MSE-GCN: Multiscale spatiotemporal feature aggregation enhanced Efficient Graph Convolutional Network for Dynamic Sign Language Recognition

Neelma Naz, Hasan Sajid, Sara Ali, Osman Hasan, and Muhammad Khurram Ehsan

Abstract— **Graph convolution networks have emerged as an active area of research for skeleton-based sign language recognition (SLR). One essential problem in this approach is to extract the most discriminative features capable of modeling short-range and long-range spatial and temporal information over all skeleton joints while ensuring low inference costs. To address this issue, we propose a novel multi-scale efficient graph convolutional network (MSE-GCN) for skeleton-based SLR. The proposed network makes use of separable convolution layers set in a multi-scale setting and embedded in a multi branch (MB) network along with an early fusion scheme, resulting in an accurate, computationally efficient, and faster system. In addition, we have proposed a novel hybrid attention module, named Spatial Temporal Joint Part attention (ST-JPA) to distinguish the most important body parts as well as most informative joints in the specific frames from the whole sign sequence. The proposed network (MSE-GCN) is evaluated on five challenging sign language datasets, WLASL-100, WLASL-300, WLASL-1000, MINDS-Libras, and LIBRAS-UFOP achieving state-of-the-art (SOTA) accuracies of 85.27%, 81.59%, 71.75%, 97.442 ± 1.01%, and 88.59±3.60%, respectively while incurring significantly lower computational costs.**

Index Terms— Sign Language Recognition, Skeleton Modeling, Graph Convolution Network, Multiscale Architectures, Separable Convolution, Efficient Net**,** Visualization

I. INTRODUCTION

In the realm of communication, sign language (SL) plays a vital role as the primary means of interaction for the deaf community. SLs are visual communication systems that transmit information through hand movements, body gestures, head motions, facial expressions, and eye gaze. With their distinctive grammar and lexicon, SLs pose a significant challenge for deaf and non-deaf individuals to acquire proficient skills. To promote unhindered communication between hearing and deaf individuals, extensive research has been conducted on automatic visual sign language understanding. This study focuses on the task of isolated sign language recognition (ISLR), which involves the classification of isolated signs from skeleton input into a predefined set of glosses (unique labels of signs, recognized by the words representing signs semantic meanings). ISLR holds significant potential for various applications, such as sign spotting, continuous sign language recognition [1], sign language translation, and sign video retrieval [2].

Due to the defining characteristics of sign language, including handshapes and movements, the possible combinations of these visual elements are inherently limited which results in a multitude of visually indistinguishable signs, posing a challenge for learning approaches to accurately discern and identify them. The extraction of discriminative and comprehensive features that effectively capture the spatial configurations and temporal dynamics intrinsic to SL, poses a significant challenge in SLR. Consequently, this aspect has garnered substantial research attention, leading to the development of numerous approaches, including appearance based (RGB), skeleton based (pose), and hybrid methods, aimed at addressing these challenges. Recent ISLR approaches [3] use the I3D [4] network with RGB sequences as input but face limited success, as I3D primarily relies on global appearance features and fails to capture fine-grained movements like finger motions. Moreover, appearance-based approaches are susceptible to the influence of illumination variations, camera viewpoints, and other background changes.

Currently, skeleton-based approaches have gained significant popularity in the field of SLR due to their ability to effectively capture dynamic changes in human body movements and enhanced robustness against illumination changes, viewpoints, and background changes. The progress in skeleton based SLR can be divided into two phases. Initially, conventional methods utilized Recurrent neural network (RNN) [3, 5] or Convolutional Neural Network (CNN) for the analysis of skeleton sequence. In recent years, graph-based models [3, 6-8] have gained attention due to their ability to represent structural data effectively. However, these SOTA approaches for extracting informative and diverse features face two notable limitations. Firstly, these approaches tend to be excessively complex and over-parametrized resulting in substantial computational requirements. Secondly, they predominantly prioritize short-range connections, overlooking the importance of long-range dependencies in SLR. For instance, as shown in Fig. 1, the sign for 'bed' requires the

(Corresponding author: Neelma Naz).

Neelma Naz, Hasan Sajid, Sara Ali, Osman Hasan are with National University of Sciences and Technology, Islamabad 44000, Pakistan (e-mail: neelma.naz@seecs.edu.pk; hasan.sajid@smme.nust.edu.pk;sarababer@smme.nust.edu.pk; osman.hasan@seecs.edu.pk)

Muhammad Khurram Ehsan is with Faculty of Engineering Sciences, Bahria University Islamabad Campus, Islamabad 44000, Pakistan (email: mkehsan.buic@bahria.edu.pk)

In our experiments, we have used publicly available datasets to evaluate our model's performance. We have followed all protocols provided by datasets authors.

coordination of left hand and head joints which are distant from each other and long-range dependencies between these joints need to be modeled for accurate recognition and sign for ‘about’ require coordination between both hands.

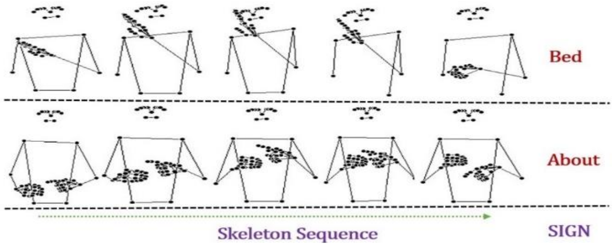


Figure 1. Samples from WLASL dataset

To tackle the first problem, we have developed an early fused multi branch (MB) network that effectively captures comprehensive characteristics derived from both spatial configurations and temporal dynamics of joints in skeleton sequence. To further reduce the model complexity and extract spatial and temporal dynamics, we have used Separable Layer (SepLayer) [9], and SandGlass Layer (SGLayer) [10], in the GCN network. To address the second problem, we have incorporated a multi-scale approach [11] along both spatial and temporal dimensions, which can capture short-range and long-range joint dependencies quite effectively. The multi-scale graph convolution (MS-GCN) replaces the graph convolution with a set of sub-graph convolutions that establishes a hierarchical structure incorporating residual connections between neighboring states. As the features traverse through this module, information exchanges with nearby nodes result in an expansion of receptive field and thus captures long range dependencies. Finally, In order to achieve improved recognition accuracy, we have introduced a novel attention module called spatiotemporal joints part attention (ST-JPA) which serves the purpose of identifying crucial joints within the complete skeleton sequence. By integrating all these modules, an efficient and lightweight MSE-GCN network is proposed that outperforms the most recent SLR methods by achieving SOTA recognition accuracy while incurring lower computational complexity. The key contributions of our work can be summarized as follows:

- To enhance SLR accuracy by efficiently capturing spatiotemporal short- and long-range dependencies between skeleton joints, we propose a novel Multi-Scale Efficient Graph Convolution Network (MSE-GCN). This network leverages a multi-input, multi-scale architecture in conjunction with an early fusion strategy.
- To ensure computational efficiency and high accuracy, separable convolutions have been introduced in spatial and temporal convolution layers instead of commonly used bottleneck layers.
- To enhance model’s ability to capture most significant spatial and temporal joints information, we propose a novel Spatiotemporal Joints Part Attention (ST-JPA) module.
- Through comprehensive evaluation on three challenging SL datasets: WLASL, LIBRAS-UFOP, and MINDs-Libras, MSE-GCN demonstrates SOTA results in terms of accuracy and computational cost.

The rest of the paper is structured as follows: In Section II, we discuss recent studies related to our work. Section III provides details of our proposed network MSE-GCN. We report extensive experiments conducted on three large scale datasets in Section IV. Finally, Section V concludes the paper.

II. Related Work

To address the problem of SLR, various approaches have been put forth in literature. In this section, we provide a brief review of the RGB-based, Skeleton-based, Hybrid and attention-based methodologies proposed for SLR.

A. RGB-Based SLR

In SLR frameworks, Spatiotemporal feature extraction plays a pivotal role and various approaches have been proposed to extract the most discriminative visual representations from RGB videos.

