attributes.
- Extensive evaluation of the proposed method Para-X on AFLFP, YFP, FDFPI, and we also introduce our new dataset FPD for the paralysis faces and DISFA, DISFA+, OuluCASIA, Caltech face, RML for the normal circumstances, while increasing automated recognition accuracy.

The remaining content of this paper is structured as follows: Section II briefly reviews some related works. Section III provides a detailed explanation of the Para-X framework. Section IV presents the details of experimental evaluations, including benchmark datasets, and discusses the results. Finally, we conclude in §V.

II. RELATED WORK

The automatic identification and diagnosis of facial palsy have traditionally relied on handcrafted features and private datasets, limiting reproducibility and adaptability to real-world conditions. Early efforts focused on asymmetry detection using tailored algorithms. Ngo *et al.* [33] employed limited-orientation modified circular Gabor filters (LO-MCGFs) to extract facial asymmetry features from the Osaka police hospital dataset. Kim *et al.* [5] developed a smartphone-based diagnostic system that included facial landmark detection, feature extraction, and classification. Similarly, Asthana *et al.* [11] used incremental face alignment to calculate asymmetry indices, applying linear discriminant analysis (LDA) and support vector machines (SVM) for classification. Wang *et al.* [34] used an active shape model (ASM) to analyze static and dynamic asymmetry across eight facial regions, employing an SVM with an RBF kernel for classification. However, the above approaches were limited due to the reliance on hand-made features, which can hinder adaptability to different clinical settings. Praveen *et al.* [12] further explored facial dynamics using Gaussian mixture models (GMMs) with dynamic kernels. However, the inherent limitations of private datasets and the need for generalized, robust approaches underscore the importance of transitioning to deep learning and larger, publicly available datasets. The rise of deep learning techniques

