MIPA-ResGCN: A Multi-Input Part Attention based Residual Graph Convolution Network for Sign Language Recognition

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

***Abstract*— Sign language (SL) is used as primary mode of communication by individuals who experience deafness and speech disorders. However, SL creates an inordinate communication barrier as most people are not acquainted with it. To solve this problem, many technological solutions using wearable devices, video, and depth cameras have been put forth. However, the ubiquitous nature of cameras in contemporary devices has resulted in the emergence of sign language recognition (SLR) using video sequence as a viable and unobtrusive substitute. In this study, we utilized 2D videos as the primary sensing modality from which we extract pose sequences. Our proposed approach comprises three key stages: pose extraction, handcrafted feature generation, and feature space mapping and recognition. Initially, an efficient off-the-shelf pose extraction algorithm is employed to extract pose information of different body parts of each subject in a video. Then, a multi-input stream has been generated using handcrafted features i.e., joints, bone lengths, and bone angles. Finally, an efficient and lightweight model based on a residual graph convolution network (ResGCN) along with efficient attention mechanisms is proposed to encode body's spatial and temporal motion in a compact feature space and recognize the signs performed. In addition to enabling effective learning during model training and offering cutting-edge accuracy, the proposed model significantly reduces computational complexity. Our proposed method is assessed on five challenging SL datasets, WLASL-100, WLASL-300, WLASL-1000, LSA-64, and MINDS-Libras, achieving state of the art (SOTA) accuracies of 83.33%, 72.90%, 64.92%, 100±0%, and 96.70± 1.07%, respectively using pose-based method. Compared to previous approaches, we achieve superior performance while incurring a lower computational cost.**

***Index Terms*—** ResGCN, Pose Sequence Modeling, SLR, Part Attention, Visualization, Muli Input Architecture.

(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.

I. INTRODUCTION

Sign languages (SLs) are non-verbal forms of communications used by deaf and speech impaired people all over the world to communicate with audially unaffected individuals. These languages are largely communicated by physical movements of hands and arms, but head, lip, eye, and brow movements are also very helpful. The visual signals are mainly generated by hands and body and are decoded by the eyes [1]. Sign language recognition (SLR) endeavors to translate these visual signals produced by individuals who communicate using sign language into speech or text, ultimately serving as a medium to establish effective communication between them and audially unaffected people. This, in turn, increases accessibility of resources for the deaf and speech impaired population, providing them with more opportunities. Because of this, automated SLR is a particularly intriguing area of research that calls for knowledge in both computer vision and natural language processing to effectively comprehend the spatiotemporal linguistic constructions of performed signs. There are more than 300 sign languages used worldwide and each one of them has its own fundamental structure, grammar, as well as subtleties and articulators that enable its users to express themselves successfully. SLR can also play a significant role for human-computer interaction (HCI) to encourage interaction between people and machines. Isolated SLR (ISLR) and Continuous SLR (CSLR) are two subcategories of Sign Language Recognition (SLR). While CSLR processes entire utterances comprising multiple sign glosses for translation, ISLR classifies individual sign records into corresponding gloss categories.

A sign gloss is made up of manual features i.e., hand shape, palm orientation, and precise hand motions and non-manual features i.e., facial expressions and body posture [2] as shown in Fig.1.

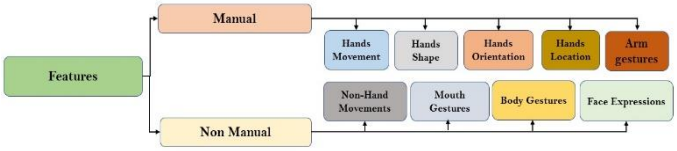


Fig. 1. Manual and Non-Manual sign language (SL) features

As these features cover small areas in the entire video frame, the background clutter can easily distract the network from learning discriminative spatiotemporal features and degrading model's performance. Additionally, videos have redundancy along temporal dimension and motions between adjacent frames are not significant which makes it difficult for models

to learn embeddings by focusing on significant spatial and temporal regions. Despite the significant strides made by deep learning in advancing SLR systems, there remain these significant challenges that impede the full realization of their potential and continue to pose significant obstacles to the development of highly accurate and generalizable SLR systems. Prior studies have demonstrated that the utilization of pre trained 2-Dimensional convolutional neural networks (2-D CNNs) as frame level spatial features extractors, along with a subsequent late fusion of these extracted features, can enhance the performance of video classification tasks to a significant extent. However, this approach ignores the temporal dependencies between neighboring frames and thus leads to poor recognition performance [3, 4]. To accurately capture the temporal information, the idea of feeding high level spatial features extracted using 2-D CNNs to recurrent neural networks (RNNs) has been investigated in [5-8]. To effectively capture low level and high level spatial and temporal features, 3-D CNNs have also been employed in [9-11]. However, 3-D CNNs have the drawback that they have a large computational cost and are thus difficult to tune. Moreover, they also suffer from optimization problems because of joint time and space modeling.

Some studies have focused only on the hand's motions. Hands are segmented from video frames using external tools such as hand detection [4, 5, 12] and fed as input to the deep learning models, but the performance of these approaches is strongly reliant on these techniques and overlooks nonmanual aspects. Various attention mechanisms have also been investigated and have shown significant improvements for SLR [13-17]. Several studies [18-21] leverage human pose information obtained using efficient pose extraction algorithms. A Human pose is composed of skeletal joints landmarks and bones connecting these joints. Models that employ skeletal data as input are light, compute-efficient, and have comparable accuracies all of which greatly boost their potential for usage in daily life.

In this work, we propose a novel three-step approach for enhancing the accuracy and efficiency of SLR. Our proposed approach leverages the pose information of human hands and upper body as inputs. To effectively model the spatial features and temporal dependencies in SL, we propose a multi-input graph convolution network with enhanced attention mechanism. As a baseline, we adopt the ResGCN [22] model which is based on spatiotemporal graph convolution (ST-GCN) blocks and employs residual connections for dimensionality reduction. We employ a multi-input network and an early fusion scheme to reduce computational complexity. Moreover, a novel part-based attention model is developed to eliminate irrelevant information and extract additional discriminative spatiotemporal features by focusing on most significant body parts and joints in a sign sequence. Experiments demonstrate that our proposed model exhibits substantial performance enhancement over SOTA SLR methods on the WLASL, LSA-64, and MINDS-Libras datasets while being computationally efficient.

The rest of this article is structured as follows. Section-II reviews the related works for SLR. In Section-III, each subpart of our proposed SLR method is described in detail. Experimental results, computational efficiency analysis, ablation studies, and Visualization and Explanations are presented in Section-IV. Finally, section-V concludes the paper.

II. Related Work

Advancements in deep learning architectures, coupled with the availability of high-performance computing resources, have enabled the development of deep models capable of processing multimodal data for SLR. The field of automatic SLR shares certain areas of overlap with action recognition, leading to a considerable influence of action recognition network designs on methods proposed for addressing the SLR problem. The three primary components of the automatic SLR problem involve the selection of the suitable input modality, the extraction of spatiotemporal features from the input data, and classification based on these features. Various approaches have been proposed for each of these phases which can be explained under four categories:

- Input Modality
- Spatiotemporal Feature Extraction
- Attention Mechanisms for SLR
- Sign Language (SL) Datasets

A. Input Modality

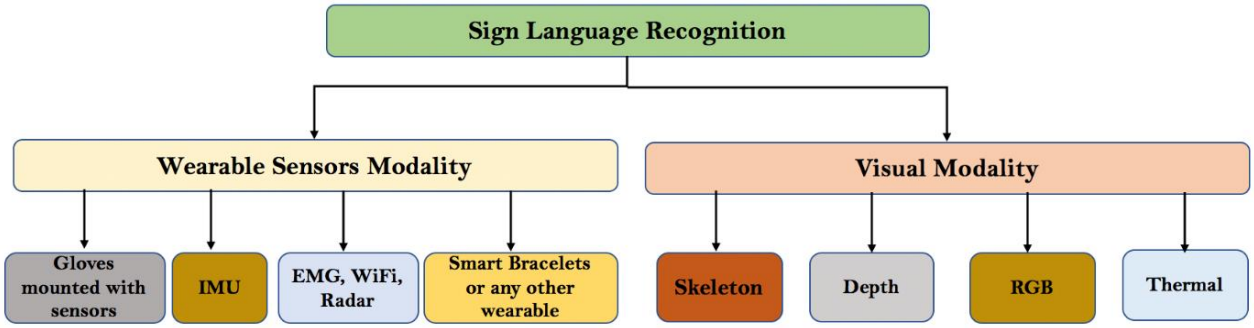


Fig. 2. Data Modalities used for sign language recognition.

In general, the two main input modalities considered for SLR are wearable sensors based and vision based as shown in Fig. 2. Glove-based models utilize specific mechanical or optical sensors affixed to a glove to leverage electrical signals for identifying hand positions, whereas vision-based models employ video data of the signers to recognize different signs. Further details are provided in respective subsections.