2D-CNNs and RNNs based architectures: In recent times, deep CNNs have demonstrated remarkable capability in learning representations and have gained extensive usage in SLR. For instance, a 2- CNN was utilized as a backbone in [12] to extract spatial features which was followed by an LSTM network to model temporal clues. An end-to-end neural framework comprising a combination of RNNs and temporal convolutions was proposed in [13]. In [14], a deep network consisting of 2D-CNN as backbone for spatial feature extraction and a transformer for temporal representations was proposed. A variant of 2D-CNNs is the 3D-CNN model which because of its capability to model spatial and temporal cues simultaneously, has found an extensive use in the task of video recognition.

3D-CNNs based Architectures: In the domain of action recognition, the inflated 3D-CNN (I3D) [4] has emerged as one of the most popular 3DCNN architecture. Some other noteworthy 3D-CNN based approaches are Slowfast [15] and S3D[10] architecture. Most recent works on SLR [3, 16] have predominantly employed these architecture to extract visual features from RGB input videos. Despite their efficiency under normal conditions, RGB-based methods are computationally heavy, and their performance degrades in case of illumination changes, viewpoints, and background changes.

B. Skeleton-Based SLR

In recent times, there has been a notable research emphasis on the exploration of methods based on skeletal or pose data for SLR. The proposed approaches using pose data as input can be split into two categories.

CNN and RNN based Architectures: Pose-based approaches employ different CNN [17] and RNN [18] based baselines for modeling the spatiotemporal characteristics of key point sequences. For instance, all the estimated body and hands key points are combined together in the form of a matrix and fed as input to 2-layered stacked GRU in [3] for sign identification. A modified GRU is used to learn spatiotemporal patterns for efficient SLR in [5]. Another transformer based architecture that utilizes pose information as input is proposed in [19].

Graph Convolution Network architectures: Given the structured nature of pose data, an increasing number of studies have embraced the utilization of graph convolution networks (GCNs) [6, 11]. An exemplar in this domain is ST-GCN framework [6], which structures the pose sequence as predefined graph and employs GCN for recognition purpose. In [7], this approach has been extended by incorporating an unpretrained transformer model (BERT) for temporal features extraction. A

GCN based model employing bottleneck layers and residual connections was proposed in [20]. Although pose based SLR techniques effectively reduce computational complexity, they currently exhibit lower accuracy compared to RGB input based approaches.

C. Hybrid Methods and Attention Enhanced Models

RGB-based SLR models face challenges due to varying video backgrounds while pose-based methods tend to have lower accuracy when used alone. To address these limitations, certain SLR studies [21-24] have endeavored to jointly model RGB videos and key points. For instance, the hybrid model (SAM-SLR) presented in [24] uses multiple input modalities i.e., Pose, RGB, Optical Flow (OF), and Depth for SLR. Similarly HMA[23], SignBERT [22], and SignBERT+ [21] propose the utilization of pose data in addition to RGB sequences to guide the model to learn more elaborated representations. However, a prevalent drawback inherent in these approaches is the augmented computational cost arising from the simultaneous processing of multiple modalities. In the realm of SLR, numerous attention based schemes have been introduced, leveraging channel, spatial, and temporal attention mechanisms [14, 25].

In our work, we exclusively rely on pose information of hands and body joints as input. We develop an efficient and lightweight model that surpasses the performance achieved by SOTA methods utilizing RGB, Pose, and Hybrid modalities. The details of proposed architecture are provided in Section III.

III. Methodology

A. Preliminary Work

In this section, we discuss several important techniques used in our proposed Multi-Scale Efficient Graph Convolution network (MSE-GCN).

1) Standard Graph Convolution

A human skeleton graph can be represented as a graph ***G = {V, E}*** consisting of human joints ***V= {v_{ti} | t = 1, …., T, i=1, …, N}*** represented as nodes and intra-frame and inter-frame connections of joints represented as edges ***E.*** Here T and N represent the total number of frames and joints respectively. The graph convolutional layer in spatial dimension is implemented using Eq. (1) :

$$X_{OS} = \sum_{d=0}^{D} \theta_d X_I (D_S^{(d)}{}^{-1/2} (A_S^{(d)} + I) D_S^{(d)}{}^{1/2} \odot \omega_d \qquad (1)$$

Where X_I and X_{OS} represent the input and output feature maps in spatial dimension, D denotes maximum graph distance, $A_S^{(d)}$ is the spatial adjacency matrix of order d, $D_S^{(d)}$ is spatial degree matrix used to normalize $A_S^{(d)}$, and I is the identity matrix used to model self-loops. θ_d and ω_d are trainable parameters matrices used for the implementation of graph convolution. To ensure the coherence between the spatial temporal graph and the video across time, the temporal graph convolution can be implemented with the classical 2-dimensional (2D) convolution operation using Eq. (2):

$$X_{OT} = Conv2D[K_T \times 1](X_I) \qquad (2)$$

Where $Conv2D[K_T \times 1]$ represents 2D temporal convolution operation with kernel size K_T.

2) Multi-Scale Graph Convolution

The Multi-scale graph convolutions[11] can be applied in both spatial and temporal dimensions. For a multi-scale spatial graph convolution (MS-SGC), an input feature vector $\boldsymbol{X} \in \mathcal{R}^{C\times T\times V}$, is partitioned into k splits along the channel dimension, represented as $\boldsymbol{x_j}$ *where* $j \in \{1,2,\dots,k\}$ and each $\boldsymbol{x_j} \in \mathcal{R}^{(C/k)\times T\times V}$. For each split $\boldsymbol{x_j}$, a distinct sub-spatial graph convolution $\mathcal{G}_j$ with 1/k number of channels in comparison to original ones, is implemented using Eq. (1). This results in a reduction in the number of parameters by a factor of $1/k^2$ as compared to original convolutions. Moreover, to increase the variety of receptive fields and capture the dependencies between short range and long-range joints, a residual mechanism is utilized to connect two adjacent splits. The whole operation can be formally represented as Eq. (3):

$$x_o = \begin{cases} \mathcal{G}_j(x_j) & if\ j = 1 \\ \mathcal{G}_j(x_j + x_{o-1}) & if\ j > 1 \end{cases} \qquad (3)$$

Here the output of *jth* sub-spatial graph convolution is denoted as $x_o \in \mathcal{R}^{(C/k)\times T\times V}$. The multi-scale graph convolution ensures enlarged receptive field by aggregating the information from sub-spatial convolutions. The outputs of all the splits are concatenated and finally an additional residual connection is introduced to ensure model convergence. Eq. (4) is used to compute the output of MS-SGC.

$$X_O = \delta(concat(\{x_{oi} | i = 1,2,..,k\}) \oplus X) \qquad (4)$$

Where δ and $\oplus$ represents an activation function and element wise addition, respectively. The MS-SGC module can be naturally extended to multi scale temporal graph convolution (MS-TGC) module by using a similar structure with the only difference that sub-spatial graph convolutions are replaced with sub-temporal convolutions represented as T_i.

3) Separable Convolutions

To reduce computation costs, separable convolution has been introduced that decomposes standard convolutions into two distinct steps: depthwise convolution, applied to individual channels, and point-wise convolution, which adjusts channel numbers using 1×1 convolutions [9]. To compare the computational complexity of standard and separable convolution approaches, consider an input feature map of size $L_f \times L_f \times C_i$ where L_f represents spatial width and height and C_i represents number of input channels. Upon the application of a standard convolution with a filter size of $L_k \times L_k \times C_o$, an output feature map of size $L_f \times L_f \times C_o$ is generated at the computational cost of

$$L_k \times L_k \times C_i \times C_o \times L_f \times L_f \qquad (5)$$

In contrast, the separable convolution approach has two components depthwise convolution with a computational cost of $L_k \times L_k \times C_i \times L_f \times L_f$ and a pointwise convolution with computational cost of $C_i \times C_o \times L_f \times L_f$ resulting in a total computational cost computed as the summation of two.

$$L_k \times L_k \times C_i \times L_f \times L_f + C_i \times C_o \times L_f \times L_f \qquad (6)$$

Eq. (6) shows a substantial decrease in computational complexity, which is especially advantageous when the number of input and output channels is high.