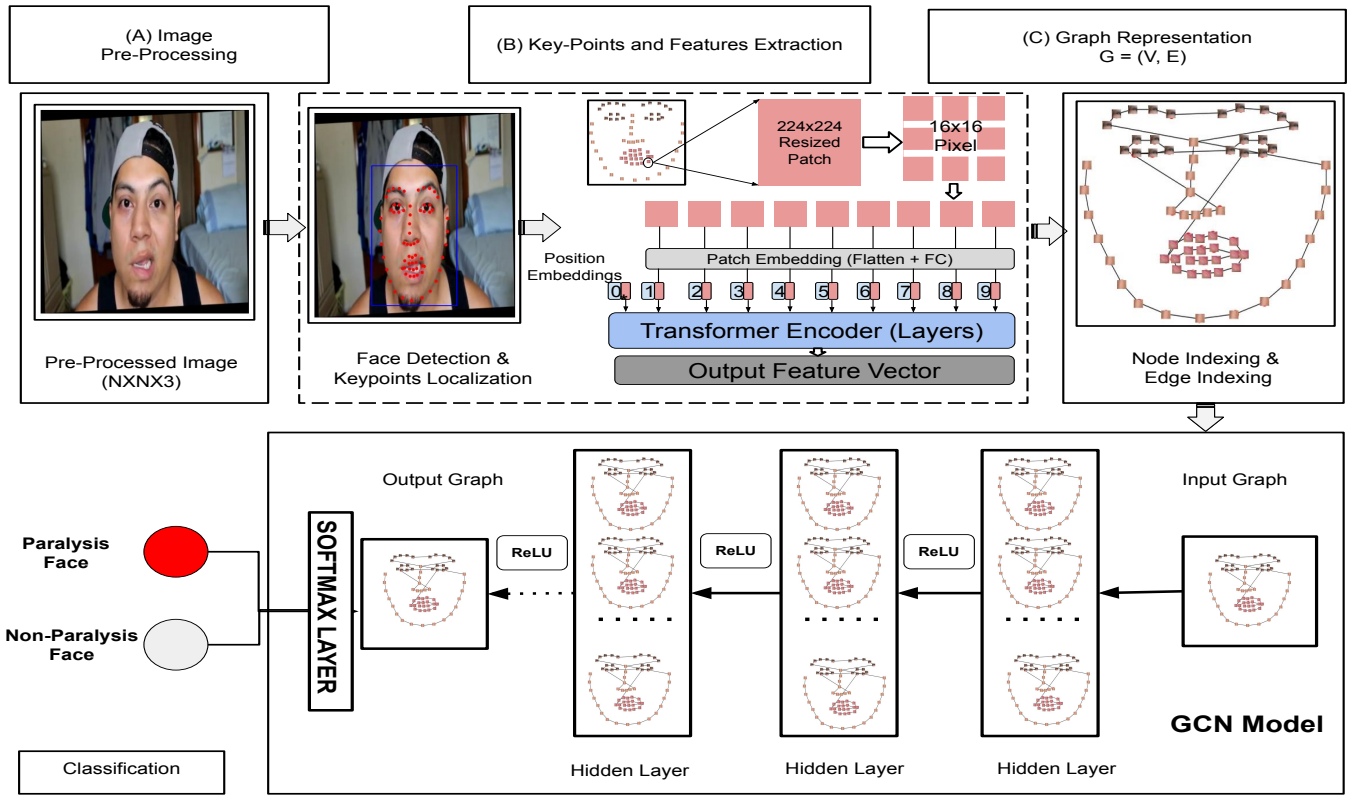

Fig. 2: An outline of the proposed framework (Para-X) for detecting and recognizing facial paralysis in the figure. Our Para-X framework is composed of four primary steps: (i) initial processing for face detection, local patch, and its feature extraction; (ii) graph representation of facial attributes; (iii) GCN blocks for categorized learning; and (iv) output as classification Paralysis or Non-paralysis face. The classification algorithm of fully connected layers is then used to discriminate between different Paralysis or Non-paralysis face labels once the graph and GCN modules have iterated over the facial graph to retrieve information about the paralysis of the face.

has significantly advanced the field, addressing some limitations of earlier methods. Hsu *et al.* [10] introduced hierarchical detection networks (HDN), which integrate local palsy region analysis and facial key-points using deep learning, evaluated on a limited dataset of 32 videos from 22 patients. The song *et al.* [35] utilized CNNs trained on 1049 clinical images, while Hossain *et al.* [36] compared ResNet50, InceptionV3, and VGG16, demonstrating the effectiveness of ResNet50 for facial paralysis detection. Storey *et al.* [37] developed 3DPalsyNet, a 3D CNN that combined joint supervised learning and transfer learning to grade facial palsy and detect motion. Liu *et al.* [6] enhanced performance by introducing a parallel hierarchical convolutional neural network (PHCNN) with LSTM, which analyzed temporal fluctuations and region-based asymmetry. Despite the significant advancements, these models often suffered from tiny sample sizes and limited clinical validation, which affected their generalizability and real-world applicability.

Standardized datasets and robust evaluation frameworks have been identified as crucial for improving the reliability and adoption of automated facial palsy diagnosis. Xia *et al.* [27] addressed this gap by creating the annotated facial key-points for the facial palsy (AFLFP) database, consisting of 16-class asymmetric facial expressions annotated with 68 facial key-points from 88 subjects. AFLEP database facilitated the development of a DNNs baseline using two-stage cascaded fully convolutional networks (FCNs) for facial landmark detection. Although AFLFP provides a valuable resource, its manual annotation process is time-consuming and prone to human error.

Jia *et al.* [38] explored graph-in-graph GCNs for hyperspectral image categorization, a method adaptable for analyzing facial asymmetry through spatial relationships. Guan *et al.* [39] introduced node-aligned GCNs for whole-slide image representation, which could model interlinked facial muscle movements associated with facial paralysis. Gao *et al.* [40] proposed a unique representation-learning approach for dynamic graphs suitable for modeling facial region's temporal and spatial dynamics. GCNs have emerged as a promising tool for capturing spatial correlations in facial features, offering a new avenue to advance facial palsy recognition.

In summary, traditional methods for extracting facial features, such as LDA, GMMs, LBP, SVM, and CNNs, have limitations in capturing subtle correlations and changes in fa-

cial paralysis [41], [42]. Advanced techniques like hypergraph-guided feature embedding and vision transformers (ViTs) are better suited for managing these complexities, especially when dealing with large and diverse datasets. Vision transformers, in particular, can capture global contextual information and intricate interactions between features more effectively, making them robust for the FER task. Recent research has explored combining ViT and GCNs to enhance FPR by integrating global visual and geometric information, using patch-specific convolutional branches for local appearance features, and employing attention mechanisms on facial components in graph-based learning.

III. Para-X Framework

The proposed Para-X system has three primary steps: (i)face detection and key-points localization, extracting features of cropped patches around the key-points via pre-trained vision transformers, (ii) graph representation of facial attributes using algorithm 1, and (iii) facial paralysis and non-paralysis classification with GCNs.

A. Defining Facial Attributes and Graph Construction Phase

Initially, image processing is performed to standardize dimensions and enhance quality. Face detection [43] and key-points localization [44] via utilizing the Dlib, and the regions surrounding the key-points are cropped patches to encoded with a pre-trained ViT [45]. The graph representation of face attributes is created by constructing an adjacency matrix $\boldsymbol{W}$ that denotes links between feature-encoded cropped patches around the key-points based on feature similarity and spatial proximity. Our method accurately represents face features and associates them with different expressions. The procedure initiates by normalizing each feature vector using the L_2 norm, followed by the computation of a similarity function $\mathcal{K}(\mathbf{f}_i, \mathbf{f}_j)$ for pairs of feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$. A distance matrix is concurrently calculated using the squared Euclidean distances between the spatial coordinates of the key-points. The initial adjacency matrix $\boldsymbol{W_{ij}}$ is derived by dividing the similarity function by an exponential function of the Euclidean distance, as seen in Equation 1. This approach efficiently amalgamates both feature and geographical information within the graph framework.

