National College of Ireland

A Lightweight SLR System Based on MobileNetV3, BiLSTM, and Attention Mechanism

MSc Research Project
Master of Science in Artificial Intelligence

HAIYAN HU
Student ID: x22247327

School of Computing
National College of Ireland

Supervisor: Kislay Raj

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……HAIYAN HU……

Student ID: ……x22247327……

Programme: Master of Science in Artificial Intelligence **Year:** ……2025……..

Module: ……MSc Research Project……

Supervisor: ……Kislay Raj……

Submission Due Date: ……15/09/2025……

Project Title: A Lightweight SLR System Based on MobileNetV3, BiLSTM, and Attention Mechanism

Word Count: ……6266…… **Page Count**……19……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……HAIYAN HU……

Date: ……11/09/2025……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	HAIYAN HU
Date:	11/09/2025
Penalty Applied (if applicable):	

A Lightweight SLR System Based on MobileNetV3, BiLSTM, and Attention Mechanism

Student: HAIYAN HU
StudentID: x22247327

Abstract

This project seeks to develop a system that recognizes American Sign Language (ASL) word level videos. I begin by using MobileNetV3-large to extract image features before using a BiLSTM(Bidirectional Long Short Term Memory Network) to model temporal information. Furthermore, I also add an attention mechanism to bring better attention to the key frames in the video to improve the recognition. Methods such as image augmentation and label smoothing were used when testing the model to make it more generalizable and stable. In general, the sum of the project was developed and validated based on the WLASL-300 dataset. Ultimately, I got good recognition results while keeping the model structure relatively simple which aligned with the project's main focus: to adapt an efficient model structure and optimize the training strategy to build a base for further applications in the future.

Keywords: Sign Language Recognition; MobileNetV3; BiLSTM; Attention Mechanism; DataAugmentation;

I. INTRODUCTION

Sign language is a form of visual as well as spatial representations of language with its unique grammar and syntax through gestures and shapes (usually by hand), body language and facial expressions used by deaf and hearing impaired people to communicate in a linguistic way. As digitization and visual computing continues to advance and accelerate computer based systems of sign language are becoming increasingly important in facilitating modes of communication between hearing impaired people and the broader community. This is important because it enables hearing impaired people to participate in society as active citizens in an inclusive way. Computer science and developments in information technology means thatsign language is becoming an important focus of research at the intersection of artificial intelligence and accessibility technologies. Through sign language recognition systems, hearing impaired people can communicate through technology with others more easily, and it also helps to promote the development of education, healthcare and public service accessibility.

This study focuses on a word-level sign language recognition which is the recognition of a specific word from a short duration video. Unlike speech recognition sign language recognition requires simultaneous understanding of the temporal continuity of image content and actions. The word level recognition task has practical applications. However, it faces many challenges in the modeling process. These challenges

include inadequate image feature extraction, limited temporal modeling capability, excessive model size and unstable training.

Currently, in the existing work, Convolutional Neural Networks (CNNs) are commonly used to extract hand and body movement features from images. LSTM and BiLSTM are widely used to model the temporal structure of sign language movements [15]. Attentional mechanisms have also been shown to be effective in enhancing the model's attention to key frames, thus improving the accuracy of recognition [1]. Furthermore, in order to reduce the computational complexity, lightweight models such as MobileNetV3 have been introduced to sign language recognition tasks, and they have shown good performance and efficiency in mobile vision tasks [2] [13].

Existing methods have made some progress in feature extraction and temporal modeling. However, simplifying the model structure while ensuring the recognition accuracy is remains challenging. In addition, some of the work ignores the impact of training strategies on model stability. For example, a single way of data enhancement or lack of regularization mechanism can easily cause overfitting. In order to solve these problems, this study proposes a lightweight word-level sign language recognition model. The model uses MobileNetV3 as the image feature extractor, combines with BiLSTM to deal with temporal relationships, and introduces an attention mechanism to optimize frame-level feature fusion. During the training process, various image augmentation methods are added and label smoothing techniques are combined to improve the robustness of the model. In its complete formthe project is based on the publicly available WLASL-300 dataset. The focus is on model structure design and training strategy optimization, and does not involve actual deployment or cross-modal fusion.

A. *The main contributions of this study include:*

- A word-level sign language recognition model integrating MobileNetV3, BiLSTM and attention mechanism.
- Introducing image enhancement and label smoothing strategies to improve the training stability and generalization ability.
- Conducting systematic experiments on the WLASL-300 dataset to validate the effectiveness of the model.

B. *The structure of the paper is as follows:*

Section 2: Introduces the research background and existing methods.
Section 3: Describes the model architecture and the training process in detail.
Section 4: Reports the experimental setup and the result analysis.
Section 5: Summarizes and proposes the direction of future research.

II. Literature Review

Sign language recognition (SLR) is an important technology that connects the deaf community with the outside world. It is a key area of research in artificial intelligence and human computer interaction. As

AI and computer vision have developed, SLR has evolved from an approach that relied on sensors to an intelligent system that recognises images and videos using deep neural networks. A variety of solutions for SLR have been proposed by researchers. These include feature acquisition, modelling methods, network structure and training strategies.

This section reviews related work from three categories of mainstream approaches to SLR. These are traditional sensor based approaches, vision based deep learning approaches and lightweight and deployment optimisation directions.

A. Traditional sensor-based approaches

Early sign language recognition systems relied heavily on external sensors. Examples include data gloves, IMUs (inertial measurement units) and Kinect, which capture the trajectory, strength and spatial position of gesture movements. These methods are highly accurate in closed environments. However, the high cost of the equipment and the poor user experience make it difficult to extend their use to everyday scenarios. Additionally, sensor systems are highly dependent on the user's operating mode and present an individual adaptation problem. Although some studies have attempted to reduce the computational burden by compressing the sensor model, its dependence on specialised hardware determines its limitations [5] [12] [14]. As camera performance improves and deep learning develops, research is gradually shifting towards more natural and convenient visual recognition methods.