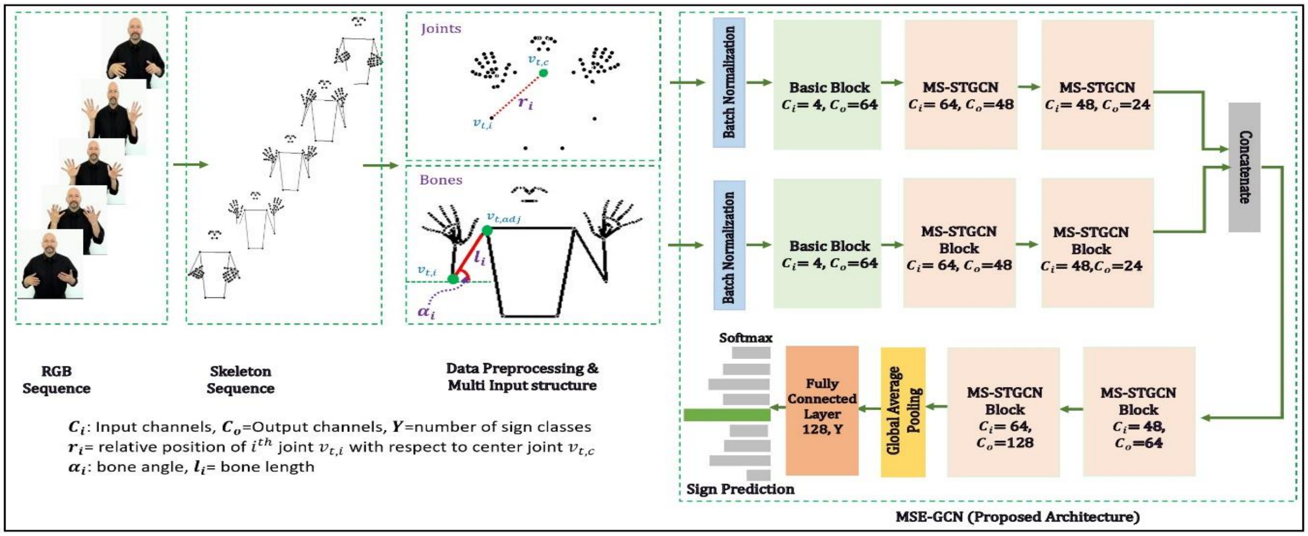

Joints
$v_{t,c}$
r_i
$v_{t,i}$
Bones
$v_{t,adj}$
l_i
$v_{t,i}$
α_i
RGB Sequence
Skeleton Sequence
Data Preprocessing & Multi Input structure
C_i: Input channels, C_o=Output channels, Y=number of sign classes
r_i= relative position of i^{th} joint $v_{t,i}$ with respect to center joint $v_{t,c}$
α_i: bone angle, l_i= bone length
Batch Normalization
Basic Block C_i= 4, C_o=64
MS-STGCN C_i= 64, C_o=48
MS-STGCN C_i= 48, C_o=24
Batch Normalization
Basic Block C_i= 4, C_o=64
MS-STGCN Block C_i= 64, C_o=48
MS-STGCN Block C_i= 48,C_o=24
Concatenate
Softmax
Fully Connected Layer 128, Y
Global Average Pooling
MS-STGCN Block C_i= 64, C_o=128
MS-STGCN Block C_i= 48, C_o=64
Sign Prediction
MSE-GCN (Proposed Architecture)

Figure 2. Complete Pipeline of Proposed architecture

B. MSE-GCN Framework

1) Skeleton Extraction and Details

Provided the RGB video sequence of dimension $\boldsymbol{X_{video}} \in \boldsymbol{\mathcal{R}^{C\times T\times H\times W}}$, where C, T, H, and W represent channels, frames, height, and width of each frame respectively, we use an off-the-shelf pose extractor MediaPipe [26] to extract the data of skeleton joints. We used MediaPipe hand and pose extractor modules to acquire 2D joints data for both hands (21 joints for each hand), both arms , and facial joints (mouth, eyebrows, and nose). In total, data of 65 joints is used and a skeleton sequence of dimension $\boldsymbol{X_{skeleton}} \in \boldsymbol{\mathcal{R}^{C\times T\times V}}$, where C is joints features (x and y coordinates), T and V represent frames and number of joints respectively, is fed as input to the pre-processing module to generate a multi-input structure.

2) Data Pre-Processing and Multi input Structure

Data preprocessing is an essential step for pose-based sign language recognition (SLR). Suppose the 2D skeleton sequence of a sign is represented as $\boldsymbol{X} \in \boldsymbol{\mathcal{R}^{C\times T\times V}}$, then to construct a multi-input architecture, the sign sequence is distributed into two inputs representing joints and bones. The joints input is formed by concatenating X (original joints) with R where R represents the relative position set of each sequence and is obtained as Eq. (7).

$$R = \{\{ r_i = v_{t,i} - v_{t,c}\}|\ i = 1,2, \ldots., V, t \le T\} \qquad (7)$$

Where $v_{t,i}$ represents the ith joint in tth frame and $v_{t,c}$ represents the center node computed as the center point between left and right shoulder joints i.e.

$$v_{t,c} = \frac{v_{t,left\_shoulder} + v_{t,right\_shoulder}}{2}$$

The input for bones stream is formed by computing bones features consisting of bones lengths B_L and bones angles B_A. These bones features are calculated using Eqs. (8) and (9).

$$B_L = \{\{ l_i = v_{t,i} - v_{t,adj}\}|\ i = 1,2, \ldots., V, t \le T\} \qquad (8)$$

where $\boldsymbol{v_{t,adj}}$ represents the adjacent joint of the ith joint.

$$B_A = \{\{ \alpha_i = \arccos (\frac{v_{t,i} - v_{t,adj}}{\sqrt{\Sigma v_{t,i}^2}})\}|\ i = 1,2, \ldots., V, t \le T\} \ (9)$$

These bone features are sent as input to the second stream after concatenation.

3) Multi-Scale Efficient GCN Architecture

Following the data preprocessing module, two distinct outputs: joints and bones are generated which are fed to the proposed multi-input architecture inspired by [27, 28] using an early fusion strategy. The proposed architecture preserves the input information and reduces model complexity significantly as compared to late fusion schemes. A complete pipeline of proposed architecture is presented in Fig. 2.

4) Sub Blocks Details

In this work, we introduce the MSE-GCN architecture, which comprises a basic block and several MS-STGCN blocks. Further details of each subblock are presented in subsequent sections.

Basic Block: A basic block is implemented using a standard graph convolution followed by a standard 2D temporal convolution layer as explained in Eqs. (1) and (2). Each layer is followed by a batch normalization and a ReLU activation layer.

MS-STGCN Block: The proposed architecture for MS-STGCN block draws inspiration from MS-G3D [29], Efficient GCN [28], and MS-GCN [11]. For the implementation of MS-STGCN blocks, we employ an ordered stacking approach that involves combining multi-scale spatial graph convolution layer (MS-SGCN) layer, Multiple multi-scale temporal convolution (MS-TCN) layer, and a spatiotemporal joint part attention (ST-JPA) module as shown in Fig. 3. Additionally, to facilitate model optimization and stable training, residual links are used for each layer. The depth of each block is dependent upon the number of MS-TCN layers stacked inside. Specifically, the depth is set as 1

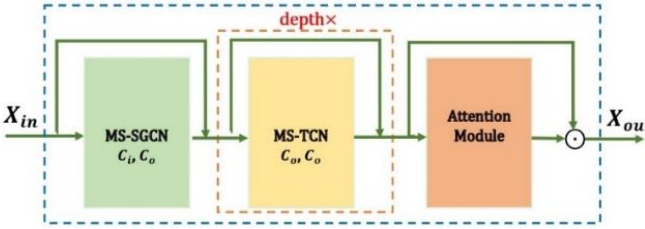


Figure 3. Implementation of MS-STGCN block (depth is the number of stacked MS-TCN layers inside a MS-STGCN block)

(MS-TCN layer is stacked once) for the basic block and is established as 2 for pre-concatenation MS-STGCN blocks. Post-Concatenation MS-STGCN blocks, on the other hand, are designated a depth of 3. The selection of these depth values adheres to the guidelines outlined in the EfficientGCN-B4 architecture [28].