1. Glove-Based/Wearable Sensors Modality

In literature, various wearable sensors-based modalities have been used to capture the spatiotemporal motion patterns of performed signs. i.e. Hand tracking is done using an electromyography (EMG) bracelet called MYO in [23] and K-NN and SVM were used for sign classification. In [24], a wearable system for SLR has been proposed that fuses the information from surface EMG (sEMG) sensors and an inertial sensor. A wearable system consisting of six inertial sensors and a hand glove equipped with ten flex sensors has been proposed in [25] for signer independent SLR. Nevertheless, these techniques are highly invasive, limit movement, and encroach upon daily activities. In [26], a multi-frequency RF sensor network was suggested for the recognition of American sign language. Subsequently, a Short-Time Fourier Transform was implemented to detect distinct motion patterns in RF data attributed to the micro-Doppler effect, and machine learning algorithms were employed to analyze the linguistic properties of this data. The recognition accuracy of this technique is notably lower in comparison to other methodologies.

2. Visual Modality

In recent years, several visual data modalities have been used in the field of SLR. Two popular forms of data utilized in models for SLR are RGB and depth. While high-resolution material is included in RGB photos or videos, depth inputs encompass accurate details regarding the displacement between the image plane and the corresponding object. Some hybrid models have combined these modalities. Another comparatively less popular modality is thermal modality [27] that uses infrared thermal sensors for imaging objects and scenes. Another data modality that has been frequently employed by researchers is flow information, which refers to the motion aspects of each pixel in a video sequence. Mainly, two types of flow information i.e., optical flow (OF): A displacement vector of pixel coordinates in RGB sequence and scene flow (SF): A dense or semi dense 3D motion field of a scene in a depth sequence, are used. However, all these modalities suffer from high data dimensionality problem. In recent years, skeleton/pose based modality [18-21] has gained significant attention because of low data dimensionality. Pose consists of encoded form of joint sequences. However, till now pose based models were not able to surpass appearance-based methods in terms of accuracy. For our work, we have used pose data as input modality and proposed model is able to achieve SOTA accuracies for SLR.

B. Spatiotemporal Feature Extraction for SLR

In the conventional approaches, spatial representations were created using manually crafted features such as Histogram of Gradients, Scale Invariant Feature Transform, motion velocity vector, and frequency domain features [28-31]. To address disparate frame rates and account for temporal dependencies, approaches such as Condition Random Fields, Hidden Markov Models [32, 33], and Dynamic Time Warping were employed. With the emergence of deep learning frameworks, several vision tasks including object detection, image classification, and action recognition have greatly benefitted from 2-D CNNs. Image-based 2-D CNN models have been extended to video tasks using late fusion methods for action class prediction [3].

In recent years, 3-D CNNs are able to encode spatial and temporal information effectively and accurately making them a suitable choice for appearance based SLR [13, 18, 34]. The C3D [35] model was the inaugural 3-D CNN introduced for action recognition. The I3D [9] design, which was employed for SLR in [18, 36], was one of many 3D CNN action recognition architectures for SLR adaptations that quickly followed. A method for recognizing and teaching sign language is proposed in [37]. This SLR system uses a spatiotemporal network to perform the semantic classification of a provided sign language video, and an educational system is proposed to detect the learners' failure modes and provide instruction on the appropriate signing techniques. In [38], the problem of SLR is solved under a zero-shot learning paradigm. From sign language dictionaries, auxiliary information in the form of textual sign descriptions and characteristics is gathered and is used for knowledge transfer. Some architectures have utilized a single-colored motion history image (MHI) [39] to encode the entire sign video and then applied an I3D model to capture spatiotemporal dependencies. To learn more complicated motions inside the signing area and to disregard the background of the videos, depth cameras have also been examined as a possible tool for this job. Previous studies have utilized ensemble models such as conditional random fields [32] or multi-layered random forests [40] on top of depth representations. However, methods using 3-D CNNs and RGB or depth modalities are compute intensive.

Due to reduced dimensionality of human skeletal joints, pose-based SLR is gaining researchers interest. These methods are predicated on the idea that the signer's body, hands, and, in certain cases, face may provide sufficient information required to identify the performed sign. Using Pose data, two baselines using gated recurrent unit (GRU) and temporal graph convolution network (TGCN) have been proposed in [18]. In [19], spatial features are extracted using GCNs and temporal features are extracted using a BERT model, and final predictions are generated using late fusion scheme. A pose-based transformer architecture is proposed for SLR in [20]. A modified GRU is proposed in [21] to encode spatiotemporal relationships and is tested with pose based SL data. Although these methods significantly reduce computational complexity but are currently less accurate than appearance-based methods.

C. Attention Mechanisms

Attention plays a vital role in the way humans perceive information. Concentration on task-critical discriminative

TABLE I
SIGN LANGUAGE DATASETS INCLUDING RGB VIDEOS

Year	Dataset	Country	Class Numbers	Subjects	Samples
2012	DGS [41]	Germany	40	15	3000
2016	LSA-64 [42]	Argentina	64	10	3200
2019	MS-ASL [36]	USA	1000	222	25,513
2020	AUTSL [43]	Turkey	226	43	36,302
2020	WLASL [18]	USA	2000	119	21,803
2021	MINDS-Libras [44]	Brazil	20	12	1155

information is a hallmark of the attention process. Attention mechanisms i.e., spatial, temporal, channel, and self-attentions assist models to focus on the most significant information and thereby improving model's performance. Various attention schemes have been proposed for SLR. For large-vocabulary isolated SLR, an attention-based C3D [13] has been proposed which employs multimodal inputs. A multi-head attention-based transformer encoder was proposed by [14] for SLR. In [15], a self-attention mechanism has been proposed for efficient aggregation of hand features with their appropriate spatiotemporal context to effectively recognize sign language. Transformer-based encoder-decoder structures with channel wise self-attention and multichannel attentions were proposed in [16, 17] for sign language translation.

D. Sign Language Datasets

There are more than 300 sign languages used by individuals with hearing and speech disabilities. A variety of publicly accessible datasets are available for SLR. These datasets vary based on regional sign languages, continuous or isolated sign languages, data sizes, signer counts, data collection methods, and signer dependencies. Table I lists the most recent and relevant visual isolated SL datasets. For each dataset, six variables including Year, name of dataset, Country, total count of sign classes, number of signers, and total count of video samples, are specified. Although these datasets target various sign languages, American Sign Language (ASL) has garnered increasing attention owing to its popularity and usage. As shown in Table I, it would be preferable to increase the number of sign categories to have a more accurate generalization of the proposed approaches for practical applications. Therefore, we have tested our model on WLASL-100, WLASL-300, WLASL-1000, LSA-64, and MINDS-Libras datasets and have established SOTA accuracies.

III. METHODOLOGY

In this section, we provide details of our proposed pipeline and individual components of proposed architecture. Our proposed pipeline as shown in Fig. 3 consists of three stages. The first stage is extracting hand and body pose information from RGB video sequence. The second stage deals with data pre-processing and frame sampling. After pre-processing, multi-inputs consisting of joints and bones information are created and forwarded to MIPA-ResGCN architecture for spatiotemporal feature extraction. Finally, a class label is predicted for the provided sequence. Further details of the proposed approach are provided in the subsequent sections.

A. Stage 1: Pose Extraction

In the past, several techniques have been proposed to determine a human pose from RGB photos or video sequences [45-48]. It is crucial to have a reliable pose estimator because SLR is reliant on hand shapes and locations. The proposed approach utilizes an open-source framework called MediaPipe Holistic [48] which uses a hybrid architecture to construct pipelines for processing perceptual data, such as images and videos. To estimate the pose of the face, body, and hands regions for every frame of the input video, the MediaPipe Holistic incorporates three distinct models: MediaPipe holistic face landmarks, pose landmarks, and hand landmarks detector. In our study, we utilized the hands and body pose information extracted using these submodules. The output of MediaPipe hands and pose landmarks detector is shown in Fig. 4.