B. Vision based deep learning methods

As computer vision and deep learning have developed rapidly, image-and video-based visual recognition methods have become the main direction of SLR. These methods typically comprise two phases: spatial feature extraction and temporal modelling. Some studies also introduce an attention mechanism to improve the model's discriminative ability.

In terms of spatial feature extraction, early research mostly used deep convolutional networks (e.g. ResNet or VGG) to extract features from video frames individually, which has the advantages of strong expressive ability and good generalisation. However, the model parameters are very large, and training and deployment costs are high [1]. In order to address the issue of computational resources, lightweight convolutional networks have emerged as a popular area of research. MobileNetV2 is one of the earlier examples of this approach, using depth-separable convolution to reduce complexity, making it suitable for mobile applications. However, its recognition performance is limited [12] . Subsequently, MobileNetV3 was developed, integrating the SE (Squeeze and Excitation) module and the H-Swish activation function to improve expressive capability while maintaining a lightweight structure [13]. Building on this, some researchers have constructed an improved MobileNetV3 model for feature extraction, achieving greater accuracy while reducing the number of parameters [2].

In addition to convolutional networks, research has also been conducted into modelling the human body based on keypoints. Using Google MediaPipe to extract and classify the 21 points hand skeleton keypoints by a lightweight neural network resulted in a model with a very low number of parameters, achieving

an accuracy of 94.88% on a specific dataset containing 26 American Sign Language (ASL) categories [6]. However, this method discards a large amount of pixel-level spatial information by converting high dimensional image data into low dimensional keypoint data [6] . Although the keypoint based approach aims to reduce the influence of visual characteristics such as background, illumination and clothing in order to improve model generalisation, keypoint extraction tools such as MediaPipe may be inadequate for hand interactions involving complex gestures, resulting in the loss of key information. Furthermore, MediaPipe's gesture estimation performance may degrade in challenging visual conditions such as occlusion, poor lighting, or low resolution, affecting the overall recognition results [5] [6] .

In temporal modelling, recurrent neural networks (RNNs) such as long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) are widely used to process video sequences. LSTM can capture inter-frame dynamics and excel in processing temporal information, making it suitable for action recognition in sign language [1] [15] . BiLSTM introduces a bidirectional learning mechanism that can capture context information from before and after simultaneously [15] . However, traditional recurrent networks have problems with long-distance dependence and poor parallelism compared with Transformer structures. In recent years, some studies have adopted the Transformer structure to enhance global context modelling capabilities [12] . For instance, the TSPNet model, which is based on hierarchical semantic modelling and an attention mechanism, significantly improves accuracy in sentence-level translation tasks [7]. However, its high training overhead and complex structure mean that it is not suitable for direct deployment in lightweight scenarios with limited resources, when compared with some models that are designed to be lightweight.

In addition, introducing the attention mechanism improves the model's ability to identify key frames. Some researchers have added attention weighting to the LSTM output layer to help the model focus on the frames at gesture transitions, thereby improving recognition accuracy [1] . Further research has embedded the self-attention mechanism into the feature modelling process, enabling the model to automatically assign temporal weights and suppress redundant frames more effectively [7] [10]. While the attention mechanism enhances model performance and interpretability, it is primarily an add-on module that is not tightly integrated with the backbone network, leaving room for optimisation.

C. Direction of model lightweight and training optimisation

As the computational power of edge devices increases, the deployability of SLR models is receiving more and more attention. The design of lightweight structures is becoming the focus of research. This includes methods such as network compression, shallow structures, key point input, and efficient training strategies. A researcher proposed a lightweight CNN structure. A sign language recognition model suitable for deployment was constructed using a strategy involving structure tailoring and channel optimisation [14]. Another class of methods uses skeleton keypoints as inputs. Removing the image texture information drastically reduces the number of model parameters. This approach offers significant advantages in terms of lightness. However, it neglects local texture and gesture pose details. The accuracy of this method's recognition suffers when faced with disturbances such as complex backgrounds or occlusions [6]. In

contrast, the lightweight CNN with MobileNetV3 as its backbone structure retains strong feature expression capability while maintaining model compactness. A good balance between accuracy and model size has been achieved in several tasks [2] [13].

In addition, there is structural optimisation. The training strategy also has a significant impact on model performance. Related research systematically evaluates various image data augmentation methods. These include colour perturbation and brightness contrast adjustment, among others. The results show that these methods can effectively improve the model's ability to generalise in different scenes [3]. Additionally, studies have introduced Label Smoothing regularisation to mitigate the model's overfitting to a single category. In particular, it has demonstrated good stability with small sample data sets [10]. However, most current training techniques are designed independently and lack close integration with the network structure. Therefore, there is still room for optimisation in terms of generalisation ability and convergence efficiency.

D. Summary and research gaps

In summary, although existing studies have advanced the development of SLR from various angles, the following gaps still remain:

- Some methods are insensitive to model volume and lack deployment considerations.
- Temporal modelling methods are not unified, and some studies only rely on rough strategies such as average pooling.
- The training process is unstable and prone to overfitting, and there is a lack of systematic optimisation of the training strategy

E. Positioning of this study

This study attempts to address the research gaps identified from my literature review. My study focuses on a lightweight MobileNetV3-based structure which combines a BiLSTM for processing video timing and a frame-level attention mechanism for improving the model's ability to perceive key information. Image enhancement and label smoothing strategies are also introduced to improve training robustness and generalisation ability. I validate the model using the WLASL_300 dataset. I use the word-level recognition task to try to strike a balance between accuracy, structural complexity and training efficiency which is a key focus of my study.

III. METHODOLOGY

This section will introduce the implementation method of this research in sign language recognition task. It includes: dataset selection, data preprocessing and enhancement strategies, model structure design, training process and parameter setting. As well as the evaluation method and index selection. The overall goal is to design a lightweight, efficient and robust word-level sign language recognition model.