MS-SGCN Layer: In detail, a MS-SGCN layer is implemented using sandglass(SG) layers which are formed by using depth convolutions and point convolutions as explained in section III.A. These separable convolution layers are arranged in a multi scale structure. The complete implementation of the MS-SGCN layer is presented in Fig.4. The spatial filter size $\mathbf{K_s}$ is selected as maximum graph distance D, which is set as 3, $\boldsymbol{r}$ represents reduction rate and is chosen as 2, and $\boldsymbol{k_s}$ represent spatial scale which is set at 4. This configuration reduces the number of trainable parameters significantly and efficiently captures local and global spatial relationships amongst skeleton joints.

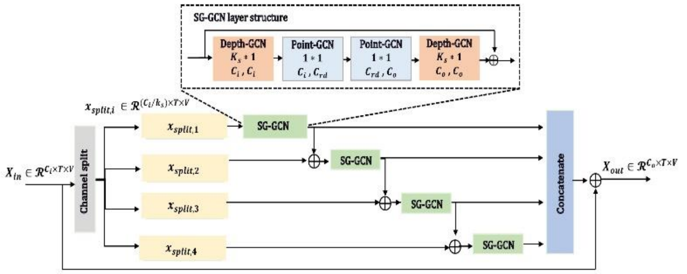


Figure 4. Implementation of MS-SGCN layer (K_s = spatial filter size, C_i=input channels, C_o=output channels, $C_{rd} = C_o/r$ (r is reduction rate), k_s= spatial scale)

MS-TCN Layer: For the implementation of MS-TCN layer, we have used two types of separable layer configurations: Sep Layer and SG layer which are inspired by [21] and [23] and are composed of depth convolution and point wise convolutions. These convolution layers are set in a multi-scale structure as shown in Fig. 5. $\boldsymbol{K_T}$ represents temporal kernel size and is selected as 5, $\boldsymbol{r}$ represents reduction rate and k_t represents temporal scale which is set at 4. The use of separable layer helps reducing the number of parameters and floating-point operations (FLOPs) significantly, whereas multi scale structure ensures generation of output feature maps with well captured long-range and short-range dependencies amongst the nodes along the temporal dimension.

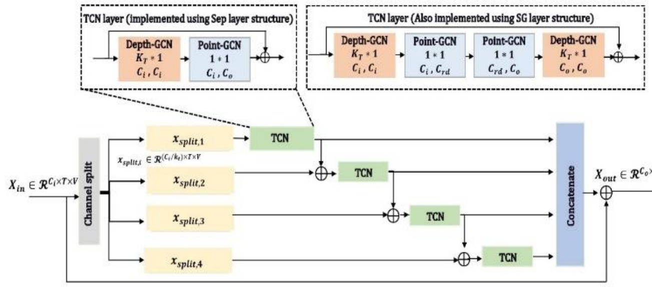


Figure 5. Implementation of MS-TCN layer (K_T= Temporal kernel, C_i=input channels, C_o=output channels, C_{rd}=C_o/r (r is reduction rate), k_t= temporal scale)

Spatiotemporal Joints Part Attention Module: Attention mechanism plays a significant role to identify the importance of each node for accurate recognition of sign. In literature, various attention modules i.e., joint attention, channel attention, frame attention, and part attention have been proposed for the task of activity recognition and gesture recognition. These modules learn attention weights on a single dimension i.e., channel, spatial or temporal and all other dimensions are globally averaged. In our work, we have proposed a novel ST-JP attention (ST-JPA) mechanism by efficiently combining a STJA mechanism proposed by [28] and a part attention mechanism [27]. STJA mechanism treats all the body parts present in an input sequence equally, whereas in sign recognition, some body parts i.e., hands have more significant movements as compared to facial and torso movements. Whereas part attention mechanism can learn the significance of each body part but cannot focus on individual joints. Our proposed ST-JPA mechanism overcomes these problems by intelligently combining the positive attributes of both mechanisms. It allocates higher attention weights to significant joints found in crucial temporal frames, particularly those pertaining to the most important body parts. This integration leads to notable improvements in recognition accuracy. Fig. 6 presents an overview of ST-JPA module.

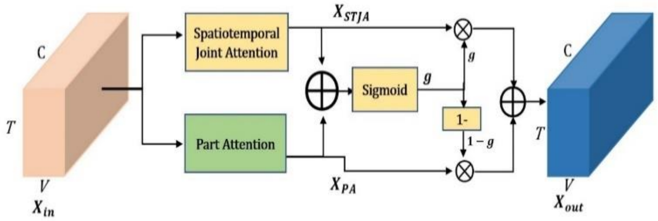


Figure 6. The overview of the proposed ST-JPA layer

IV. Experimental Results

In this section, we evaluate the performance of our proposed MSE-GCN architecture by comparing it against SOTA methods using RGB and skeleton-based features in terms of accuracy and computational efficiency. We also conduct ablation studies and offer visualizations and explanations to support the effectiveness of our proposed approach.

A. Datasets

We evaluate our proposed model on three challenging sign language datasets: WLASL [3], LIBRAS-UFOP [30], and MINDS-Libras [31]. The details of each dataset are provided in the subsequent section.

1) WLASL:

The WLASL dataset [3], is a comprehensively compiled collection of data sourced from various online platforms encompassing a wide range of signers, lighting conditions, and background variations. It is organized into four subsets: WLASL-100, WLASL-300, WLASL-1000, and WLASL-2000, with the numerical suffix indicating the number of sign classes present in each subset. Our study follows the same training, validation , and testing protocols as specified by dataset authors [3].

2) LIBRAS-UFOP:

LIBRAS-UFOP [30] is a Brazilian sign language dataset and comprises 3,040 data sequences categorized into 56 different sign classes. The 56 sign classes are distributed into four categories (Cat-1, Cat-2, Cat-3, and Cat-4), based upon the similarity in movement, hand configuration, and articulation point. In order to effectively compare our findings to those presented in [30], we adhered to the suggested training, validation, and testing protocols provided by the authors. The dataset is divided into five distinct sets by organizing the signers in such a way that the same signer cannot appear in more than one set. Each split consists of 3 signers grouped into a training set, one signer in the validation set, and one signer in the test set as presented in Table. 1. In each set, optimal parameters are found by training and validating the model using the data of signers present in train and validation set, respectively.

Table 1: Train, Validation, and Test Experimental sets for LIBRAS-UFOP dataset . **Note:** s1, s2, s3, s4, s5 represent signer 1, signer 2, signer 3, signer 4, and signer 5 respectively.

Experimental Sets	**Signer**		
	Train	**Validation**	**Test**
#01	s1, s2, s3	s4	s5
#02	s2, s3, s4	s5	s1
#03	s1, s4, s5	s2	s3
#04	s1, s2, s5	s3	s4
#05	s3, s4, s5	s1	s2

3) MINDS-Libras:

MINDS-Libras [31] is a Brazilian sign language dataset and comprises 1,200 data sequences categorized into 20 classes. We adopt the recommended setup proposed by [31] for training and evaluation of our model. The dataset is split into a 75:25 ratio for training and testing, following the protocol established by the dataset authors. To determine the optimal model parameters, we employ a k-fold cross validation with k=3 on the training set.