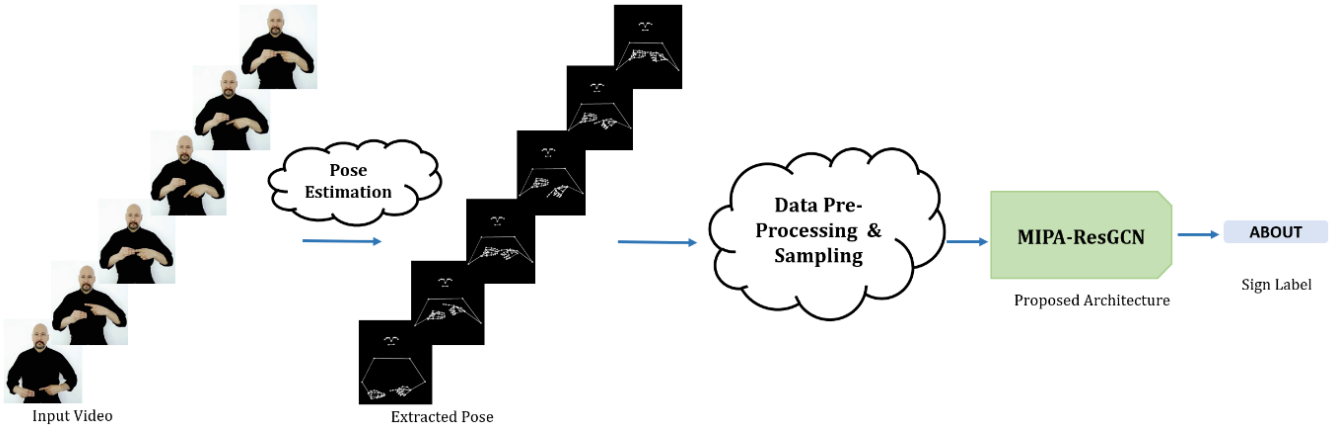


Fig. 3. A complete overview of proposed approach

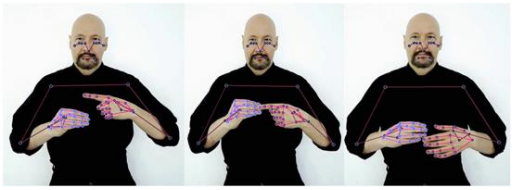


Fig. 4. Pose Estimated using MediaPipe Holistic Framework

Thus, for an input video $\boldsymbol{x_i}$ where $\{\boldsymbol{x_i} \in \boldsymbol{\mathcal{R}^{T\times H\times W\times C}}\}$; with T, H, W, & C represent number of frames, height, width and number of channels in each frame respectively, the extracted pose will have the dimensions of $\{\boldsymbol{x_i} \in \boldsymbol{\mathcal{R}^{T\times V\times F}}\}$: where T, V and F represent number of frames, number of joints per frame and number of features per joint respectively.

B. Stage 2: Data Preprocessing

Out of 75 landmarks generated using this model, we only used data for 65 landmarks. The set of 65 landmarks includes 23 landmarks each for left and right arms, torso, as well as significant face nodes including the lips, eyes, ears, and nose and 21 landmarks for each hand (4 landmarks for each finger and one for wrist). Since lower body joints do not play a significant role in sign language recognition, they were discarded. Although MediaPipe provides 3D landmarks for each joint, the depth coordinates represented by z dimension are not very accurate [48] and introduce noise. Hence, in our approach, we have exclusively utilized 2D coordinates (x & y) for each joint. The MediaPipe Holistic pose estimation technique offers landmark coordinates that are normalized to [0, 1] with respect to the image width and height. In order to maintain a consistent scale, we shift these coordinates by [-0.5, -0.5] and multiply them by 2. This process also ensures that the man is zero and standard deviation is unity. A Two noise transform is used to augment the data. It creates a copy of the input sequence. As the input videos lengths may vary and model requires a fixed length video as input, we have chosen a sequence length of 64. The 64 frames are obtained from the input video via a random start sampling strategy. Thus, the pose sequence $\boldsymbol{x_i}$ forwarded as input to the model has dimension of $\{\boldsymbol{x_i} \in \boldsymbol{\mathcal{R}^{64\times 65\times 2}}\}$.

B. Stage 3: Proposed Architecture

In this section, the details of our proposed architecture MIPA-ResGCN are illustrated. We build on our previous work SIGNGRAPH [49], in which we introduced the ResGCN [22] approach for skeleton-based sign language recognition. The implementation of SIGNGRAPH resulted in a significant performance improvement for pose-based SLR. In this work, We have extended SIGNGRAPH [49] by introducing a multi input architecture and an efficient part attention mechanism. In order to exhibit the efficacy of our model in learning the most distinctive features, we have also introduced the class activation map technique (CMAP) [50] to compute the activation of individual joints while performing a sign. In this section, the graph convolutions (GCN) being the fundamental component of our architecture will be firstly introduced followed by the details of MIPA-ResGCN architecture.

1. Graph Convolution

Firstly, we represent the human pose as a unidirectional graph $G = (V, \mathcal{E})$ where V = $\{v_1, v_2, \ldots, v_n\}$ is n number of nodes representing body and hand joints and $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$ is m number of edges representing bones connecting these joints. The relationships between nodes and edges are modeled by an adjacency matrix $\boldsymbol{A} \boldsymbol{\epsilon} \mathbb{R}^{\boldsymbol{n\times n}}$. An entry A_{ij} is equal to 1 if node *i* is connected to *j* otherwise it is zero. Each node in this graph has two channels representing x and y coordinates. In a pose sequence, the spatial graph convolution (SGC) for each frame can be defined as:

$$\boldsymbol{X_t^{(l+1)}} = \sum_{d=0}^{2} \boldsymbol{\mathcal{M}_d^{(l)}}\, \boldsymbol{X_t^{(l)}}\left(\boldsymbol{D_d}^{-1/2}(\boldsymbol{A_d}+\boldsymbol{I})\boldsymbol{D_d}^{-1/2} \otimes \boldsymbol{\theta_d}^{(l)}\right) \qquad (1)$$

Where 2 is the maximum graph distance, $\boldsymbol{X_t^l}$ and $\boldsymbol{X_t^{(l+1)}}$ are input and output features for frame t and layer *l*, $\otimes$ represents element wise multiplication, $\boldsymbol{A_d}$ is adjacency matrix of order d and I is the identity matrix to model self-loops. $\boldsymbol{D_d}$ is degree matrix used to normalize $\boldsymbol{A_d}$ and $\boldsymbol{\mathcal{M}_d^{(l)}}$ and $\boldsymbol{\theta_d}^{(l)}$ are learnable weights. An L×1 temporal convolutional layer (TCN) is used to collect the contextual cues embedded in adjacent frames for the purpose of extracting temporal characteristics. L is a hyper parameter representing temporal window size. These SGC and 2-D TCN are used to construct basic and bottleneck blocks for constructing a ResGCN architecture.

2. ResGCN Architecture

To address the problem of SLR, we have used ResGCN [22] architecture as baseline method. The spatiotemporal graph convolution (ST-GCN) serves as the foundation to construct basic and bottleneck blocks of this architecture.

Basic Block: The basic block consists of a spatial basic block connected in series with a temporal basic block. Spatial basic block is constructed using a GCN layer followed by a batch normalization (BN) and Relu activation layer whereas temporal basic block is composed of a 2-D TCN layer followed by a BN and Relu activation layer. A block residual is used to feed the input of each block at its output.

Bottleneck Block: Inspired by ResNet [51], the ResGCN architecture incorporates a bottleneck structure that enables a faster implementation of the model for both training and inference. The bottleneck block consists of a spatial bottleneck block connected in series with a temporal bottleneck block. By implementing the bottleneck with a reduction rate *R*, the number of feature channels is reduced. The spatial bottleneck block is constructed by adding two 1×1 convolution layers before and after a graph convolution layer. The first 1×1 convolution layer reduces the number of channels at its output by (input channels/R) and forwards the output to graph convolution layer and the second 1×1 convolution layer increases the channels by (GCN layer output*R). The temporal bottleneck follows the same structure with the difference that GCN layer is replaced with a temporal 2-D convolution layer. Each bottleneck block uses a block residual mechanism to feed the input of each block at its output. The sizes of spatial and temporal kernels are kept as 3 and 9 respectively and reduction rate R is set as 4.

Multi-Branch Input: Most multi-branch networks operate by independently feeding data from various modalities to the same model, and subsequently merging the outcomes of these streams to form the ultimate decision. Although this approach

is effective for data augmentation and enhances the model performance but might lead to high computational expenses and difficulties in hyper parameters tuning for extensive datasets. Thus, we have extracted features from both input branches using bottleneck blocks and concatenated them at an early stage of our model as presented in Fig.7. The concatenated features are then fed to one main branch to extract discriminative features. The pose-based action recognition frameworks [22, 52] divide the input features into three categories: joints, velocities, and bones. Velocity doesn't play a significant role in SLR because a sign can be performed faster or slower, but its context doesn't change. Considering the mentioned reason, we have divided the input features into two categories: 1) joint positions and 2) bone features (lengths & angles).

Suppose that the original 2D coordinate set of a sign sequence $\boldsymbol{X_t} \in \mathcal{R}^{C\times T\times V}$, where C, T, and V represent coordinates, frames, and nodes is extracted using pose extractor. The relative position '*r'* of each joint is calculated with respect to the center node 'c' of the pose as follows.

$$\boldsymbol{r} = \{\boldsymbol{v_{t,i}} - \boldsymbol{v_{t,c}} |\ \boldsymbol{i} = \boldsymbol{1, 2, \dots, V, t < T}\}$$
$$\boldsymbol{s.t.\, v_{t,c}} = \boldsymbol{mean}(\boldsymbol{v_{t,rightshoulder}, v_{t,leftshoulder}}) \qquad (\boldsymbol{2})$$

The original nodes and these relative positions are concatenated and sent as joint position input to the first branch. Next, bone features consisting of bone lengths and bone angles are computed. Bone length $\boldsymbol{l}$ is computed by subtracting each joint $\mathbf{v_{t,i}}$ from its adjacent joints $\mathbf{v_{t,adj}}$.

$$\boldsymbol{l} = \{\boldsymbol{v_{t,i}} - \boldsymbol{v_{t,adj}} |\ \boldsymbol{i} = \boldsymbol{1, 2, \dots, V, t < T}\} \qquad (\boldsymbol{3})$$

Finally, Bone angle is computed as:

$$\boldsymbol{\alpha} = \{\boldsymbol{arccos}(\frac{\boldsymbol{v_{t,i}} - \boldsymbol{v_{t,adj}}}{\sqrt{\sum \boldsymbol{v_{t,i}^2}}}) |\ \boldsymbol{i} = \boldsymbol{1, 2, \dots, V, t < T}\} \qquad (\boldsymbol{4})$$

These bone features are concatenated together and sent as input to the second branch.