A. *Dataset (WLASL-300)*

The WLASL-300 dataset was used in this study. This is a publicly available word-level American Sign Language video set. It contains 300 sign language lexical categories covering a wide range of movement types and gesture variations. The dataset contains three parts: a training set, a validation set, and a test set. As shown in Table I .Each video contains a signer performing a complete gesture movement, with varying number of frames, and some variations in background and lighting. This dataset has good diversity and public availability, and is a standard benchmark widely used in word-level SLR research.

TABLE I
WLASL-300 DIVISION

Dataset Split	Sample Size	Number of Classes
Training Set	2488	300
Validation Set	649	300
Test Set	530	300

B. *Data Preprocessing and Data augmentation*

In the data preprocessing stage, I extracted 12 keyframe images from each video to capture the key moments of sign language actions. As shown in Figure 1. Then the images were uniformly resized to 128×128 pixels and normalized to improve the training stability of the model.

In order to observe the generalization ability of the model, I compared two sets of training methods:

Basic augmentation (experiment 1): only have resize and normalize.

Advanced augmentation (experiment 2): RandomBrightnessContrast: randomly adjust brightness and contrast (p=0.4). GaussNoise, MotionBlur, and MedianBlur are selected to simulate the noise and blur in the shooting.

These augmentations not only improve the robustness of the model, but also effectively prevent over-fitting.

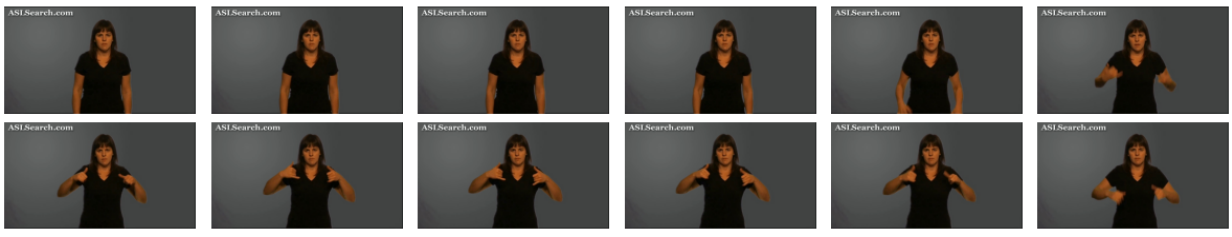

Fig. 1. extract 12 keyframe images from a video

C. *Model architecture design*

The model consists of three main parts: MobileNetV3-Large, BiLSTM and the attention mechanism. MobileNetV3 is used as a feature extractor, responsible for extracting 960-dimensional spatial features from each image frame. Next, the BiLSTM model is used to capture temporal information between frames,

producing 128-dimensional temporal features. Finally, the attention mechanism generates weighted features for classification by calculating attention weights and focusing on key frames. As shown in Figure 2, the model's overall architecture involves the data stream going through feature extraction, temporal modelling and attention weighting from the input video to the final category prediction. This architecture enhances the model's temporal modelling and keyframe understanding, while ensuring computational efficiency.

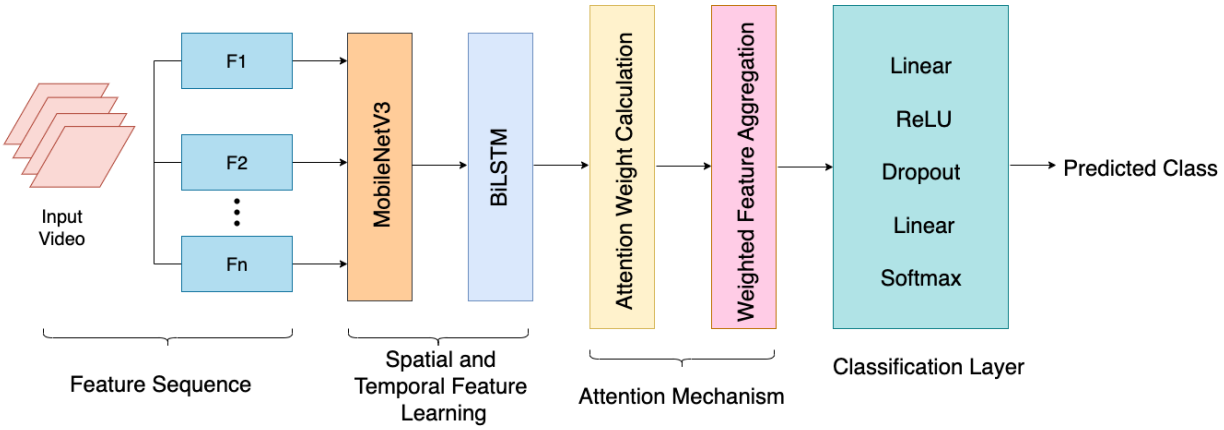

Fig. 2. model architecture design

D. Training Process and Parameter Settings

The following settings were used for the training process. Data loading was performed using PyTorch's DataLoader for batch loading. The batch size was 8. The optimizer was selected as Adam, the learning rate was set to 1e-4 and the weight_decay was 1e-4. 60 rounds of training were performed. An early stopping strategy was used to prevent overfitting. The learning rate scheduling was dynamically adjusted based on the validation set performance. The convergence and generalization ability of the model is ensured. During training, training and validation losses were monitored to ensure model stability and effectiveness. And experiments are conducted under two augmentation conditions separately to observe the effect of image perturbation on model stability. As shown in Table II , lists the detailed training parameter settings.