A thresholding hyperparameter τ enhances the adjacency matrix, eliminating weak connections, and preserving only significant interactions between key-points. The refinement involves comparing each element $\boldsymbol{W}_{ij}$ of the matrix to a threshold value determined by $\mu_K + \tau \cdot \sigma_K$, where μ_K and σ_K represent the mean and standard deviation of the initial adjacency matrix. If $\boldsymbol{W}_{ij}$ surpasses the threshold, it is assigned a value of 1; otherwise, it is assigned a value of 0, as seen in Equation 2.

$$\boldsymbol{W}_{ij} = \frac{\mathcal{K}(\mathbf{f}_i, \mathbf{f}_j)}{e^{\|\mathbf{q}_i - \mathbf{q}_j\|}} \tag{1}$$

This formulation guarantees that spatially proximate key-points and homologous features exhibit deeper connections within the matrix.

$$\boldsymbol{W} = \begin{cases} 1, & \text{if } \boldsymbol{W}_{ij} > \mu_K + \tau \cdot \sigma_K \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The equations 1, 2 facilitate the development of a graph-based representation of a face, wherein facial key-points serve as nodes, and both feature similarity and spatial proximity weight the edges (connections) between them. The refined adjacency matrix is applicable for facial paralysis recognition in graph construction.

Algorithm 1 Graph Construction from Image

Input: Image $\mathbf{I}$
Output: Graph $\mathcal{G} = (\mathbf{Q}, \mathbf{F}, \mathbf{W})$

Region $\leftarrow$ DetectFace($\mathbf{I}$) ▷ Identify face region in the image
$\mathbf{Q} \leftarrow$ GetKeypoints(Region) ▷ Extract keypoint coordinates
for each keypoint $\mathbf{q}_j$ in $\mathbf{Q}$ **do**
    $\mathbf{I}_j \leftarrow$ CropPatch($\mathbf{I}, \mathbf{q}_j$) ▷ Crop patch centered at keypoint
    $\mathbf{f}_j \leftarrow$ ComputeFeatures($\mathbf{I}_j$) ▷ Generate feature descriptor
    Append $\mathbf{f}_j$ to $\mathbf{F}$
end for
Compute adjacency matrix $\mathbf{W}$ for $\mathbf{Q}$ and $\mathbf{F}$ via using equations (1) and (2).
return $\mathcal{G} = (\mathbf{Q}, \mathbf{F}, \mathbf{W})$

B. Facial paralysis Classification Using GCNs

The matrix $\boldsymbol{W_{ij}}$ for constructing a graph defines the spatial relationships between facial key-points. Let $\boldsymbol{I}$ denote the processed facial expression and paralysis image, $\mathbf{Q}$ denote the set of facial key-points, $\boldsymbol{F}$ denote the set of graph-based features, and $\boldsymbol{W}$ denote the adjacency matrices of the image. Algorithm 2 delineates the GCNs model.

Algorithm 2 GCN Model for our Para-X framework

Input: Feature matrix $\mathbf{F}$, Adjacency matrix $\mathbf{W}$, Number of layers K
Output: Node feature matrix $\mathbf{H}^{(K)}$

Initialization: Set the initial feature matrix $\mathbf{H}^{(0)} \leftarrow \mathbf{F}$
    $\tilde{\boldsymbol{W}} \leftarrow \hat{\boldsymbol{D}}^{-1/2}(\boldsymbol{W} + \boldsymbol{I})\hat{\boldsymbol{D}}^{-1/2}$
    where $\hat{\boldsymbol{D}}$ is the degree matrix of $\boldsymbol{W} + \boldsymbol{I}$.
for $k = 0$ to $K - 1$ **do**
        Feature Propagation: Update the feature matrix
        $\mathbf{H}^{(k+1)} \leftarrow \phi\left(\tilde{\mathbf{W}}\mathbf{H}^{(k)}\mathbf{M}^{(k)}\right)$
        where $\mathbf{M}^{(k)}$ is the weight matrix for layer k.
end for
return $\mathbf{H}^{(K)}$

The operation at each layer involves two main steps:

1) **Feature Aggregation**: The features of the node, graph structure encoded in $\tilde{\boldsymbol{W}}$ is used to aggregate $\boldsymbol{H}^{(k)}$.
2) **Feature Transformation**: The aggregated features are subsequently transformed through a learnable weight matrix $\boldsymbol{W}^{(k)}$ and passed through a non-linear activation function $\phi(\cdot)$.

This procedure is repeated for each layer in the GCNs, acquiring increasingly abstract representations of the nodes in the graph. The binary cross-entropy loss ($\mathcal{L}$) is employed to train the Para-X as follows:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{j=1}^{N}[y_j \log(\hat{y}_j) + (1 - y_j)\log(1 - \hat{y}_j)] \quad (3)$$

Here, y_j and $\hat{y}_j$ represent the true label and the predicted probability for the paralysis class of the j-th instance, respectively.