B. Implementation Details

In our study, the learning rate is initialized at 0.1 and follows a cosine schedule, decaying after the 10th epoch. To ensure a stable training process, a warmup strategy is implemented, that gradually increases the learning rate from 0 to the initial value over the first 10 epochs. we have utilized a stochastic gradient descent (SGD) with weight decay of 0.0001 and Nesterov momentum of 0.9 for model optimization. The model is trained for 350 epochs. The hyperparameters maximum graph distance D and temporal kernel size K_T are chosen as 3 and 5, respectively. To prevent overfitting, a dropout layer with the drop probability of 0.25 is inserted after the GAP layer. Swish activation is used as an activation function in all convolution blocks because of its smoothness and differentiability. All experiments are conducted on a single Nvidia RTX-3080 GPU using a PyTorch framework. The batch size is set as 16 and 64 frames are sampled from the input sequence using random sampling strategy and fed as input to the model.

C. Comparison with state-of-the-art methods

1) Results on WLASL Dataset

We compare our proposed model MSE-GCN with previously proposed SOTA methods on three subsets of WLASL dataset: WLASL-100, WLASL-300, and WLASL-1000. The comparison is performed based on the top-1 and top-5 per instance and per class accuracies for these subsets. The results obtained are presented in Table. 2.

Table 2. Experimental Results on the WLASL-100, WLASL-300, and WLASL-1000. Top-1 (T-1) and Top-5 (T-5)). ***Note:*** "-" means that test values have not been reported and bold representation highlights the best accuracies.

Method	**WLASL-100**				**WLASL-300**				**WLASL-1000**			
	Per-instance		**Per-class**		**Per-instance**		**Per-class**		**Per-instance**		**Per-class**	
	T-1	**T-5**	**T-1**	**T-5**	**T-1**	**T-5**	**T-1**	**T-5**	**T-1**	**T-5**	**T-1**	**T-5**
RGB-based												
I3D [3]	65.89	84.11	67.01	84.58	56.14	79.94	56.24	78.38	47.33	76.44	-	-
TCK [16]	77.52	91.08	77.55	91.42	68.56	89.52	68.75	89.41	-	-	-	-
Fusion 3 [32]	75.67	86.00	-	-	68.30	83.19	-	-	56.68	79.85	-	-
SignBERT[22]	82.56	94.96	83.30	95.00	74.40	91.32	75.27	91.72	-	-	-	-
BEST[33]	81.01	94.19	81.63	94.67	75.60	92.81	76.12	93.07	-	-	-	-
SignBERT+[21]	84.11	**96.51**	85.05	**96.83**	78.44	**94.31**	79.12	**94.43**	-	-	-	-
Skeleton-based												
Pose-GRU [3]	46.51	76.74	-	-	33.68	64.37	-	-	30.01	58.42	-	-
ST-GCN[6]	50.78	79.07	51.62	79.47	44.46	73.05	45.29	73.16	-	-	-	-
Pose-TGCN [3]	55.43	78.68	-	-	38.32	67.51	-	-	34.86	61.73	-	-
PSLR [7]	60.15	83.98	-	-	42.18	71.71	-	-	-	-	-	-
MOPGRU [5]	63.18	-	-	-	-	-	-	-	-	-	-	-
SPOTER [19]	63.18	-	-	-	43.78	-	-	-	-	-	-	-
SignBERT[22]	76.36	91.09	77.68	91.67	62.72	85.18	63.43	85.71	-	-	-	-
BEST[33]	77.91	91.47	77.83	92.50	67.66	89.22	68.31	89.57	-	-	-	-
SIGNGRAPH[20]	72.09	88.76	-	-	71.40	92.26	-	-	61.83	85.87	-	-
SignBERT+[21]	79.84	92.64	80.72	93.08	73.20	90.42	73.77	90.58	-	-	-	-
MSE-GCN(Ours)	**85.27**	94.96	**86.00**	95.58	**81.59**	93.41	**82.17**	93.81	**71.75**	**90.83**	**71.52**	**90.69**

In Table. *2*, I3D, TCK, Fusion 3, SignBERT, SignBERT+, and BEST are the representative recent appearance-based methods, and our model outperform these methods in terms of top-1 per instance accuracy by 1.16%, 3.15%, and 15.07% for WLASL-100, WLASL-300, and WLASL-1000 respectively. We also assess our model against the most recent skeleton-based methods, which include Pose-GRU, Pose-TGCN, GCN-BERT, MOPGRU, SPOTER, SignBERT, SignBERT+, BEST, and SIGNGRAPH. Our model exhibits significant performance improvements in comparison to these methods, with gains of 5.43%, 8.39%, and 9.92% for WLASL-100, WLASL-300, and WLASL-1000, respectively. The experimental results imply that the proposed architecture MSE-GCN exhibits robust performance, comparable to SOTA methods, indicating its effectiveness in the task at hand. We believe that the improved performance is caused by efficient aggregation of multi-scale local and global relationships between body joints, and the optimally scaled network. Moreover, the proposed ST-JPA module helps the model in identifying most informative joints that results in a significant increase in accuracy.

Confusion Matrix and Error Analysis:

From Table.2, it is evident that our proposed model demonstrates exceptional recognition performance when applied to extensive WLASL dataset. However, there are certain signs that our model struggles to identify accurately. To gain a more comprehensive understanding of these problematic cases, we constructed a confusion matrix as shown in Fig. 7.

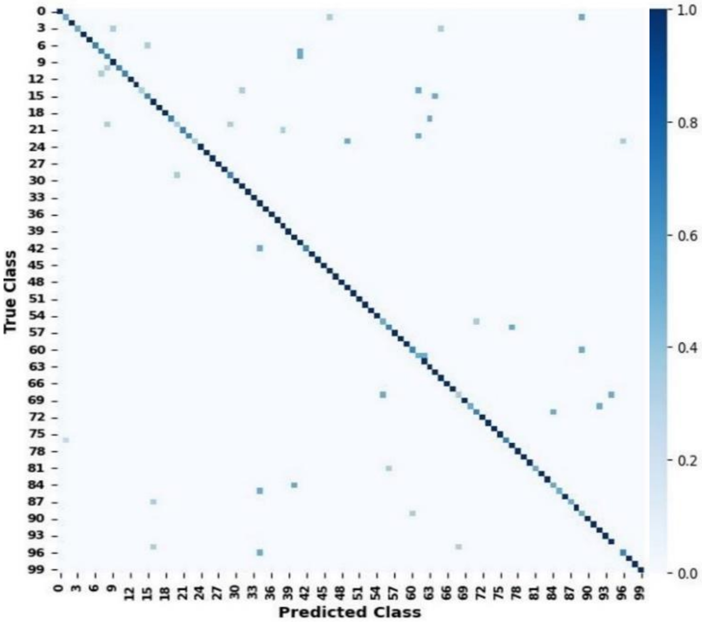


Figure 7. Confusion Matrix for WLASL-100 dataset.

The analysis of this matrix reveals that there are a few signs, namely ‘Right’, ‘Medicine’, and ‘Pizza’, for which recognition accuracies are comparatively low. Further investigation shed light on the underlying reasons for these inaccuracies. It was observed that these sign classes exhibit spatiotemporal movements that either closely resemble those of other classes or lack distinctiveness, thereby resulting in inaccurate predictions. For instance, the sign ‘Right’ shares an indistinguishable spatiotemporal pattern with the sign ‘Year’, leading to misclassification , as demonstrated in video 3 and video 4 presented in Fig. 8. Similarly, another instance of ‘Right’ is erroneously recognized as ‘Pizza’, as shown in video 1 and video 2 present in Fig. 8. A similar issue arises with the signs representing ‘Medicine’ and ‘Doctor’, as they possess identical spatiotemporal movement patterns, leading to misclassification. Collectively, these findings indicate that the main factor contributing to the failure in sign recognition is the existence of signs that share extremely similar movement dynamics in both space and time.