Part Attention Module: Inspired by split attention model in [53], a part based attention module has been designed to capture the significance of each body part throughout the entire sign sequence. We have manually divided input features into five individual body parts P=5: face, left arm, right arm, left hand, and right hand based upon each part's corresponding joints as shown in Fig. 5.

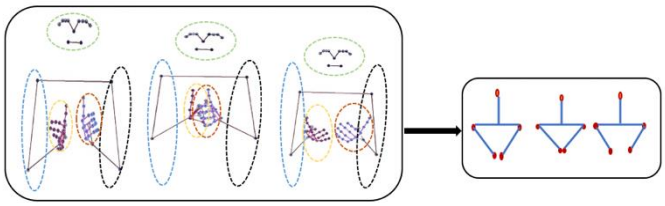


Fig. 5. An illustration depicting manually designed body parts.

To compute part attentions, the first step involves applying average pooling in the temporal dimension to the entire skeleton. The resulting feature maps are passed through a 2D convolution layer, followed by a BN and ReLU Layer. Afterward, attention matrices are calculated using five 2D convolution layers (one corresponding to each part) , and a part-level SoftMax is used to identify the most essential part. A final skeleton representation is formed by concatenating features of five parts with different attention weights. The overall framework of part attention module is presented in Fig. 6. This attention block shown in Fig. 6 can be mathematically formulated as explained in Eq. (5a & b):

$$\boldsymbol{x_p} = \boldsymbol{\gamma}\big(\boldsymbol{\delta}(\boldsymbol{p}(\boldsymbol{x_{in}})\boldsymbol{\theta})\boldsymbol{\theta_p}\big) \qquad (\boldsymbol{5.a})$$

$$\boldsymbol{x_{out}} = \boldsymbol{x_{in}} \otimes (\boldsymbol{Concat}(\{\boldsymbol{x_p}|\boldsymbol{p} = \boldsymbol{1, 2, \dots, P}\})) \qquad (\boldsymbol{5.b})$$

Where $\boldsymbol{x_{in}}$ and $\boldsymbol{x_{out}}$ represent input and output feature maps, $\boldsymbol{\gamma}(.)$, $\boldsymbol{\delta}(.)$ $\boldsymbol{and}$ $\boldsymbol{p}(.)$ denote Softmax, ReLU and temporal average pooling operations respectively. Where $\boldsymbol{\theta \epsilon \mathcal{R}^{C\times\frac{C}{R}}}$ and $\boldsymbol{\theta_p \epsilon \mathcal{R}^{\frac{C}{R}\times C}}$ are learnable parameters and C represents total channels in the layer and R is the reduction rate which is set as 4.

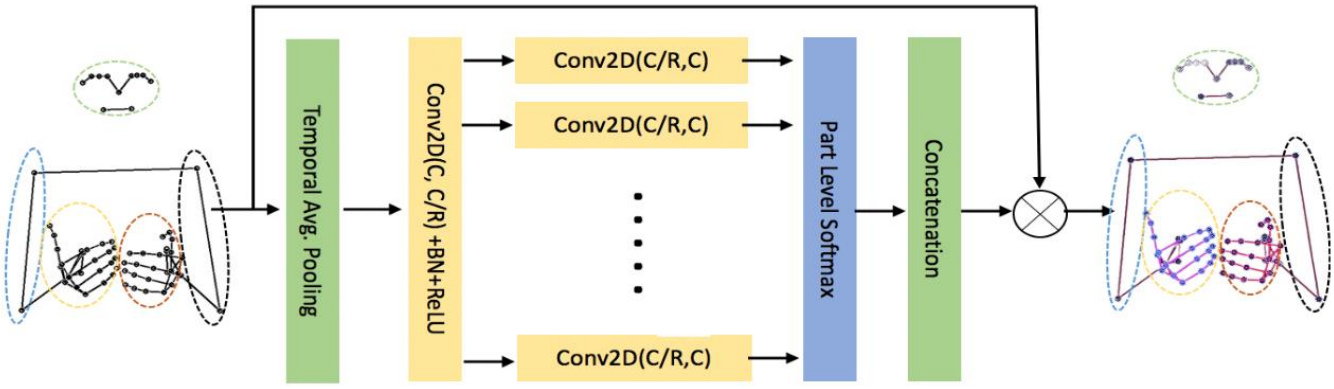


Fig. 6. The structure of proposed part attention block, where R(residual)=4 and C (channels in each layer)

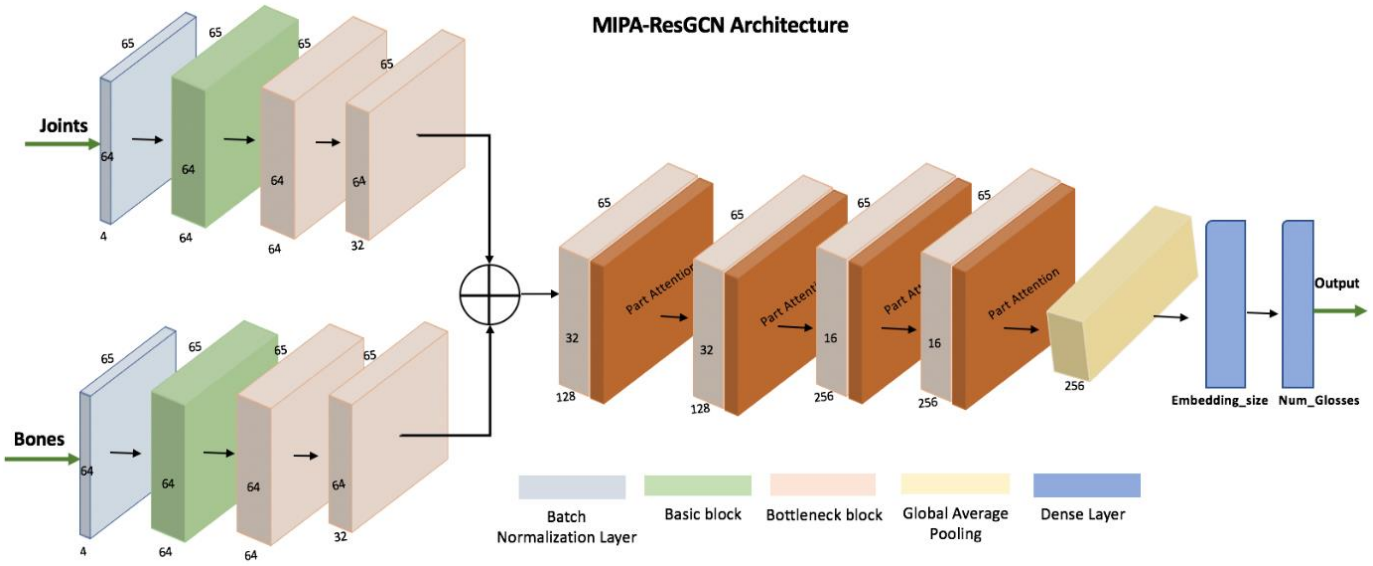


Fig. 7. Overview of Multi Input Part Attention ResGCN (MIPA-ResGCN) architecture

The complete overview of the proposed architecture MIPA-ResGCN is presented in Fig. 7.

IV. Results and Discussions

In this section, we evaluate our proposed architecture on five publicly available SL datasets: WLASL-100, WLASL-300, and WLASL-1000 [18], LSA-64 [42], and MINDS-Libras [44], using four type of evaluation matrices accuracy, precision, recall, and F1 score. These matrices are computed using four values: True Positive Predictions (TPP), True Negative Predictions (TNP), False Positive Predictions (FPP), and False Negative Predictions (FNP). The following equations are used to compute accuracy, precision, recall, and F1 score.

$$\boldsymbol{Accuarcy} = \frac{\boldsymbol{TPP+TNP}}{\boldsymbol{TPP+TNP+FPP+FNP}} \quad \textbf{(6)}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TPP}}{\boldsymbol{TPP+FPP}} \quad \textbf{(7)}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TPP}}{\boldsymbol{TPP+FNP}} \quad \textbf{(8)}$$

$$\boldsymbol{F1_{Score}} = \boldsymbol{2} * \frac{\boldsymbol{Precision * Recall}}{\boldsymbol{Precision+Recall}} \quad \textbf{(9)}$$

We compare our results with SOTA approaches based on appearance and pose features. Additionally, we conduct ablation studies to elaborate the contribution of each component in the proposed architecture to the overall performance.