TABLE II
TRAINING PARAMETER SET

Parameter	**Value**
Optimizer	Adam
Initial Learning Rate	1e-4
Weight Decay	1e-4
Learning Rate Scheduler	ReduceLROnPlateau (reduce learning rate when validation set does not improve)
Batch Size	8
Maximum Epochs	60
Early Stopping	Stop training if validation set does not improve for 5 consecutive epochs
Loss Function	CrossEntropyLoss (with Label Smoothing = 0.1)

E. Evaluation Methodology and Metrics Selection

Model performance is evaluated by a variety of metrics, including accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of the prediction. Precision rate assesses the accuracy of predicting the target category. Recall measures the degree of coverage of the samples in the target category.The F1 score is the harmonic mean of precision and recall. In addition, a detailed analysis was performed using the confusion matrix and classification reports to identify patterns of model error and room for improvement.

IV. Experimental Results and Analysis

This section will show and analyze the key experimental results of this research. The main topics include: comparison experiments of model input configurations and augmentation strategies. Performance metrics evaluation of the final model. The difference in performance between the base model and the augmented model in the classification task. The experiment follow the idea of "from simple to complex, from comparison to optimization" to gradually build and verify the effectiveness of the lightweight sign language recognition system.

A. Experimental Design and Procedure

To verify the effects of different designs on model performance I completed the following experimental stages which I describe in the following sections.

- **Dataset selection**

WLASL-100 and WLASL-300 data sets were used respectively. This was important to compare their effects on model training stability and validation accuracy. The experimental results show that although WLASL-100 is less complex, it has limitations in generalization ability and category coverage.

Benefits of the WLASL-300 in sign language recognition research:

The amount of data is significantly increased. As shown in Table III. The training, validation, and test sets contain 2488, 649, and 530 samples, respectively, which is approximately 2.5 times the size of WLASL-100. This not only provides more training examples for the model, but also bring a more challenging evaluation environment. With more categories and more complex gesture differences, WLASL-300 is more suitable as a benchmark for evaluating lightweight models, which helps the model learn more discriminative features and reduces the risk of overfitting.

TABLE III
WLASL-300 AND WLASL-100 DIVISION

Dataset Split	WLASL-300	WLASL-100
Training Set	2488	999
Validation Set	649	242
Test Set	530	202

The larger data size also improves the generalization ability and robustness of the model. Bigger test sets can more reliably evaluate the performance of models on unseen data. It is closer to real usage scenarios. Concurrently, the model is more stable and has less performance fluctuation when choosing hyperparameters or early stopping judgments. This facilitates the controllability of the training process.

Although WLASL-300 belongs to a medium-sized dataset. However, the 300 words it covers can already meet many daily communication needs. It has a certain semantic breadth and practicality. As a training and evaluation target in the intermediate stage, it strikes a good balance between model complexity and task difficulty, and lays the foundation for subsequent scaling to larger vocabularies (e.g., WLASL-1000). This incremental strategy is well suited to the research path of lightweight modeling, and is a reasonable choice for balancing feasibility and scalability.

- **Input frame selection**

Every videos is extracted from 8, 10, 12 and 16 frames respectively for comparison experiments. The results show that under the same model structure, the 12 frame configuration achieves a better balance between accuracy and computational efficiency, and performs best in the validation set. So finally, choose 12 frames as the keyframe input. The choice of 12 frames as the number of keyframes for each video is relevant for the following reasons.

1) **Selection of the number of keyframes (12 frames)**
   First of all, 12 frames strikes a good balance between computational efficiency and information retention which captures the key changes in gesture movements without imposing too much computational burden on the model. Second, considering the GPU memory limitation and the training efficiency. 12 frames is a reasonable choice after trade-offs to ensure the stability of the training process while maintaining the performance of the model.

2) **Keyframe selection strategy**
   A hybrid sampling strategy is used to determine the location of keyframes rather than simple equal interval sampling. This strategy consists of three layers. First, it keeps the first and last frames of the video. This ensures getting the start and end states of the gesture action. It is important for understanding the complete gesture sequence. Second, dynamic sampling is performed based on inter-frame motion differences. The moments with the largest changes in motion are identified by calculating the difference in the grayscale maps between neighboring frames. These moments usually correspond to key turning points of the gesture. Finally, if the number of frames obtained by dynamic sampling is less than 12 frames. Then equal interval sampling is used to supplement the remaining frames, ensuring that a fixed number of key frames are obtained for each video.

3) **Dynamic sampling experiments**
   A dynamic sampling algorithm based on motion variations is implemented. This algorithm identifies the most significant moments of motion in a video by calculating the difference between neighboring frames. The specific implementation consists of calculating the mean square error (MSE) of the inter-frame grayscale maps, represented in equation (1). and selecting the most representative frames

based on these difference values. This dynamic sampling method has significant advantages over simple equal interval sampling. This is because it can better capture the critical moments of gesture movements. Especially those frames with drastic changes in motion, these frames usually contain richer gesture information.

$$D_t = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(I_t(i,j) - I_{t+1}(i,j) \right)^2 \tag{1}$$

$I_t(i,j)$ denotes the grayscale value of the pixel at position (i,j) in frame t, and H, W are the height and width of the image, respectively. D_t represents the inter-frame difference metric, which is used as a reference for dynamic sampling.

4) **Visual Validation**

In order to verify the effectiveness of the keyframe extraction strategy. I implemented the keyframe visualization function.As shown in Figure 3, a video sample is randomly selected to show the results of extracting 8 frames, 10 frames, 12 frames and 16 frames. As can be observed from the figure, 12 frames not only completely cover the start, middle change and end states of the gesture, but also avoids the inter-frame redundancy as in the 16 frame sampling. In addition, 12 frames can more accurately reproduce the details of gesture transitions than 8 and 10 frames. A good balance between expressiveness and computational efficiency. This experiment shows that the hybrid sampling strategy can effectively recognize and retain the key information in gesture actions. Thus, the temporal modeling ability of the model is further improved.