IV. Experimental Evaluation

In this section, we provide essential details of the experiments conducted to assess the performance of existing methods. We compared Para-X with recent and state-of-the-art methods such as [6], [46]–[48]. Our framework uses Dlib [43] for face detection; PyTorch2.1.2+cu121 is part of the software stack, allowing deep learning. The NVIDIA RTX A6000 GPU with 48 GB memory was used for the evaluation. In our implementation, the hyperparameters are configured as follows: the image size is set to 224x224 pixels, with an initial learning rate of 0.001, gradually decreasing to a final learning rate of interval 1e-6. We apply a weight decay of 5e-4 and utilize the Adam optimizer. The model includes a hidden layer with 256 units, a dropout rate of 0.2, and a seed value of 1000 to ensure reproducibility. For better generalization, we pre-trained our model on the FER dataset. We trained on YFP+OuluCASIA, YFP+CalTech face, and AFLFP+DISFA+ dataset as shown in the Fig. 4, fine-tuning datasets shown in the TABLE II, III. We tested on the FPR dataset because the availability of the FER dataset is broader than that of FPR datasets. To explore the generalizability of our method, we conducted extensive experiments with the standard benchmark datasets used to evaluate FER using the Para-X method. The

TABLE I: Details of Facial Paralysis and Expression Datasets

Dataset	Resolution	Samples	Modality
AFLFP [27]	640x480	88 subjects	RGB
YFP [10]	640x640	32 videos	RGB
FDFPI [16]	Variable	1024 images	RGB
DISFA [29]	1024x768	27 subjects	RGB
DISFA+ [29]	1024x768	75 subjects	RGB
OuluCASIA [30]	320x240	80 subjects	RGB/NIR
Caltech face [31]	896x592	450 images	RGB
RML [32]	640x480	1,500	RGB
FPD (Our)	Variable	694 images	RGB

datasets and details used in our experiments are shown in the TABLE I. The YFP dataset comprises 32 videos from 21 patients, processed into six fps image sequences. Local paralysis regions were identified based on high deformity intensity, annotated by three physicians, with their intersection serving as the ground truth. Each region was categorized by intensity—0.5 (low) or 1.0 (high) and classified as either the eye or mouth area. The FDFPI dataset includes curated images of facial deformities, capturing asymmetries in the eyes, mouth, and other facial regions affected by paralysis. Details of the AFLFP dataset are provided in §II. Our facial paralysis detection dataset comprises 694 images sourced from Google and YouTube, featuring diverse subjects across different age groups, including children, teenagers, adults, and elderly individuals. The dataset includes both single-face and multi-face scenarios. Using the YFP dataset as a reference, we annotated images focusing on the eye and mouth regions. The dataset is categorized into six classes: *normal_eye*, *normal_mouth*, *slightly_palsy_eye*, *slightly_palsy_mouth*, *strongly_palsy_eye*, and *strongly_palsy_mouth*. Our dataset aims to advance research in detecting and analyzing facial paralysis across various age groups.

The DISFA dataset consists of videos from 27 participants exhibiting spontaneous facial expressions. Each frame is annotated for action unit (AU) intensity on a scale from 0 to 5, making it a valuable resource for facial expression analysis and AU recognition. An extension, DISFA+, enhances the original dataset with additional annotations, improved quality, and a broader range of expressions, making it more comprehensive for studying facial movements and emotions. Both datasets contribute significantly to affective computing and facial behaviour research. The Oulu-CASIA dataset contains videos of 80 subjects displaying six basic emotions (happiness, sadness, anger, surprise, fear, and disgust) under three lighting conditions—normal, weak, and dark—using visible and near-infrared (NIR) imaging. The RML dataset integrates facial expressions, speech, and physiological signals, supporting multimodal emotion recognition research. The Caltech Face dataset consists of 450 color images of 27 unique individuals, each captured in frontal view under varying lighting, expressions, and backgrounds.

TABLE II: Comparison of Methods

Method	Datasets	Class	Accuracy (%)
MobileNetV2 [48]	YFP+CK	2	98.93
SVM [49]	YFP+CK	2	95.59
VDRRE [46]	YFP+Caltech	2	99.34
PHCNN+LSTM [6]	YFP+CK	2	94.61
RC-SSELM-VC [47]	YFP+CK	2	85.50
ResNet50 [36]	FPFDI+UTK	2	96.6
Para-X (our)	AFLFP+DISFA	2	91.15
Para-X (our)	YFP+DISFA	2	96.09
Para-X (our)	FDFPI + RML	2	94.47
Para-X (our)	FPD+ RML	2	97.09
Para-X (our)	YFP+CalTech face	2	**99.99**

TABLE III: Performance Metrics for Different Datasets

Dataset	τ	Loss	Acc	P	R	F1
FDFPI + RML	0.35	0.37	94.47	90.61	94.47	92.5
FPD + RML	0.35	0.34	97.09	92.84	97.09	94.92
YFP + DISFA	0.2	0.35	96.09	98	98.04	96.04
YFP + CalTech	0.2,0.25	0.31	99.99	99.99	99.99	99.99
AFLFP + DISFA	0.25	0.40	91.15	95.57	95.57	91.15