A. Experimental Setup

For our Experiments, we have represented the human pose sequence as a graph, based on a spatial configuration described in [54]. Our experiments are conducted using the PyTorch framework on a single NVIDIA 3080 RTX GPU. The model is trained for 350 epochs for WLASL-100, LSA-64, & MINDS-Libras and for 600 epochs for WLASL-300 and WLASL-1000. The embedding size is set at 128 for WLASL-100, LSA-64, and MINDS-Libras and set at 300 & 512 for WLASL-300 & WLASL-1000 respectively. Further details are provided below:

- Optimizer: Adam optimizer [55]
- Batch size: 32.
- Temporal kernel size L: 9
- Maximum graph distance: 2
- Spatial kernel size: 3
- Learning Rate: Cyclic learning rate scheduler with a learning rate set to 0.01.
- For model optimization, a cross-entropy loss measured by Eq. (10) is used as an objective function.

$$\boldsymbol{Cross\ Entropy\ Loss} = -\sum_{i=1}^{C} \boldsymbol{y_i log \hat{y_i}} \quad \textbf{(10)}$$

where $\boldsymbol{\hat{y_i}}$ is the Softmax probability for the i[th] class and C represents the total number of classes.

B. Results on WLASL Dataset

Dataset Description: WLASL (word level American sign language) [18] is a recent and comprehensive dataset compiled from various online open sources, featuring a diverse range of signers, lighting conditions, and background variations. The dataset is divided into four subsets of data: WLASL-100, WLASL-300, WLASL-1000, & WLASL-2000. The number in each subset's name corresponds to the number of sign glosses it contains. Further details are provided in Table II. Our study employs the same training, validation, and testing protocols as those specified by the authors of the dataset [18].

TABLE II
Details of WLASL Subsets

Dataset -Subset	Gloss count	Video count
WLASL-100	100	2,038
WLASL-300	300	5117
WLASL-1000	1000	13168
WLASL-2000	2000	21,083

TABLE III
PERFORMANCE COMPARISON OF PROPOSED ARCHITECTURE WITH SOTA METHODS ON WLASL-100, WLASL-300, WLASL-1000

Data Type	Model	WLASL-100			WLASL-300			WLASL-1000		
		top 1	top 5	top 10	top 1	top 5	top 10	top 1	top 5	top10
Appearance	I3D [18]	65.89	84.11	89.92	56.14	79.94	86.98	47.33	76.44	84.33
	TK-3D Convnet [56]	77.55	91.42	-	68.75	89..41	-	-	-	-
	Fusion 3 [57]	75.67	86.00	90.16	68.30	83.19	86.22	56.68	79.85	84.71
Pose	Pose-GRU [18]	46.51	76.74	85.66	33.68	64.37	76.05	30.01	58.42	70.15
	Pose-TGCN [18]	55.43	78.68	87.60	38.32	67.51	79.64	34.86	61.73	71.91
	GCN-BERT [19]	60.15	83.98	88.67	42.18	71.71	80.93	-	-	-
	MOPGRU [21]	63.18	-	-	-	-	-	-	-	-
	SPOTER [20]	63.18	-	-	43.78	-	-	-	-	-
	SIGNGRAPH [49]	72.09	88.76	92.64	71.40	92.26	94.16	61.83	85.87	91.04
	MIPA-ResGCN (Ours)	**83.33**	**92.64**	**95.35**	**72.90**	**88.92**	**93.41**	**64.92**	**88.37**	**92.16**

Comparison with SOTA methods: This paper presents the top-1, top-5, and top-10 accuracy achieved by our architecture on the WLASL-100, WLASL-300, and WLASL-1000 datasets. A comparative analysis between the proposed method and SOTA appearance-based and pose-based approaches is provided in Table III.

Vs. Pose based: From Table III, the MIPA-ResGCN obtains an excellent performance of, 83.33%, 72.90%, and 64.92% thereby outperforming SOTA pose based method by 10.43%, 1.5%, and 3.09% for WLASL-100, WLASL-300, and WLASL-1000 respectively.

Vs. Appearance based: The proposed method exhibits superior performance compared to the SOTA appearance-based method by 5.78%, 4.15%, and 8.24% for WLASL-100, WLASL-300, and WLASL-1000 respectively. Confusion matrix for WLASL-100 dataset, is shown in Figure 8. Confusion matrix represents a complete picture of sign recognition accuracy.

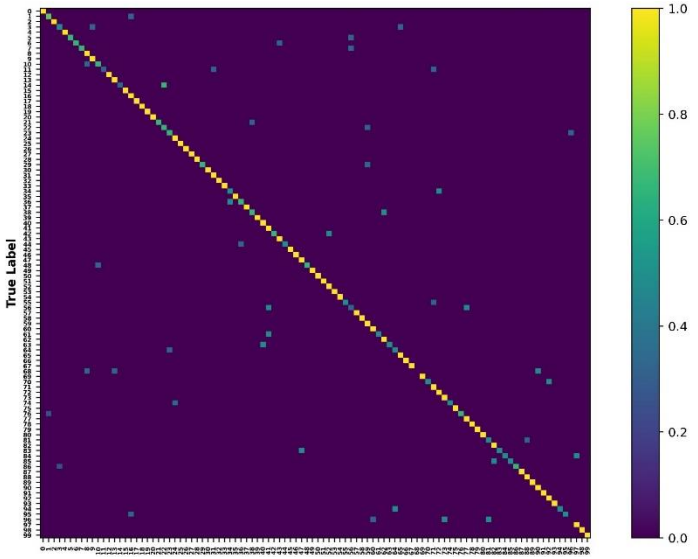


Fig. 8. Confusion Matrix on the WLASL-100 dataset

Failure Cases: Our model shows excellent recognition performance across approximately the entire WLASL-100 dataset, however, there are three signs: 'same', 'pizza', and 'school' which our model finds difficult to recognize. All instants of sign "School" are recognized as "Paper" by the model. Upon investigation, it was observed that both the signs are performed in the same way as shown in Fig. 9, leading to difficulties in accurate recognition by the model. Moreover, a significant variability in the execution of the signs "1. Same" and "2. Pizza" by various signers was observed upon investigation also shown in Fig. 10. These ambiguities in the dataset: 1. different words signed in the same way, 2. same word signed differently by signers, lead the model to inaccurate gloss predictions.

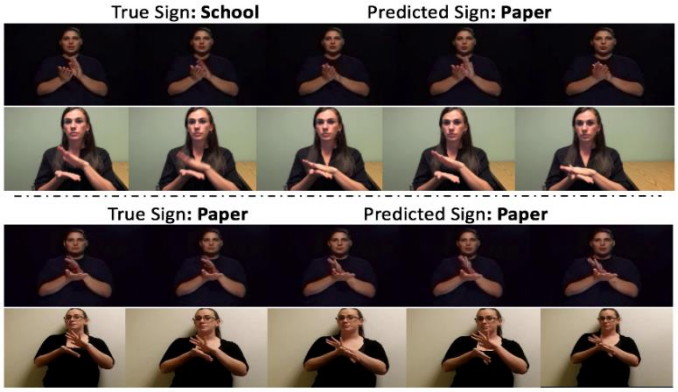


Fig. 9. Two different words (1. School, 2. Paper) performed in the same way by signers.

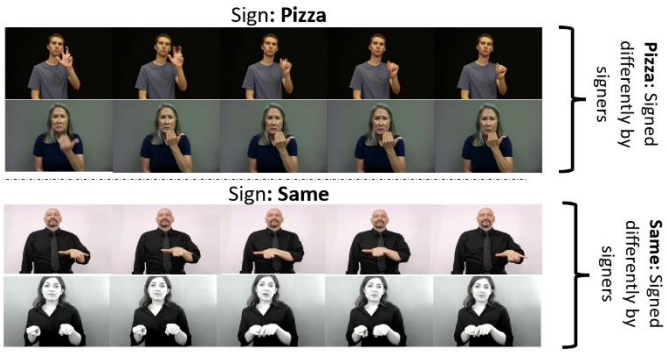


Fig. 10. Words (1. Pizza, 2. Same) Various instances of each word performed in entirely different ways by signers.

C. Results on LSA-64 Dataset

Dataset Description: LSA-64 dataset targets vocabulary of 64 different glosses from Argentinian sign language. It comprises 50 video samples per class, with each class signed by 10 non-expert signers. The 64 glosses include both verbs and nouns used in Argentinian sign language. The following setup is used for model training and evaluation. The data is divided into 80:20 for training and test purposes as per the protocol used by dataset authors. To determine the optimal model parameters, a k-fold cross-validation is employed on the training set with k=4. The results are evaluated over an average of five repetitions.

Comparison with SOTA methods: We present the top-1 accuracy of our architecture on LSA-64 dataset. We provide a comparison of our method to the current SOTA Appearance based, pose based, and hybrid approaches in Table IV. Our model can achieve SOTA 100±**0**% accuracy on LSA-64 dataset. The results represented with * are obtained as an average of 5 iterations.