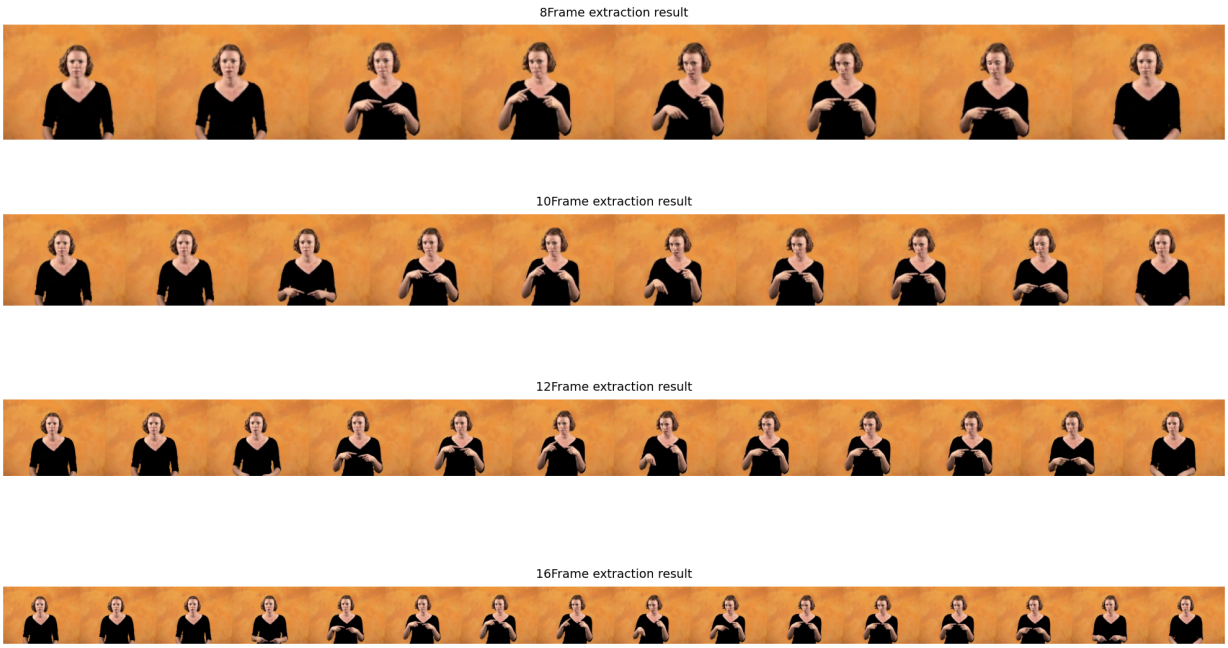

Fig. 3. from top to bottom are sampled 8, 10, 12 and 16 frames

- **Comparison of enhancement strategies**

Experiment 1 uses base enhancement (Resize and Normalize only). Experiment 2 adds multiple image perturbations (e.g., luminance perturbation, blurring, and noise), combined with label smoothing to improve robustness.

1) **Data augmentation probability selection strategy**
   The selection of data augmentation probability is based on the balanced consideration of the specificity of the sign language recognition task and the computational efficiency. For 'RandomBrightnessContrast', a probability value of 'p=0.4' is chosen, and this choice is based on several factors. First, the sign language recognition task is very sensitive to detailed information (e.g., gesture shape, finger position, hand movement trajectory). Too high a probability of augmentation may destroy these key features. And so a relatively conservative augmentation strategy was adopted. Second, the probability value of 0.4 provides enough data augmentation to improve the generalization ability of the model without overly altering the original image features. The integrity of the key information of the sign language movement is maintained.

2) **Data augmentation strategy design**
   The data augmentation policy adopts a hierarchical design. It contains two configurations, the base policy and the augmentation policy. The base policy contains only the necessary preprocessing operations (resize and normalize). The augmentation strategy, on the other hand, introduces a variety of data augmentation techniques. It includes a combination of techniques such as brightness and contrast adjustment (p=0.4), gaussian noise (p=0.2), motion blur and median blur. And the'A.OneOf()'combination mechanism was used to allow selection of one of the multiple augmentation techniques. This design increases the diversity of the data while avoiding information loss due to over augmentation.

- **Network Architecture Unification**

Both sets of experiments use the architecture of MobileNetV3-large, BiLSTM , and Attention. Both sets of experiments also ensure that the variables can be controlled, and only compare the performance changes brought by the enhancement and training strategies.

- **Evaluation metrics**

Evaluation metrics include Accuracy, Precision, Recall, F1-Score, and adding confusion matrix and attention visualization.

B. Experimental Results

- **Accuracy Comparison**
  Table IV shows the performance of the two experimental schemes in the training and validation phases:13 From the results, it can be seen that the augmented model performs better on both the training set and the validation set. The accuracy in the training set is improved. This indicates that the augmentation strategy can improve the model generalization ability and training effect.
- **Classification Performance Metrics**

TABLE IV
COMPARISON OF TRAINING AND VALIDATION ACCURACY

Metric	Exp1 (Basic)	Exp2 (Augmented)	Difference
Best Val Acc	66.44%	67.46%	+1.02%
Final Train Acc	92.89%	95.94%	+3.05%
Final Val Acc	66.27%	67.29%	+1.02%

In addition, Accuracy, Precision, Recall and F1-Score were computed to evaluate the model on the test set. As shown in Table V, the enhanced model shows a small improvement in all four metrics. This indicates that the added data enhancement and label smoothing strategies improve the model's sensitivity to boundary samples and a few categories. The overall generalization ability is improved.

TABLE V
COMPARISON OF PERFORMANCE METRICS

Model	Accuracy	Precision	Recall	F1-Score
Experiment 1 (Basic)	0.6644	0.3817	0.3967	0.3771
Experiment 2 (Augmented)	0.6746	0.3963	0.4215	0.3963

- **Confusion Matrix Analysis**

  Figure 4 and Figure 5 show the confusion matrices for the base and augmented models, respectively. The confusion matrix is used to show the relationship between real labels and predicted labels, and the closer to the diagonal line indicates the more accurate classification.