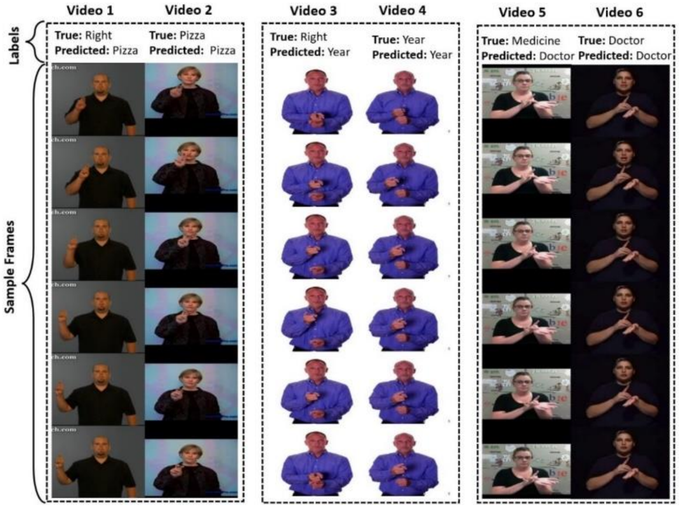

Video 1
Video 2
Video 3
Video 4
Video 5
Video 6
Labels
True: Right Predicted: Pizza
True: Pizza Predicted: Pizza
True: Right Predicted: Year
True: Year Predicted: Year
True: Medicine Predicted: Doctor
True: Doctor Predicted: Doctor
Sample Frames

Figure 8. Failure cases analysis

2) Results on LIBRAS-UFOP Dataset

In pursuit of a rigorous comparison with state-of-the-art (SOTA) approaches on the LIBRAS-UFOP dataset, our study meticulously adhered to recommended experimental protocols (Section IV.A). Our results, presented in Table 3, demonstrate the superior performance of our proposed method over SOTA techniques across various dataset schemes. Notably, when considering the skeleton input data (SCH1), our method exhibits a substantial performance advantage. Furthermore, we surpass the historically promising SCH6 approach, which employs a resource-intensive late fusion scheme, offering a more cost-effective solution.

Confusion Matrix for LIBRAS-UFOP dataset:

Table 3. Experimental results on LIBRAS-UFOP dataset. The best results are presented in bold.

Method	**Cat.1 (%)**	**Cat.2 (%)**	**Cat.3 (%)**	**Cat.4 (%)**	**All (%)±SD**
(SCH1) [30]	61.45±2.97	60.11±1.25	80.35±3.33	61.57±0.91	60.28±3.11
(SCH2) [30]	62.26±3.11	60.27±2.48	75.85±1.15	60.42±1.56	61.25±2.75
(SCH3) [30]	65.84±1.85	62.32±2.01	79.48±0.95	60.92±0.11	60.86±0.72
(SCH4) [30]	73.91±2.87	68.14±3.75	87.07±2.10	67.54±1.21	71.80±4.74
(SCH5) [30]	76.05±3.19	70.24±4.21	90.46±1.70	69.19±1.49	72.44±3.35
(SCH6) [30]	78.60±4.48	72.34±3.03	92.48±1.75	71.58±1.57	74.25±3.28
GEI [34]	58.27±2.88	72.22±0.65	**93.17±5.79**	82.36±7.21	64.91±3.79
MSE-GCN (Ours)	**81.28±6.10**	**91.27±5.50**	87.02±6.3	**89.58±5.76**	**88.59±3.60**

To further understand the performance of our proposed architecture, we constructed a confusion matrix as presented in Fig. 9. As explained in section IV.A, the signs in this dataset are grouped into four categories. The most significant recognition inaccuracies are observed for Cat.1 and for one class of Cat.4 . In Cat.1, the signs share a common type of movement at specific points of articulation, varying only in the way the hands are configured. The discrepancy in hand configuration can be as minimal as a single finger, i.e., in the signs for year1, year 2, and year 3 [30]. Therefore, 'year-1' is either recognized as 'year-1' or 'year-3'. Similarly, most of the instances of 'Takes a little care' are recognized as 'safe' which shares the same hand movements at the same articulation points with a difference in hand configurations. Upon further investigation, it was observed that while hand is pointing outwards, because of its own shadow, hand pose is not properly extracted thereby resulting in inaccurate predictions. Per class Precision, Recall, and F1 score can be found in Fig. 11(a).

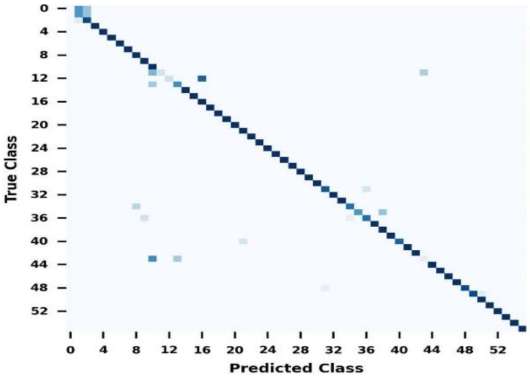


Figure 9. Confusion Matrix for LIBRAS-UFOP dataset.

3) Results on MINDS-Libras Dataset

Utilizing the same methodology as elucidated in section IV-A, we carried out an evaluation of our proposed method on the MINDS-Libras dataset. The ensuing comprehensive evaluation, presented in Table. 4 compares our approach against SOTA techniques employed on the same dataset. Our findings exhibit a superiority of our method over the SOTA methods, thereby substantiating the efficacy of our approach.

Table 4. Experimental results on MINDS-Libras dataset

Type	Method	Accuracy (%) ± SD
RGB-Based	CNN3D [35]	72.6
	CNN 3D [36]	93.3 ± 1.69
	GEI+SVD+SVM [34]	84.66 ± 1.78
Pose-Based	**MSE- GCN (Ours)**	**97.44 ± 1.01**

Confusion Matrix for MINDS-Libras dataset:

we constructed a confusion matrix for MINDS-Libras dataset as presented in Fig. 10, to elaborate the performance of our model on per class basis. Per class Precision, Recall, and F1 score can also be found in Fig. 11(b). Notably, our proposed approach attains a minimum performance of 82% across all assessed metrics, demonstrating an excellent outcome. This observation is further supported by a very high classification rate of 97.44% , evident in the confusion matrix depicted in Fig. 10.

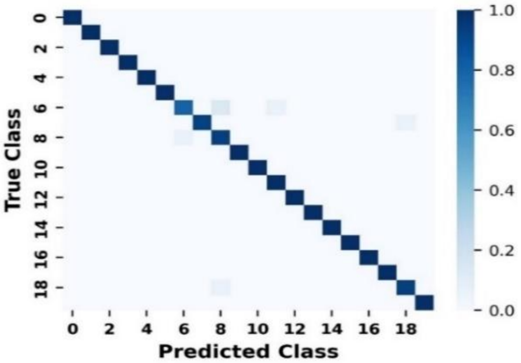
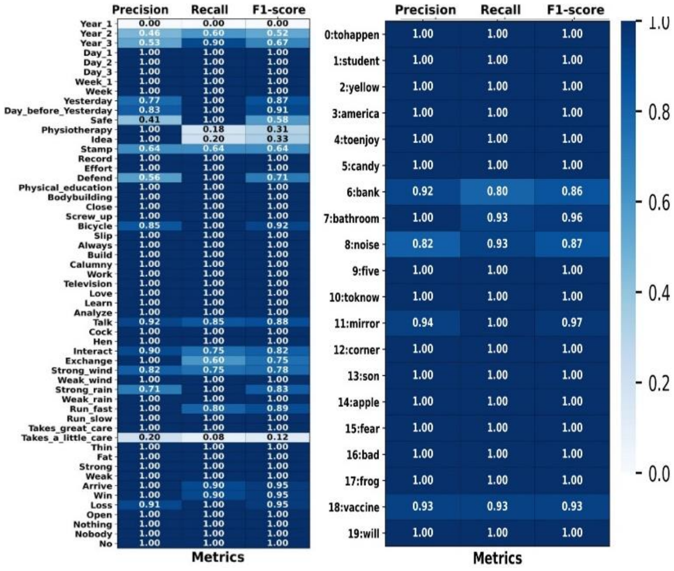


Figure 10. Confusion Matrix for MINDS-Libras dataset.



(a) (b)

Figure 11: Precision, Recall, and F1 score for (a) LIBRAS-UFOP dataset and (b) MINDS-Libras dataset.