Acc: Accuracy(%), P: Precision(%), R: Recall(%), F1: F1-Score(%)

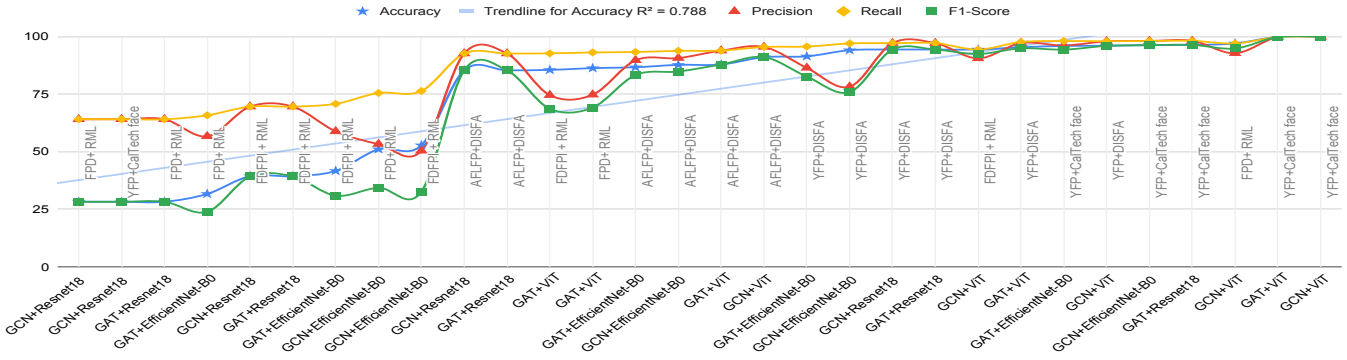

Fig. 3: Ablation study on different Models

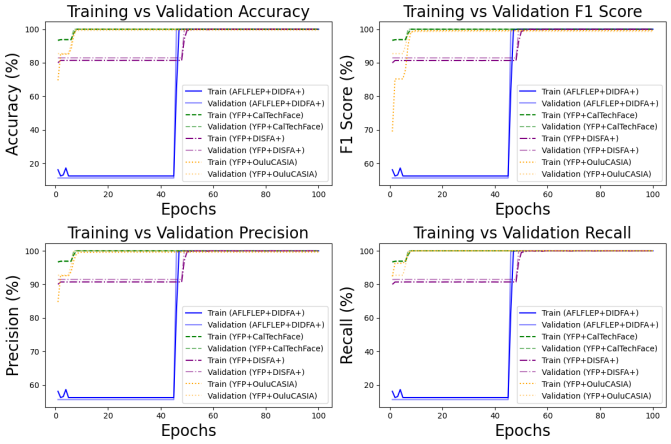

Fig. 4: Training vs. validation performance metrics over epoch.

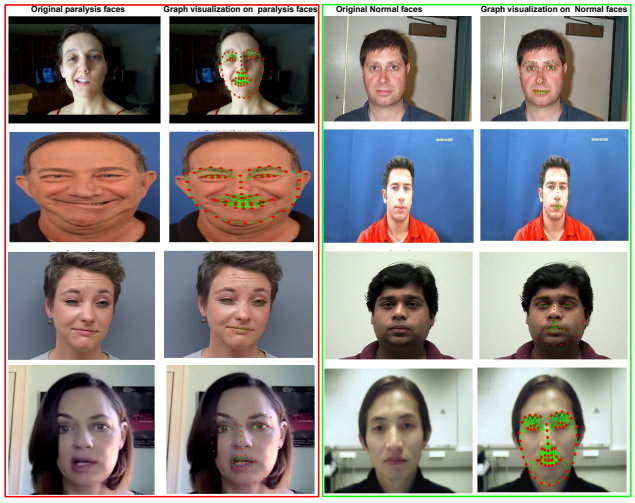

Fig. 5: Graph visualization on the test datasets.

TABLE II compares the classification methods applied to various datasets, analyzing their accuracy. Several methods, including MobileNetV2, SVM, VDRRE, PHCNN+LSTM, and RC-SSELM-VC, are compared against the proposed approach (denoted as "our"). The results show that deep learning-based methods, such as MobileNetV2 (98.93% accuracy) and SVM (95.59% accuracy), generally perform better than approaches like PHCNN+LSTM (85.90%). Notably, the proposed method outperforms all other methods, achieving 99.99% accuracy on the YFP + CalTech face dataset. Our method also attains high accuracy on other datasets, such as 97.09% on FDP + RML and 96.09% on YFP + DISFA, proving its superior performance. TABLE III provides more details with other metrics and threshold values, where YFP + CalTech face achieves the highest with intense precision (99.99%) and recall (99.99%) and F1-score (99.99%). FDPFI + RML also performs well, achieving 94.47% accuracy. At the same time, AFLFP + DISFA records the lowest accuracy at 91.15% and the corresponding training and validation results shown in the Fig. 4, Fig. 3 presents a graph of an ablation study analyzing the performance of different models, and Fig. 5 graph visualization with a threshold of 0.30. TABLE IV presents a comparative analysis of different object detection models such as YOLO-based [50], RT-DETR [51], and YOLOv8-AM [52].