  As can be seen from Figure 4 (Experiment1_Basic). The model is able to make relatively accurate predictions on most categories. Most of the categories have reached more than 10 correctly predicted samples. This indicates that the model has good discrimination ability for these types of gestures. However it can also be observed that a few categories (e.g. "accident") perform slightly weaker in the validation set. It indicates that the model may still be confusing for certain gestures with similar shapes or inconspicuous movements.

  In contrast, Figure 5 (Experiment2_Augmented) shows the confusion matrix of the augmented model. In general, the number of correct predictions is basically the same or even slightly improved, and the classification is more balanced. It indicates that the augmented model has more stable cross-category generalization ability.

C. *Interpretability Analysis of Model Decisions*

- **Interpretable Design of Attention Mechanisms**

  The model utilizes a bidirectional LSTM and attentional mechanism architecture. This makes the model's decision-making process highly interpretable. The attentional weights can clearly show which key frames the model pays attention to when making classification decisions. This design not only improves the performance of the model but also provides a transparent decision-making process. It enables us to understand how the model makes classification decisions based on temporal information.

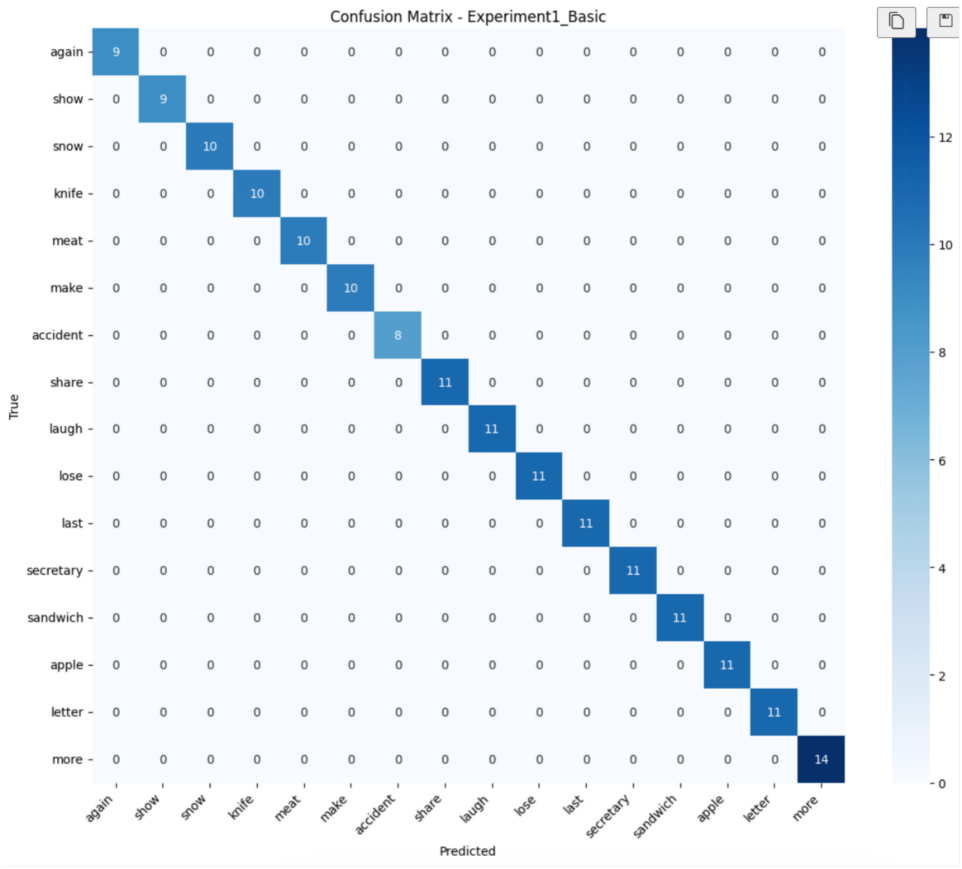

Fig. 4. Experiment1_Basic confusion matrix

- **Attention weight visualization implementation**
  As shown in Figure 6, a complete attention weight visualization is implemented to intuitively show how much attention the model pays to different frames.

- **How the Attention Mechanism Works**
  In the model structure, the bidirectional LSTM is firstly responsible for processing the input temporal features and outputting a hidden state sequence of shape [B, T, 2H]. Where B is the batch size, T is the number of time steps, and 2H is the hidden dimension after bi-directional splicing. Next, using an attention module, the corresponding attention weights are computed for the features at each time step. These scores are then transformed into a probability distribution via a softmax function.
  Based on this, the model performs a weighted summation of all temporal features based on the attention weights at each time step. A weighted global feature representation is obtained and used as the final classification input. This design allows the model to automatically focus on those moments that are more critical to the recognition result.

- **Interpretable analysis results**
  By visualizing the analysis of attention weights several important patterns can be found. First, the model tends to focus on the beginning, intermediate key change points and ending frames of the gesture action. This is highly consistent with the temporal nature of sign language recognition.