D. Ablation Studies

This section primarily focuses on elucidating the individual contributions made by various components within the proposed MSE-GCN framework. These components include the selection of spatial and temporal layer structure, the significance of multi-scale strategy, the choice of an efficient attention module, and the importance of employing an early fused architecture. This section provides a rationale for incorporating these components with specific values in the main framework. The subsequent sections study the effect of each component individually.

1) Comparison of Spatial and Temporal Layers

In accordance with the details provided in section III, we have evaluated the proposed architecture using two distinct layer structures: Sep layer and SG layer. The outcomes of each configuration are presented in Table. 5. The SG layer is implemented with a reduction rate (*r*) of 2. As shown in the Table. 5, configuration 4 has comparable model size and inference speed, with 0.827 GFlops, in comparison to configuration 2. However, configuration 4 exhibits superior recognition accuracy, therefore rendering it an appropriate choice for our layer configuration. To further characterize the training process, we have sketched validation accuracy waveforms for each configuration and are shown in Fig.12(a). The sketched waveforms demonstrate the superiority of the chosen layer configuration over all others.

Table 5. Comparison of Spatial and Temporal Layer configuration on WLASL-300 dataset . M= number of parameters in millions (×10^6) and G= FLOPs in giga (×10^9).

	Spatial	Temporal	Accuracy (%)	FLOPs (G)	#Params (M)
1	SG	SG	78.74	0.937	1.383
2	Sep	Sep	80.84	0.818	1.338
3	Sep	SG	74.40	0.928	1.379
4	**SG**	**Sep**	**81.59**	**0.827**	**1.342**

2) Effective Scale of Multi Scale Architecture

To examine the effectiveness of our proposed multi-scale efficient GCN (MSE-GCN) architecture, we tested the proposed model with varying spatial and temporal scales $\boldsymbol{k_s}$ and $\boldsymbol{k_t}$. We conducted the experiments by varying $\boldsymbol{k_s}$ and $\boldsymbol{k_t}$ in a similar manner by choosing them as same values. When the scales were set to 1, the model utilized a standard spatiotemporal graph convolutional block, and when $\boldsymbol{k_s} = \boldsymbol{k_t} >1$, the model acquired the ability to aggregate information at multiple scales, resulting in multi-scale spatial and temporal representations. The outcomes of conducted experiments are presented in Table. 6. As the scale increased, MSE-GCN architecture exhibited a superior performance in terms of recognition accuracy and inference time. The most optimal outcomes were observed with scale set to 4, as this configuration effectively captured both local and global dependencies between joints. However, the further increase in scale resulted in performance degradation. The reason behind this can be attributed to the fact that when the scale is increased beyond an optimal value, the model starts to overfit on the training data and fails to generalize well. To put further light on our training process, validation accuracy waveforms under each setup are sketched and are presented in Fig. 12 (b).

Table 6. Comparison of MSE-GCN architecture for different scales on WLASL-300 dataset . M represents number of parameters in millions (×10^6) and G represents FLOPs in giga (×10^9).

$\boldsymbol{k_s}$ (Spatial Scale)	$\boldsymbol{k_t}$ (Temporal Scale)	Accuracy (%)	FLOPs (G)	#Params (M)
1	1	78.59	1.369	1.321
2	2	80.24	1.008	1.234
4	**4**	**81.59**	**0.827**	**1.342**
8	8	79.79	0.737	1.701

3) Fusion Stage

The results presented in Table. 7 show our model's performance for various fusion stages. The best results in terms of accuracy and computational complexity tradeoffs have been achieved for fusion stage 2. Fusing at the later stages increases model size as well as computational cost significantly. Moreover, recognition accuracy declines beyond fusion stage 2 because models overfits the training data and fail to generalize to validation data. Thus, we have chosen stage 2 as optimal feature fusion stage. To assess the training procedure for various fusion stages, validation accuracies are presented in Fig. 12 (c).

Table 7. Comparison of Fusion Stages on WLASL-300 dataset . M represents number of parameters in millions (×10^6) and G represents values in giga (×10^9)

Fusion Stage	Accuracy (%)	FLOPs (G)	#Params (M)
1st	80.39	0.787	1.254
2nd	**81.59**	**0.827**	**1.342**
3rd	79.04	1.004	1.568
4th	76.65	1.158	2.140

4) Attention Modules

In this work, we have proposed a novel spatiotemporal joints part attention (ST-JPA) module. In order to compare the effectiveness of proposed module with other attention modules namely: joint attention (JA) [37], part attention (PA) [27], spatiotemporal joint attention (ST-jointAtt) [28] and spatiotemporal channel attention (STCAtt) [11], we have conducted extensive experiments and the results are presented in Table. 8.

Table 8. Comparison of Attention Modules on WLASL-300 dataset . M represents number of parameters in millions (×10^6) and G represents values in giga (×10^9).

Attention Module	Accuracy (%)	FLOPs (G)	#Params (M)
No Attention	77.29	0.800	0.877
JA	77.69	0.809	0.850
PA	80.29	0.810	1.169
ST-JointAtt	79.94	0.818	0.997
STCAtt	80.19	0.819	1.054
ST-JPA	**81.59**	**0.827**	**1.342**

It is worth noting that inclusion of attention modules led to a noticeable improvement in accuracy. Among the various modules employed, the proposed ST-JPA module achieved the highest

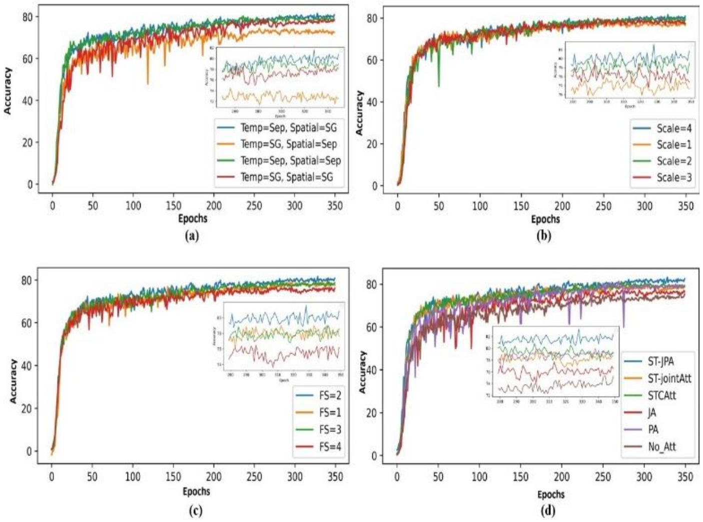


Figure 12. Ablation study waveforms: (a) Effect of separable Layers on Validation Accuracy. (b). Effect of various scales used for multi-scale architecture on Validation Accuracy. (c). Effect of Fusion stage (FS) on Validation Accuracy. (d). Effect of various attention modules on Validation Accuracy

accuracy. The waveform graphs depicting the validation accuracy for each attention module are presented in Fig. 12 (d) and provide additional clarity regarding our training process. These graphs clearly illustrate that incorporation of ST-JPA module boosts the model performance significantly because of focusing on the joints of most significant body parts in relevant frames.

E. Computational Performance Analysis

Model complexity refers to the level of intricacy and sophistication embedded within a computational model. It encompasses factors such as the number of parameters, depth of the model architecture, and computational resource required. Balance of model's accuracy and complexity is crucial, as overly complex models lead to overfitting. To assess the efficacy of our proposed model, we evaluate its performance on WLASL-300 dataset by comparing it to SOTA methods in terms of accuracy, and model complexity (number of parameters and floating-point operations (FLOPs). Fig. 13 presents the computational performance comparison of proposed model with alternative methods.