TABLE IV: Performance metrics for various best YOLO versions, YOLOv8+Attention, and RT-DTR models

Method	Dataset	P	R	m1	m2	IT
HDN	YFP	89	87	-	-	-
Yolov5l	YFP	98.4	94.6	97.7	96.3	7.9
Yolov10l	FDFPI	98.01	96.69	98.21	90.45	15.2
Yolov3	FPD	93.31	92.19	92.01	72.13	15.6
Yolov5n	FDFPI+FPD	84.8	88.6	92	75.5	4.2
RT-DETR	YFP	98.5	93.3	94.2	89.9	10
RT-DETR	FDFPI	96.9	97.5	96.9	87.2	10.1
RT-DETR	FPD	91.2	94.5	92.7	70.2	13.1
RT-DETR	FDFPI+FPD	92.4	91.7	90.1	73.9	21.4
Yolov8l +AM1	FDFPI	96.5	96.6	98.2	98.8	5.2
Yolov8l +AM2	FPD	93	90.1	94.4	74.8	7
Yolov8l +AM1	YFP	98.1	95	98.4	96.8	1

mAP50: m1, m2: mAP50-90, IT: Inference time per image (ms/img)

RT-DETR achieves the highest recall and competitive mAP values across different datasets, with inference times ranging from 10.0 to 21.4 ms. The YOLOV3 model performs relatively well on the FPD dataset, but its inference time is higher at 15.6 ms. Yolov5n, being a lightweight model, exhibits a balance between accuracy and efficiency and the trade-offs between precision, recall, and inference time among the evaluated models. The YOLOv8-AM improved version of YOLOv8

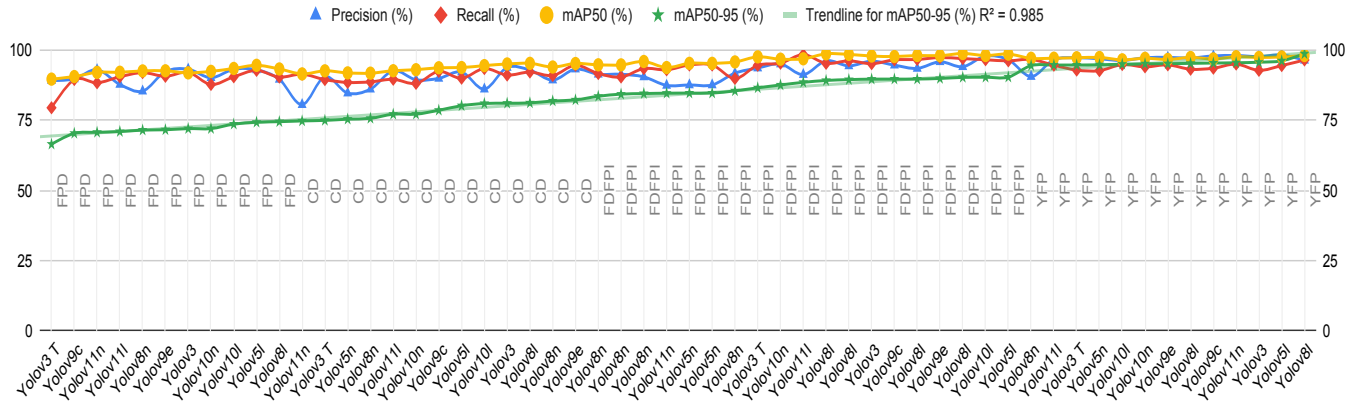

Fig. 6: All YOLO best versions results on the FPR datasets.

TABLE V: Ablation Study Results on YFP+Caltech Face

Cases	M1	M2	M3	M4	M5	Accuracy	F1-Score
E	✓	✗	✗	✗	✓	28.3	28.3
A	✓	✗	✗	✗	✗	88.38	35.99
C	✗	✗	✓	✗	✗	95.82	54.65
F	✗	✓	✗	✓	✗	96.03	94.4
G	✗	✓	✗	✗	✓	96.35	96.35
D	✓	✗	✗	✓	✗	96.46	96.46
B	✗	✓	✗	✗	✗	99.52	93.13
H	✗	✗	✓	✓	✗	99.96	99.93
I	✗	✗	✓	✗	✓	99.99	99.99

M1-5: Modules, M1: ResNet18, M2: EfficientNet-B0, M3: ViT, M4: GAT, M5: GCNs.

integrates attention mechanisms to improve performance. Our approach consists of four different attention modules: residual convolutional block attention module (ResCBAM: AM1), efficient channel attention (ECA: AM2), shuffle attention (SA: AM3), and global-context attention mechanism (GAM: AM4). By integrating these modules, we refine the model's architecture and optimize its training process in Fig. 6, where Yolov8l + AM1 outperforms mAP50-90 98. 8%, 96. 8%, 74. 4% to FDFPI and YFP, FPD, respectively.