TABLE IV
ACCURACY COMPARISON WITH SOTA METHODS ON LSA-64 DATASET

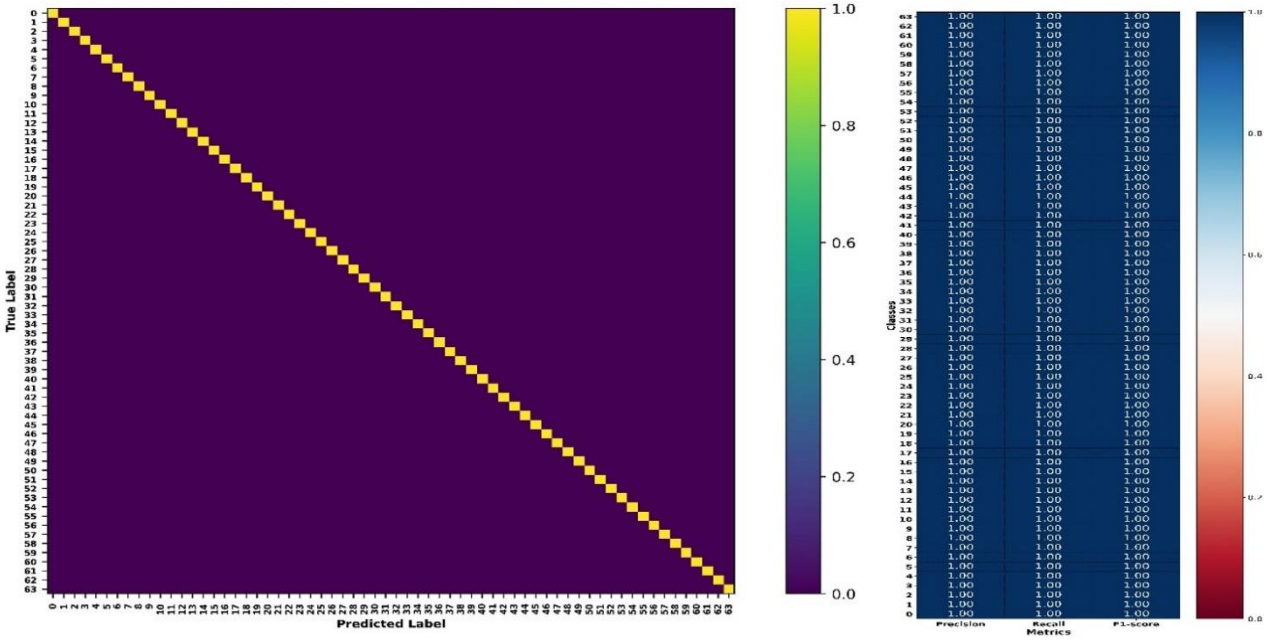
Data Type	Model	Top-1 Accuracy
Appearance Based	LSTM+LDS [58]	98.09 ± 0.59 *
	DeepSign CNN [59]	96.00
	MEMP [60]	99.06
	I3D	98.91
Appearance + Pose	LSTM+DSC [61]	99.84± 0.19*
	ELM+MN CNN [62]	97.81
Pose Only	SPOTER [20]	100.00± 0 *
	MOPGRU [21]	99.92
	MIPA-ResGCN (ours)	**100.00± 0***

Confusion matrix and metrices representing accuracies, recall and precision of each class are shown in Fig. 11, clearly demonstrating an excellent performance of our proposed architecture towards sign recognition.

D. Results on MINDS-Libras Dataset

Dataset Description: MINDS-Libras [44] dataset targets a vocabulary of 20 different glosses from Brazilian sign language. The dataset comprises of 60 video recordings per category, each performed by 12 distinct signers, and recorded in a controlled environment with a static green background using a Canon EOS Rebel t5i DSLR camera and a Microsoft Kinect v2 to capture RGB and RGB-D sequences. For our work we use only RGB sequences to extract pose information. The following setup as proposed by [63] is used for model training and evaluation. The data is divided into 75:25 for training and test purposes as per the protocol used by dataset authors. A k-fold cross validation is employed on the training set with k=3 to find the best model parameters. The results are presented as an average of ten repetitions.

Comparison with SOTA methods: We present the top-1 accuracy of our architecture on MINDS-Libras dataset. We provide a comparison of our method to the current SOTA Appearance based approaches in Table V. Our model is able to achieve SOTA accuracies of 96.70±1.07% on MINDS-Libras dataset. Fig. 12 illustrates the precision, recall, F1 score and confusion matrix for each sign in the dataset. Our proposed methodology has a minimum of 80% on all used matrices of Precision, Recall, and F1-score which can be considered as a very good model performance. The evidence supporting this claim can be seen in the confusion matrix, which demonstrates a high accuracy.



(a) (b)

Fig. 11. a). Confusion matrix b). Precision, recall, and F1-score for each sign class on LSA-64 dataset

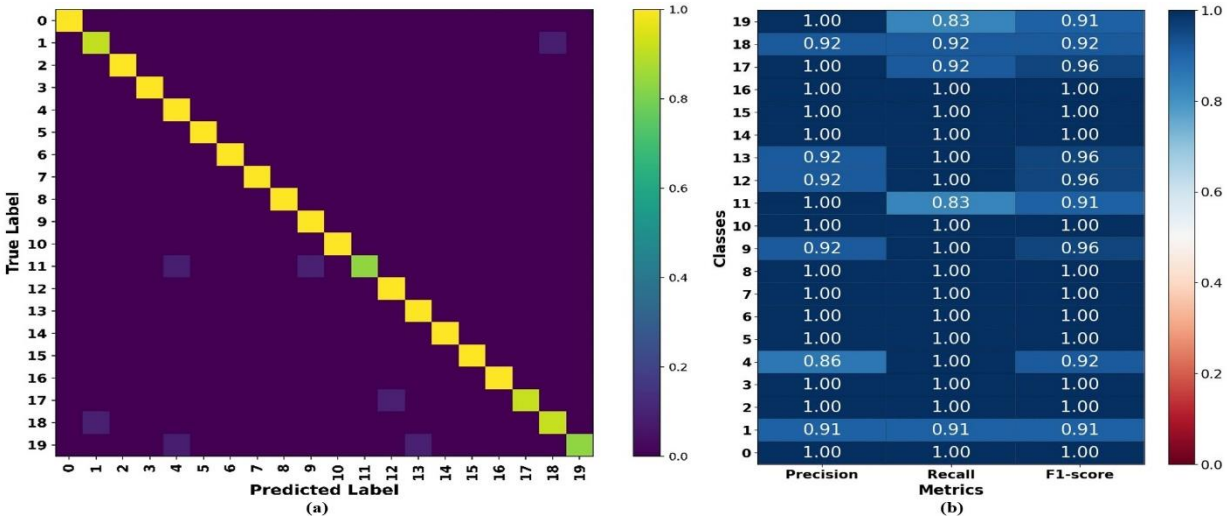


Fig. 12. a). Confusion matrix b). Precision, recall, and F1-score for each sign class on MINDS-Libras dataset

TABLE V
ACCURACY COMPARISON WITH SOTA METHODS ON MINDS-LIBRAS DATASET

Type	Model	Top-1 Accuracy
Appearance Based	CNN3D [58]	72.6
	CNN 3D [59]	93.3 ± 1.69
	GEI + SVD+SVM [63]	84.66 ± 1.78
Pose Based	**MIPA-ResGCN (Ours)**	**96.70± 1.07%**

E. Computational and Generalized Performance Analysis

In this study, we conducted a comparison of our proposed architecture (MIPA-ResGCN) with I3D (an appearance-based method), SPOTER (a pose-based method)**,** and SIGNGRAPH (a pose-based method) to evaluate the computational efficiency and model's generalization performance. To begin, we compared the number of model parameters, finding that MIPA-ResGCN has 0.99 million while I3D has 12.4 million and SPOTER has 5.92 million model parameters. Next, we evaluated the computational complexity of each model by quantifying the number of floating-point operations and average time taken to process each video during inference stage, utilizing the FLOP profiler feature of the DeepSpeed library [64]. The evaluations were performed on a single NVIDIA RTX-3080 GPU. The results shown in Fig. 13 demonstrate that our model has a much smaller number of parameters and inference time as compared to I3D and SPOTER and comparable compute performance to SIGNGRAPH with an accuracy increase by large margin. The number of GFLOPs required are much smaller as compared to I3D and comparable to SPOTER and SIGNGRAPH.