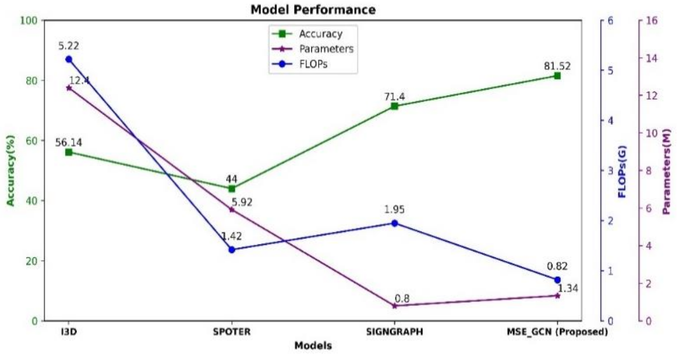


Figure 13. Computational Complexity Analysis

Our proposed model outperforms other SOTA methods by 10.12 % in accuracy with 1.73× lesser FLOPs. Although the number of parameters is comparable to SIGNGRAPH, our proposed model (MSE-GCN) has much faster inference time and higher accuracy making it a suitable choice for faster and efficient skeleton based SLR.

F. Visualization & Explanation

To demonstrate the operational characteristics of our model, we employ the class activation map [38] to calculate the activation maps of several skeleton sequences. The resultant maps are illustrated in Fig. 14, showcasing the activated joints across various frames sampled from the original sequence. Observing the Fig, it becomes apparent that the MSE-GCN model adeptly prioritizes the most informative joints that carry significant relevance. For instance, when a sign for bed is performed, the signer tilts his head towards the right and places it on the palm of his right hand. Our model demonstrates accurate activation of these highly relevant joints of head and right hand. Similarly, in the instances representing signs for 'study', 'book', and 'how', both hands exhibit substantial spatiotemporal movements. Consequently, our model appropriately concentrates on these specific parts and associated joints and assigns them the highest weights. While the sign for 'Later' is performed, the most significant motion is observed in the thumb and index finger of the right hand. While the other fingers of the right hand also undergo significant spatial and temporal changes and are consequently activated to a notable degree, the thumb and index finger exhibit the highest activation weights. These findings collectively indicate the effective functioning of our proposed model and attention mechanism.

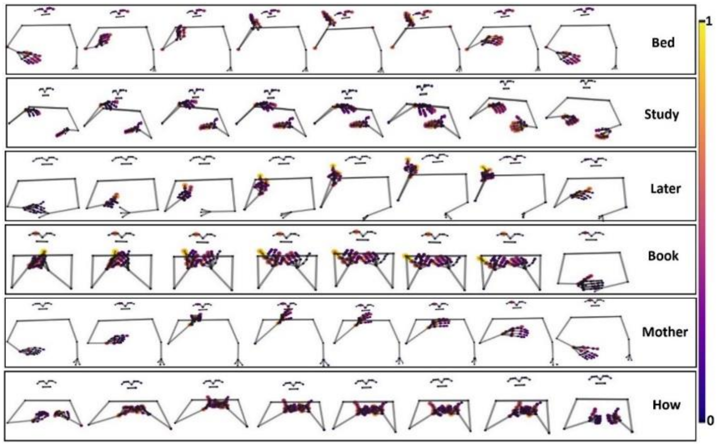


Figure 14. Activated Joints in eight frames of MSE-GCN for the sample signs: Bed, Study, Later, Book, Mother, and How. The bigger scale and brighter colors represent more activated joints.

V. Conclusion

In this paper, we propose a novel multi scale efficient graph convolution network (MSE-GCN) for skeleton-based SLR. The multi scale architecture in the proposed network enhances the spatiotemporal reception fields by the decomposition of a local graph convolution into multiple sub graph convolutions and creating a hierarchical residual architecture. It enables each node to perform multiple spatial and temporal aggregations with its adjacent nodes and in turn effectively capture both short-range and long-range spatiotemporal dependencies within the sequence. The utilization of separable convolution layers along the spatial and temporal dimensions tends to reduce the computational complexity significantly. To reduce the number of model parameters and eliminate redundant features, we employ an early fusion strategy that fuses the features from both input streams. Furthermore, we also introduce a novel hybrid Spatiotemporal Joints Part Attention (ST-JPA) mechanism that enhances the model's performance by assigning higher attention weight to the most significant parts and joints. Through extensive experiments, we demonstrate that the proposed MSE-GCN network achieves SOTA performance on three large scale SL datasets: WLASL, LIBRAS-UFOP, and MINDS Libras. Moreover, our model achieves these results with fewer FLOPs and a smaller number of parameters comparatively. In future work, we plan to extend our work by incorporating object appearance information to efficiently recognize the most similar signs.

References

[1] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, "A simple multi-modality transfer learning baseline for sign language translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5120-5130.

[2] A. Duarte, S. Albanie, X. Giró-i-Nieto, and G. Varol, "Sign language video retrieval with free-form textual queries," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14094-14104.

[3] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459-1469.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299-6308.

[5] B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K.-H. Park, and J. Kim, "An integrated mediapipe-optimized GRU model for Indian sign language recognition," *Scientific Reports,* vol. 12, pp. 1-16, 2022.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[7] A. Tunga, S. V. Nuthalapati, and J. Wachs, "Pose-based sign language recognition using gcn and bert," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 31-40.

[8] N. Naz, H. Sajid, S. Ali, O. Hasan, and M. K. Ehsan, "SIGNGRAPH: An efficient and accurate pose-based Graph Convolution approach towards Sign Language Recognition," *IEEE Access,* 2023.

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

[10] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305-321.

[11] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 1113-1122.

[12] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE transactions on pattern analysis and machine intelligence,* vol. 42, pp. 2306-2320, 2019.

[13] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *International Journal of Computer Vision,* vol. 126, pp. 430-439, 2018.

[14] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10023-10033.

[15] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202-6211.

[16] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring cross-domain knowledge for video sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6205-6214.

[17] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 2020, pp. 35-53.

[18] Y. Min, Y. Zhang, X. Chai, and X. Chen, "An efficient pointlstm for point clouds based gesture recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5761-5770.

[19] M. Boháček and M. Hrúz, "Sign Pose-based Transformer for Word-level Sign Language Recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 182-191.

[20] N. Naz, H. Sajid, S. Ali, O. Hasan, and M. K. Ehsan, "Signgraph: An Efficient and Accurate Pose-Based Graph Convolution Approach Toward Sign Language Recognition," *IEEE Access,* vol. 11, pp. 19135-19147, 2023.

[21] H. Hu, W. Zhao, W. Zhou, and H. Li, "SignBERT+: Hand-model-aware Self-supervised Pre-training for Sign Language Understanding," *IEEE transactions on pattern analysis and machine intelligence,* 2023.

[22] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, "Signbert: pre-training of hand-model-aware representation for sign language recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11087-11096.

[23] H. Hu, W. Zhou, and H. Li, "Hand-model-aware sign language recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1558-1566.

[24] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413-3423.

[25] F. B. Slimane and M. Bouguessa, "Context matters: Self-attention for sign language recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 7884-7891.

[26] Google. (2022, Deecember 31). *MediaPipe Holistic*. Available: https://google.github.io/mediapipe/solutions/holistic

[27] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1625-1633.

[28] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 45, pp. 1474-1488, 2022.

[29] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143-152.

[30] L. R. Cerna, E. E. Cardenas, D. G. Miranda, D. Menotti, and G. Camara-Chavez, "A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a microsoft Kinect sensor," *Expert Systems with Applications,* vol. 167, p. 114179, 2021.

[31] T. M. Rezende, S. G. M. Almeida, and F. G. Guimarães, "Development and validation of a Brazilian sign language database for human gesture recognition," *Neural Computing and Applications,* vol. 33, pp. 10449-10467, 2021.

[32] A. A. Hosain, P. S. Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka, "Hand pose guided 3d pooling for word-level sign language recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3429-3439.

[33] W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, "BEST: BERT Pre-Training for Sign Language Recognition with Coupling Tokenization," *arXiv preprint arXiv:2302.05075,* 2023.

[34] W. L. Passos, G. M. Araujo, J. N. Gois, and A. A. de Lima, "A gait energy image-based system for Brazilian sign language recognition," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 68, pp. 4761-4771, 2021.

[35] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Sign language recognition based on hand and body skeletal data," in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2018, pp. 1-4.

[36] J. A. Shah, "Deepsign: A deep-learning architecture for sign language," 2018.

[37] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921-2929.