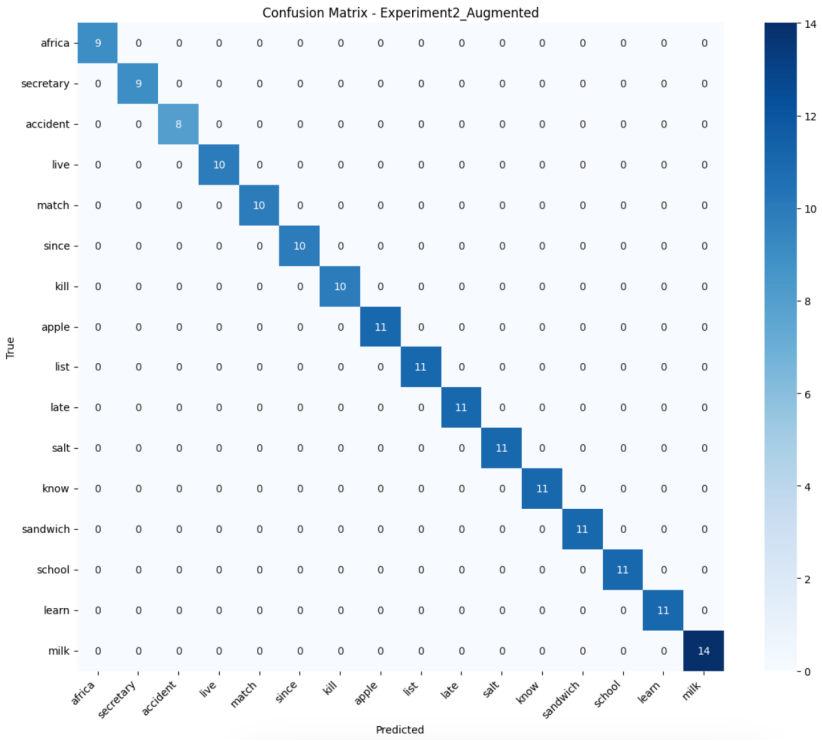

Fig. 5. Experiment2_Augmented confusion matrix

Second, as shown in Figure 7, the frames with the highest attentional weights usually correspond to the key turning points of the gesture action, such as hand shape change, movement direction change, etc. Finally, by analyzing the attentional weights, it is possible to understand why the model makes specific classification decisions, which provides important support for the reliability of the model.

- **Practical application value and significance**
  This interpretable analysis has significant practical value. First, by observing whether the attention weights are as expected (focus on keyframes). It can be verified whether the model has learned meaningful features. Second, when the model predicts incorrectly, analyzing the attention weights can help us understand the reason for the error and thus guide the model improvement. In addition, the analysis of attention weights can guide us to optimize the model architecture. For example, adjusting the keyframe extraction strategy. Most importantly, an explainable decision-making process can enhance users' trust in the model prediction results, which is crucial for user acceptance in real-world applications.

D. Comparison with the Existing Techniques

To show the performance advantages of the proposed methods in this paper more comprehensively, I have selected some sign language recognition methods used in recent years for comparison. Different types of network structures, input modalities and sequence modeling strategies are covered. As shown in Table

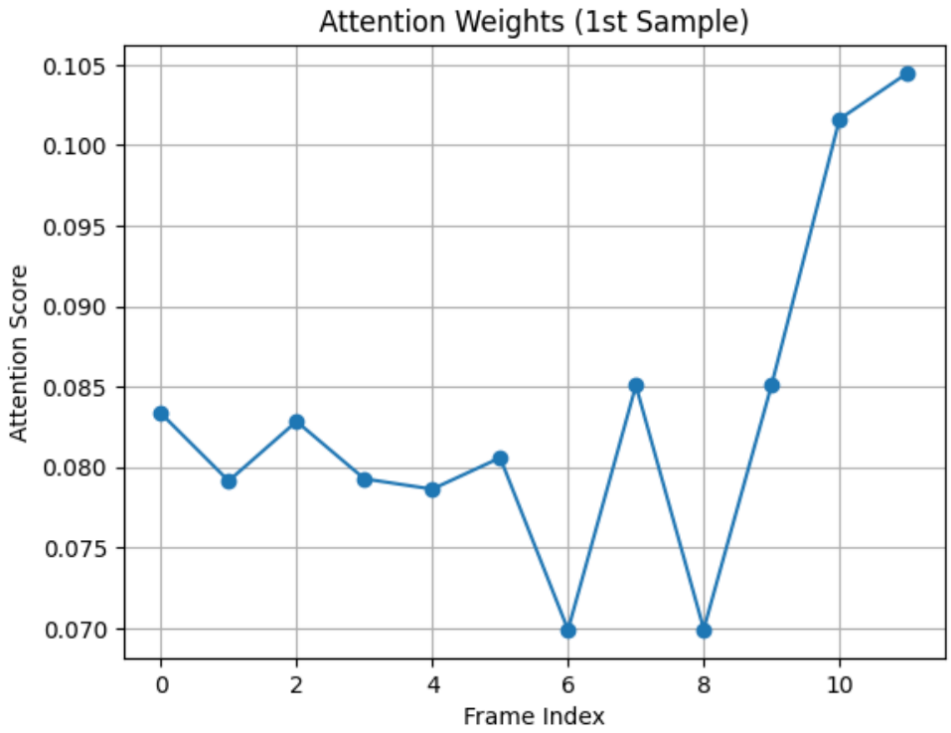

Fig. 6. attention weight

6. The backbone network, modeling temporal sequence information, the use of Attention mechanism, the dataset used and its Top-1 accuracy of each method are summarized.

As shown in Table VI, some of the methods rely on information or introduce structurally complex Transformer mechanisms. Although they have achieved high accuracy rates on specific datasets, they generally have problems such as high model complexity, high consumption of training resources or strong deployment dependency. Also, some lightweight models have also achieved better accuracy rates.

This project makes a contribution based on the MobileNetV3 backbone, combined with BiLSTM to accomplish temporal modeling, and introduces an attention mechanism to enhance the keyframe response, using only RGB image inputs. Under the condition of epoch=60, a Top-1 accuracy of 67.46% is obtained on the WLASL-300 dataset. This represents good generalization ability and structural efficiency are demonstrated. Compared with other methods with complex structures or relying on multimodal inputs. Accordingly, this model has advantages in terms of deployment friendliness and performance balance.

E. Comparative Analysis

Combining the above results, the enhancement model shows a steady improvement in key metrics such as precision rate, recall rate and F1 score. This indicates that adding image enhancement and label smoothing strategies has a positive effect on the generalization and robustness of the model. Compared with the base model, the augmented model is more robust under the conditions of illumination change, motion blur or background interference.