To assess the generalizability and robustness of our proposed model, we conducted a series of experiments wherein we trained the model using smaller subsets of the training data and evaluated its performance on a fixed test set. These experiments were conducted on the LSA-64 data, which was partitioned into training and test sets at an 80:20 ratio. MIPA-ResGCN, I3D, SIGNGRAPH and SPOTER models were trained with different splits of training data. Training was conducted using subsets of the training data ranging from 10% to the complete dataset, with 20% more data being added at each stage. To ensure uniform class distributions, training subsets were selected using a uniform sampling method. The resulting models were then evaluated on the fixed test set. The results, shown in Fig. 14, indicate that MIPA-ResGCN achieved an accuracy of 92% on test set when trained on just 10% split of the training data, while SPOTER, SIGNGRAPH, and I3D models achieved an accuracy of 88.68%, 75% and

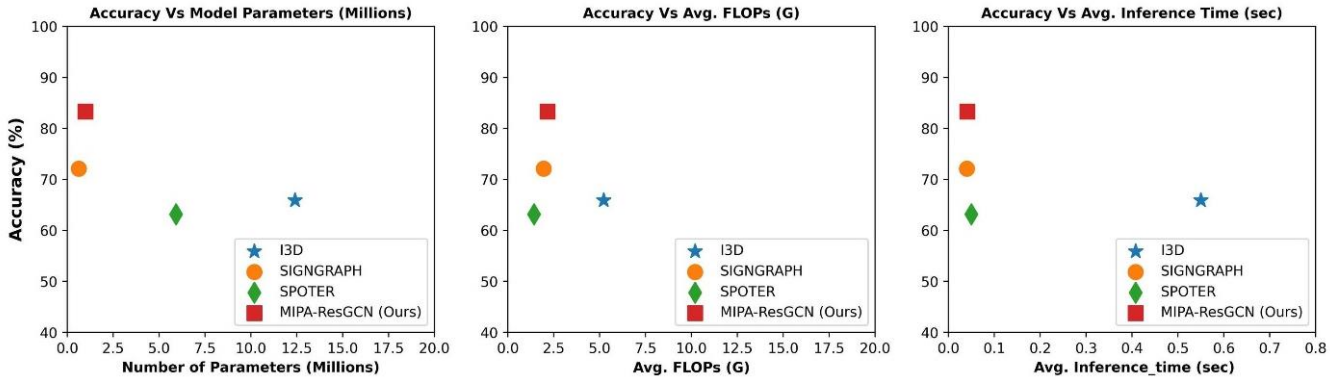


Fig. 13. Comparison of MIPA-ResGCN (Proposed model), I3D, SIGNGRAPH and SPOTER in terms of number of parameters (millions), average FLOPS (G), and average inference time (sec).

45.47% respectively. The performance of MIPA-ResGCN continued to improve as the size of the training set increased, ultimately reaching an accuracy of 100% when trained on 50% of the training data. SIGNGRAPH achieved the accuracy of 100 % at 70% split of training data, SPOTER achieved the accuracy of 100 % at 90% split of training data and I3D achieved the highest accuracy of 98.91% when trained on 100% data. The results of the conducted experiments demonstrate that MIPA-ResGCN performs significantly better than SOTA SLR models even when trained with smaller data sizes.

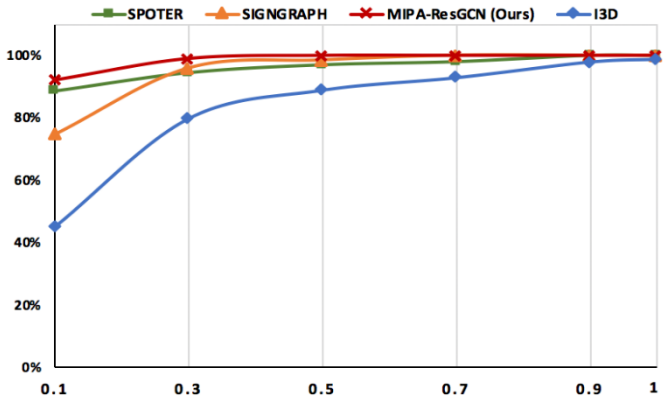


Fig. 14. Top-1 accuracies of the MIPA-ResGCN, I3D, SIGNGRAPH, and SPOTER models trained on six subsets of the training data and evaluated on a fixed 20% test split.

F. Ablation Studies

In this section, we conduct ablation studies to examine how different components, which were introduced in the baseline ResGCN model, contribute to its overall performance. The proposed model consists of ResGCN with a reduction rate (***R***) of 4 as a baseline model. The ResGCN model includes one basic and six bottleneck blocks followed by an average pooling and two fully connected layers. We have introduced two input branches consisting of joints and bones information as explained in section. III. Various attention mechanisms including part attentions, joint, and frame attentions have been tested. The results presented in Table VI clearly demonstrate that our model greatly benefits from multi-input structure. It improves the model's performance significantly for all datasets. Inclusion of part attention mechanism also enhances the model performance by a huge margin. Overall best recognition accuracies are achieved by including both part attention and multi-inputs for WLASL-100, WLASL-300, and WLASL-1000 dataset.

G. Visualizations and Explanations

To showcase the efficacy of our model in learning the most distinctive features, we applied the class activation map technique [50] to compute the activation of individual joints within a video frame. These activation maps presented in Fig. 15 depict the activated joints in various frames of a sequence. Joints with the highest activation are represented using brighter colors and on a larger scale in the visualization. It is evident from the results that out of the five body parts, our skeleton was divided in for part attention mechanism, the model pays higher attention to the left and right hands, as hands locations and movements are indeed the most distinguished features of sign language. Moreover, the model is also able to correctly capture the significance of each joint. The sign for "CHAIR" is performed by moving the index and middle fingers of right hand up and then bringing them down and touching the same two fingers of left hand. As can be seen, in the frames for sign class "CHAIR", our model gives the most attention to the joints of these fingers. The observations align perfectly with the understanding that sign language relies heavily on the shapes, locations, and orientations of the hands.

V. Conclusion

In this study, we propose an accurate and efficient method for pose-based isolated sign language recognition. The architecture uses an efficient pose extractor to extract pose information which is then divided into two branches: joints and bones, to construct the multi-input structure. This multi-input is forwarded to ResGCN consisting of basic and bottleneck blocks and a proposed part attention mechanism to

TABLE VI
Ablation Studies of model components in accuracy (%) for WLASL-100, WLASL-300, & WLASL-1000 datasets
MI: multi input, PA: Part attention, FA: Frame attention, JA: Joint attention

Model	WLASL-100	WLASL-300	WLASL-1000
Baseline (ResGCN): using Bones only	48.06	47.75	46.59
Baseline: using Joints only	72.09	71.40	61.83
Baseline: using Bones only + PA	70.16	56.14	52.40
Baseline: using Joints only + PA	75.19	71.71	63.54
Baseline + MI (Using both Joints & Bones)	77.52	72.01	62.30
Baseline + MI + FA	81.01	71.86	62.42
Baseline + MI + JA	79.07	**74.10**	64.02
MIPA-ResGCN (Baseline + MI + PA)	**83.33**	72.90	**64.92**

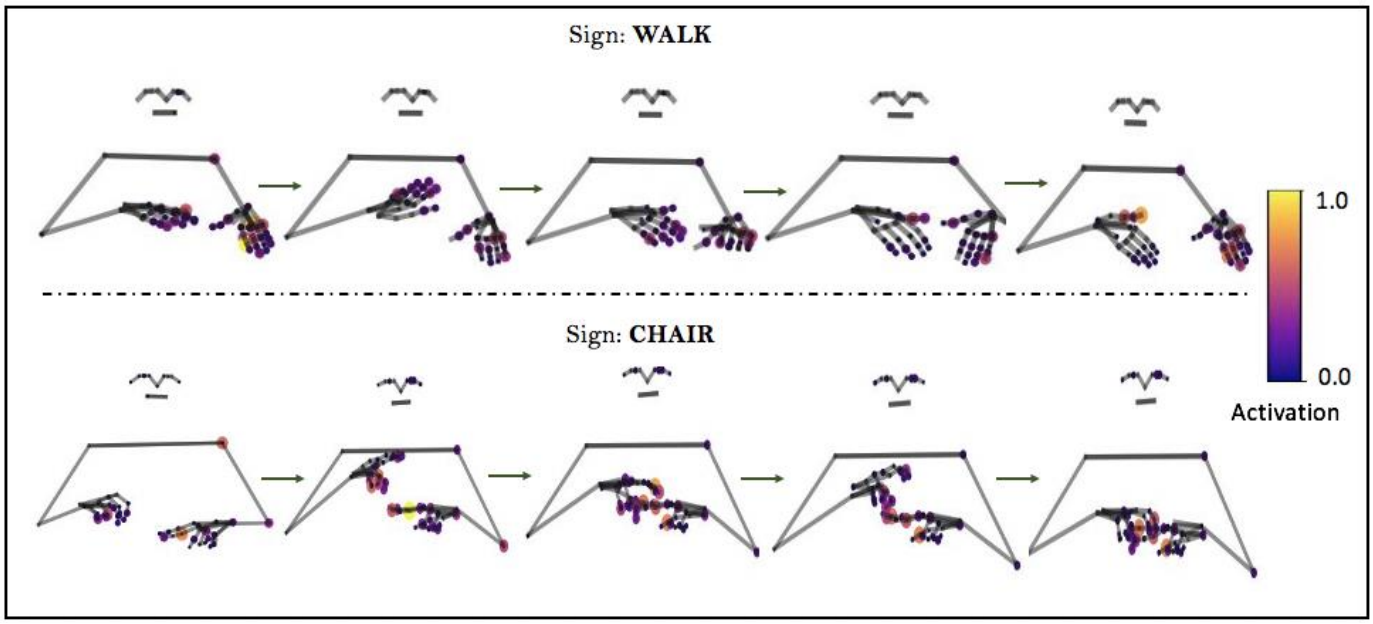


Fig. 15. Activated joints for the examples of 'WALK' and 'CHAIR' signs. A bigger scale and brighter color represent the higher activated joint.

force the model to learn the efficient spatiotemporal features by focusing on the most essential body parts and ignoring nodes with unnecessary information. Our results clearly demonstrate the model's ability to learn strong temporal dependencies thereby providing SOTA accuracies on the challenging datasets of WLASL, MINDS-Libras, and LSA-64. Our model provides significant reduction in computational complexity and provides more generalizable results. Additionally, our visualizations of activated joints effectively illustrate that the model places a strong emphasis on the most crucial body parts in sign language: hand shapes, locations, and orientations, thereby supporting the assertions made in section-I. The proposed architecture will have a large influence on applications requiring gesture recognition. Our future work involves expanding the proposed architecture to incorporate appearance-based hand features, with the aim of enhancing recognition accuracy in critical scenarios where the same signs may be signed differently.