Overall, the results suggest that the proposed approach is highly effective, surpassing existing models in classification accuracy, particularly on the YFP + CalTech face dataset. The comparison also highlights the advantage of deep learning models over conventional techniques, reinforcing their reliability in complex classification tasks. YOLO-based models, particularly YOLOv5l and YOLOv10l, demonstrate superior precision, recall, and mAP scores on the six annotated classes per the YFP dataset. However, RT-DETR models also show substantial performance, especially on FDPFI and YFP datasets, albeit with higher inference times. The findings suggest that different YOLO variants and datasets have unique strengths, with Yolo+attention models excelling in mAP while others balance accuracy and efficiency. The results presented in the TABLE V ablation study demonstrate the impact of different model architectures on accuracy (%) and F1-score (%). The configuration incorporating ViT and GCNs achieved the highest performance with 99.99% accuracy and F1-score. EfficientNet-B0 also significantly contributed to performance, as seen in Experiments B and G. Furthermore, models using ResNet18 alone showed significantly lower accuracy. These findings highlight the importance of model selection and combination in achieving optimal performance.

V. Conclusion

This work introduces Para-X, a novel framework for efficient facial paralysis recognition using a graph-based representation of facial attribute structures. Our approach captures the intricate interactions among facial features by representing key-points as graph vertices, with edges defined by proximity and similarities of local appearance. Using vision transformers and graph convolutional networks, Para-X encodes structural and global dependencies, resulting in a more robust and comprehensive representation of facial expressions. The evaluation results demonstrate the effectiveness and generalizability of our proposed method.

References

[1] A. M. Kosins, K. A. Hurvitz, G. R. Evans, and G. A. Wirth, "Facial paralysis for the plastic surgeon," *Canadian Journal of plastic surgery*, vol. 15, no. 2, pp. 77–82, 2007.

[2] A. Ciorba, V. Corazzi, V. Conz, C. Bianchini, and C. Aimoni, "Facial nerve paralysis in children," *World Journal of Clinical Cases*, vol. 3, no. 12, p. 973, 2015.

[3] N. R. Walker, R. K. Mistry, and T. Mazzoni, "Facial nerve palsy," in *StatPearls [Internet]*. StatPearls Publishing, 2023.

[4] C. S. Wang, M. Sakai, A. Khurram, and K. Lee, "Facial nerve palsy in children: a case series and literature review," *Otolaryngology Case Reports*, vol. 20, p. 100297, 2021.

[5] H. S. Kim, S. Y. Kim, Y. H. Kim, and K. S. Park, "A smartphone-based automatic diagnosis system for facial nerve palsy," *Sensors*, vol. 15, no. 10, pp. 26 756–26 768, 2015.

[6] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham, "Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2325–2332, 2020.

[7] A. Namavarian, A. Eid, H. Ziai, E. Y. Cheng, and D. Enepekides, "Facial nerve paralysis and covid-19: a systematic review," *The Laryngoscope*, vol. 133, no. 5, pp. 1007–1013, 2023.

[8] C. Xu, C. Yan, M. Jiang, F. Alenezi, A. Alhudhaif, N. Alnaim, K. Polat, and W. Wu, "A novel facial emotion recognition method for stress inference of facial nerve paralysis patients," *Expert Systems with Applications*, vol. 197, p. 116705, 2022.

[9] M. E. Davis and J. J. Greene, "Advances and future directions in the care of patients with facial paralysis," *Operative Techniques in Otolaryngology-Head and Neck Surgery*, vol. 33, no. 1, pp. 60–71, 2022.

[10] G.-S. J. Hsu, W.-F. Huang, and J.-H. Kang, "Hierarchical network for facial palsy detection." in *CVPR Workshops*, 2018, pp. 580–586.

[11] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1859–1866.

[12] N. Perveen, C. K. Mohan, and Y.-W. Chen, "Quantitative analysis of facial paralysis using gmm and dynamic kernels." in *VISIGRAPP (5: VISAPP)*, 2020, pp. 173–184.

[13] 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*, 2022.

[14] Z. Guo, W. Li, J. Dai, J. Xiang, and G. Dan, "Facial imaging and landmark detection technique for objective assessment of unilateral peripheral facial paralysis," *Enterprise Information Systems*, vol. 16, no. 10-11, pp. 1556–1572, 2022.

[15] C. Vletter, H. Burger, H. Alers, N. Sourlos, and Z. Al-Ars, "Towards an automatic diagnosis of peripheral and central palsy using machine learning on facial features," *arXiv preprint arXiv:2201.11852*, 2022.