ACKNOWLEDGMENT

The authors wish to express their gratitude towards DeafTawk (https://www.deaftawk.com/) for their valuable assistance, guidance, and support in terms of domain knowledge, which proved instrumental in the successful completion of this research project.

REFERENCES

[1] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 131-153, 2019.

[2] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13009-13016.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.

[4] A. A. Q. Mohammed, J. Lv, and M. S. Islam, "A deep learning-based end-to-end composite system for hand detection and gesture recognition," *Sensors*, vol. 19, p. 5282, 2019.

[5] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools and Applications*, vol. 79, pp. 22965-22987, 2020.

[6] 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.

[7] C. Mao, S. Huang, X. Li, and Z. Ye, "Chinese sign language recognition with sequence to sequence learning," in *CCF Chinese Conference on Computer Vision*, 2017, pp. 180-191.

[8] K. Bantupalli and Y. Xie, "American sign language recognition using deep learning and computer vision," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4896-4899.

[9] 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.

[10] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546-6555.

[11] 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.

[12] Y. Li, X. Wang, W. Liu, and B. Feng, "Pose anchor: a single-stage hand keypoint detection network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 2104-2113, 2019.

[13] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 2822-2832, 2018.

[14] M. De Coster, M. Van Herreweghe, and J. Dambre, "Sign language recognition with transformer networks," in *12th international conference on language resources and evaluation*, 2020, pp. 6018-6024.

[15] 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.

[16] 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.

[17] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," in *European Conference on Computer Vision*, 2020, pp. 301-319.

[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] M. Barreto, "Criação de uma base de dados para o alfabeto datilológico utilizando dispositivo de interação não-convencional," Universidade Tecnológica Federal do Paraná, 2017.
[24] J. Wu, L. Sun, and R. Jafari, "A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors," *IEEE journal of biomedical and health informatics*, vol. 20, pp. 1281-1290, 2016.
[25] A. Calado, V. Errico, and G. Saggio, "Toward the minimum number of wearables to recognize signer-independent Italian sign language with machine-learning algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-9, 2021.
[26] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using rf sensing," *IEEE Sensors Journal*, vol. 21, pp. 3763-3775, 2020.
[27] D. S. Breland, S. B. Skriubakken, A. Dayal, A. Jha, P. K. Yalavarthy, and L. R. Cenkeramaddi, "Deep learning-based sign language digits recognition from thermal images with edge computing system," *IEEE Sensors Journal*, vol. 21, pp. 10445-10453, 2021.
[28] P. C. Badhe and V. Kulkarni, "Indian sign language translator using gesture recognition algorithm," in *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, 2015, pp. 195-200.
[29] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2961-2968.
[30] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research*, vol. 13, pp. 2205-2231, 2012.
[31] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293-1301.
[32] H.-D. Yang, "Sign language recognition with the kinect sensor based on conditional random fields," *Sensors*, vol. 15, pp. 135-147, 2014.
[33] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *2016 IEEE international conference on multimedia and expo (ICME)*, 2016, pp. 1-6.
[34] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, Z. Ma, and J. Song, "Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model," *Pattern recognition letters*, vol. 119, pp. 187-194, 2019.
[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.
[36] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *arXiv preprint arXiv:1812.01053*, 2018.
[37] Z. Liu, L. Pang, and X. Qi, "MEN: Mutual Enhancement Networks for Sign Language Recognition and Education," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[38] Y. C. Bilge, R. G. Cinbis, and N. Ikizler-Cinbis, "Towards zero-shot sign language recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2022.
[39] O. M. Sincan and H. Y. Keles, "Using Motion History Images with 3D Convolutional Networks in Isolated Sign Language Recognition," *IEEE Access*, vol. 10, pp. 18608-18618, 2022.
[40] A. Kuznetsova, L. Leal-Taixé, and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 83-90.
[41] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2200-2207.
[42] F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. Rosete, "LSA64: an Argentinian sign language dataset," in *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*. 2016.
[43] O. M. Sincan and H. Y. Keles, "Autsl: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340-181355, 2020.
[44] 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.
[45] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10863-10872.
[46] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3383-3393.
[47] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, and X. Wang, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, pp. 3349-3364, 2020.
[48] Google. (2022, Deecember 31). *MediaPipe Holistic*. Available: https://google.github.io/mediapipe/solutions/holistic
[49] 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.
[50] 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.
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[52] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1112-1121.
[53] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, and R. Manmatha, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2736-2746.
[54] 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.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[56] 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.
[57] 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.
[58] 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.
[59] J. A. Shah, "Deepsign: A deep-learning architecture for sign language," 2018.
[60] X. Zhang and X. Li, "Dynamic gesture recognition based on MEMP network," *Future Internet*, vol. 11, p. 91, 2019.
[61] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "A deep learning approach for analyzing video and skeletal features in sign language recognition," in *2018 IEEE international conference on imaging systems and techniques (IST)*, 2018, pp. 1-6.

[62] J. Imran and B. Raman, "Deep motion templates and extreme learning machine for sign language recognition," *The Visual Computer*, vol. 36, pp. 1233-1246, 2020.
[63] 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.
[64] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3505-3506.

Neelma Naz received the M.S. degree in Electrical engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2014, where she is currently pursuing the Ph.D. degree in Robotics and Intelligent Machine Engineering with the School of Mechanical and Manufacturing Engineering.
Her current research interests include computer vision, pattern recognition, machine learning and control systems.

Hasan Sajid received the B.S. degree in mechatronics engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from University of Kentucky, USA, in 2014 and 2016, respectively. He is currently an Associate Professor at Department of Robotics & AI, NUST and Scientific Director at National Center for Artificial Intelligence. He has expertise in the areas of computer vision, machine learning and deep learning. His research interests include speech and text recognition, video analytics and application of AI in healthcare, crowd and traffic domains. He has 25+ high impact peer reviewed publications and won fundings of more than 100 M. He was a recipient of the U.S. State Department Fulbright Scholarship.

Sara Ali received her PhD degree in Robotics and Intelligent Machine Engineering from School of Mechanical and Manufacturing Engineering, National University of Sciences and Technology (NUST), Pakistan in 2020. She did her master's in research from Middlesex University, London, UK in 2013 on faculty development program (FDP). Currently she is an Assistant Professor in Robotics and Intelligent Machine Engineering Department, NUST, Pakistan. Her research interests include Human-Robot Interaction, Sensor Systems, Interactive Robotics, Virtual Reality, and Human-Machine Interaction. She is the Principal Investigator (PI) of 2 main labs i.e., Intelligent Field Robotics Lab (IFRL) and Human-Robot Interaction (HRI) lab at National University of Sciences and Technology. She has published over 40 research articles in international peer-reviewed journals and conferences including citations from Nature. She has also been appointed as session chair and member of the Scientific Advisory Board (SAB) at several prestigious international conferences. She has also authored a book titled "Introducing Therapeutic Robotics for Autism" in the UK published by Emerald Publishing Ltd. She has also been awarded a national patent on Brain Imaging Tool for Medical Diagnosis.

Osman Hasan received his BEng (Hons) degree from the University of Engineering and Technology, Peshawar Pakistan in 1997, and the MEng and PhD degrees from Concordia University, Montreal, Quebec, Canada in 2001 and 2008, respectively. Before his PhD, he worked as an ASIC Design Engineer from 2001 to 2004 at LSI Logic. He worked as a postdoctoral fellow at the Hardware Verification Group (HVG) of Concordia University for one year until August 2009. Currently, he is Pro-Rector (Academics) at National University of Science and Technology (NUST), Islamabad, Pakistan. He is the founder and director of System Analysis and Verification (SAVe) Lab at NUST, which mainly focuses on the design and formal verification of energy, embedded and e-health related systems. He has received several awards and distinctions, including the Pakistan's Higher Education Commission's Best University Teacher (2010) and Best Young Researcher Award (2011) and the President's gold medal for the best teacher of the University from NUST in 2015. Dr. Hasan is a senior member of IEEE, member of the ACM, Association for Automated Reasoning (AAR) and the Pakistan Engineering Council.

Muhammad Khurram Ehsan received his Ph.D. degree in engineering with specialization in statistical signal processing and the M.S. degree in electrical communication engineering from the University of Kassel, Germany, in 2016 and 2010, respectively. He has been an Associate Professor, Faculty of Engineering, Bahria University, Pakistan, since July 2022. He has been a Visiting Lecturer, Faculty of Electrical Engineering and Computer Science, University of Kassel, Germany, since July 2017. His research interests include statistical modeling, data analysis, and cognitive radio systems.