[16] K. Mehta, "Facial droop and facial paralysis image dataset," https://www.kaggle.com/datasets/kaitavmehta/facial-droop-and-facial-paralysis-image, accessed: 2025-01-28.

[17] W. Gao and Y. Xia, "Ccfexp: Facial image synthesis with cycle cross-fusion diffusion model for facial paralysis individuals," *CoRR*, vol. abs/2409.07271, 2024.

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

[19] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.

[20] 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.

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

[22] Y. Liu, X. Zhang, J. Zhou, and L. Fu, "Sg-dsn: A semantic graph-based dual-stream network for facial expression recognition," *Neurocomputing*, vol. 462, pp. 320–330, 2021.

[23] 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.

[24] N. Xie, J. Li, M. Guo, L. Yang, and Y. Gong, "Attention-based global-local graph learning for dynamic facial expression recognition," in *ICIG (1)*, ser. Lecture Notes in Computer Science, vol. 14355. Springer, 2023, pp. 3–15.

[25] 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.

[26] 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.

[27] Y. Xia, C. Nduka, R. Y. Kannan, E. Pescarini, J. E. Berner, and H. Yu, "Aflfp: A database with annotated facial landmarks for facial palsy," *IEEE Transactions on Computational Social Systems*, 2022.

[28] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[29] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 1–8.

[30] 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.

[31] M. Weber, "Caltech face dataset 1999," Jul 2022.

[32] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[33] T. H. Ngo, M. Seo, N. Matsushiro, W. Xiong, and Y.-W. Chen, "Quantitative analysis of facial paralysis based on limited-orientation modified circular gabor filters," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 349–354.

[34] T. Wang, J. Dong, X. Sun, S. Zhang, and S. Wang, "Automatic recognition of facial movement for paralyzed face," *Bio-medical materials and engineering*, vol. 24, no. 6, pp. 2751–2760, 2014.

[35] A. Song, Z. Wu, X. Ding, Q. Hu, and X. Di, "Neurologist standard classification of facial nerve paralysis with deep neural networks," *Future Internet*, vol. 10, no. 11, p. 111, 2018.

[36] S. M. Hossain, Z. Jamal, A. A. Noshin, and M. M. Khan, "Comparative study of deep learning algorithms for the detection of facial paralysis," in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2022, pp. 0368–0377.

[37] G. Storey, R. Jiang, S. Keogh, A. Bouridane, and C.-T. Li, "3dpalsynet: A facial palsy grading and motion recognition framework using fully 3d convolutional neural networks," *IEEE access*, vol. 7, pp. 121 655–121 664, 2019.

[38] S. Jia, S. Jiang, S. Zhang, M. Xu, and X. Jia, "Graph-in-graph convolutional network for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[39] Y. Guan, J. Zhang, K. Tian, S. Yang, P. Dong, J. Xiang, W. Yang, J. Huang, Y. Zhang, and X. Han, "Node-aligned graph convolutional network for whole-slide image representation and classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 813–18 823.

[40] C. Gao, J. Zhu, F. Zhang, Z. Wang, and X. Li, "A novel representation learning for dynamic graphs based on graph convolutional networks," *IEEE Transactions on Cybernetics*, 2022.

[41] J. Lou, H. Yu, and F.-Y. Wang, "A review on automated facial nerve function assessment from visual face capture," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, pp. 488–497, 2019.

[42] Y. Zhang, W. Gao, H. Yu, J. Dong, and Y. Xia, "Artificial intelligence-based facial palsy evaluation: a survey," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

[43] D. E. King, "dlib c++ library," accessed: January 27, 2025.

[44] 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 (CVPR)*, 2014, pp. 1867–1874.

[45] 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.

[46] O. O. Abayomi-Alli, R. Damaševičius, R. Maskeliūnas, and S. Misra, "Few-shot learning with a novel voronoi tessellation-based image augmentation method for facial palsy detection," *Electronics*, vol. 10, no. 8, p. 978, 2021.

[47] X. Tan, J. Yang, and J. Cao, "Facial nerve paralysis assessment based on regularized correntropy criterion sselm vc and cascade cnn," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2021, pp. 1043–1047.

[48] Z. M. Baig and D. Van Der Haar, "Facial paralysis recognition using face mesh-based learning." in *ICPRAM*, 2023, pp. 881–888.

[49] G. S. Parra-Dominguez, C. H. Garcia-Capulin, and R. E. Sanchez-Yanez, "Automatic facial palsy diagnosis as a classification problem using regional information extracted from a photograph," *Diagnostics*, vol. 12, no. 7, p. 1528, 2022.

[50] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[51] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "Detrs beat yolos on real-time object detection," 2023.

[52] R.-Y. Ju, C.-T. Chien, and J.-S. Chiang, "Yolov8-rescbam: Yolov8 based on an effective attention module for pediatric wrist fracture detection," *arXiv preprint arXiv:2409.18826*, 2024.