Fig. 7. the frame with the highest attention weight

F. Efficiency Analysis

The final model only 3.62M parameters and 0.93 GFLOPs, with a compact size of 14.07 MB. On an RTX 4090 GPU, the average inference latency per video is 1.33 ms, the throughput reaches 751.07 videos/s, and the peak GPU memory usage is only 40.09 MB. These results demonstrate that the system achieves both structural efficiency and suitability for deployment on resource-limited devices.

V. Conclusion and Future Work

A. Research Summary

This research addresses the word-level sign language recognition task. A lightweight MobileNetV3-large, BiLSTM, attention-based architecture was adapted and optimized for improved robustness and deployment efficiency.And a systematic experimental validation was done on the WLASL-300 dataset. In the training strategy, image augmentation and label smoothing are combined to improve the generalization ability and stability of the model. The experimental results show that the performance of the model after these augmentation has been significantly improved.

TABLE VI
COMPARISON OF SLR METHODS

Method	Backbone	Temporal Modeling	Attention Mechanism	Dataset	Top-1 Accuracy
CNN+LSTM+ Attention [1]	MobileNetV2	yes	yes	WLASL-100	84.65%
PoseFormer [4]	PoseFormer	1D Temporal Conv+Self-Attention	Multi-Head Attention	WLASL/AUTS	61.4%
Transformer [9]	CNN+Transformer	yes	Memory-augmented Temporal Attention	MSASL100	83.04%
SIGNGRAPH [16]	ResGCN	2D Temporal Convolution	–	WLASL-1000	72.73%
This project	MobileNetV3	BiLSTM	yes	WLASL-300	67.46%

B. Project Strengths and Weaknesses Analysis

This study's strengths are mainly in the lightweight structure and the effectiveness of the training strategy. The model parameter count is small, and the training and inference efficiency is high, which is suitable for scenarios with limited resources. At the same time, the frame-level attention mechanism enhances the model's ability to perceive key actions, and the data augmentation and regularization means significantly improve the stability of the model. However, the study still has some shortcomings. For gestures with highly imbalanced categories or highly similar action patterns, the confusion problem of the model is still obvious. In the case of complex backgrounds, unstable lighting conditions, or partially occluded hand movements, the model recognition performance degradation is more obvious.

C. Real-World Deployment Possibilities and Challenges

From the perspective of real-world deployment, the model is compact, fast reasoning, and has some potential for end-side deployment. However, the complexity of real scenarios still brings multiple challenges. First, the arithmetic limitations of low-power devices may affect the real-time performance. Second, the application environment has many background interferences, light fluctuations, and the model is more sensitive to external noise. Finally, there are some differences in the gesture habits of different users, and the lack of personalized adaptation may lead to a decrease in recognition accuracy. Therefore, further improving the robustness of the model in diverse usage scenarios is an important research direction in the future.

D. Future Work

Future work can be carried out in the following directions. First, in model lightweight optimization it can be explored more aggressively through compression strategies through knowledge distillation techniques to guide the training of the current model with a larger teacher model (e.g., ResNet or EfficientNet).

Second, in temporal modeling, the existing attention mechanisms can be optimized by trialling the use of a lightweight Transformer to replace LSTM in the future. Thia may offer a more efficient timing modeling capability. In addition, it can study the adaptive frame sampling method. Redundant frames are dynamically skipped based on the attention weights, thus further improving the inference efficiency.

Third, in terms of data augmentation and multimodal fusion, for the special characteristics of sign language recognition. Data augmentation strategies that are more in line with the characteristics of sign language which can be designed such as hand key point perturbation, background replacement, light change, etc., could all improve the robustness of the model in complex operating environments. At the same time, multimodal information fusion can be explored to combine the visual flow with features such as hand keypoints and facial expressions, and the keypoint information extracted by tools such as MediaPipe can be used to improve the model's recognition stability in occlusion or fast-motion scenes.

Finally, in terms of deployment and practicality, a complete mobile end evaluation system could be established which is vital to test end-to-end latency, memory usage and battery consumption on real mobile devices to establish reproducible benchmarks. At the same time, edge computing optimizations such as model pruning, dynamic batch processing, and other techniques can be explored to ensure the usability of the model in real applications.

REFERENCES

[1] D. Kumari and R. S. Anand, 'Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework Based on Attention Mechanism', *Electronics*, vol. 13, no. 7, Art. No. 7, Jan. 2024, doi: 10.3390 / electronics13071229.

[2] L. Zhao and L. Wang, 'A new lightweight network based on MobileNetV3', *KSII Transactions on Internet and Information Systems,* vol. 16, no. 1, pp. 1–15, Jan. 2022.20

[3] T. Kumar, R. Brennan, A. Mileo, and M. Bendechache, 'Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions', *IEEE Access,* vol. 12, pp. 187536–187571, 2024, doi: 10.1109 / ACCESS. 2024.3470122.

[4] T. Vandendriessche, M. D. Coster, A. Lejon, and J. Dambre, 'Representing Signs as Signs: One-Shot ISLR to Facilitate Functional Sign Language Technologies, Feb. 27, 2025, arXiv: arXiv:2502.20171. doi: 10.48550 / arXiv. 2502.20171.

[5] Z. Long, X. Liu, J. Qiao, and Z. Li, 'Sign Language Recognition Based On Facial Expression and Hand Skeleton', in 2023 38th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Aug. 2023, pp. 237–241. doi: 10.1109 / YAC59482.2023.10401630.

[6] I. Siju and P. Selvam, 'A Novel Approach for Lightweight Sign Language Recognition Leveraging Google Mediapipe and Deep Neural Net', in 2024 First International Conference on Software, Systems and Information Technology (SSITCON), Oct. 2024, Pp. 1-6. Doi: 10.1109 / SSITCON62437.2024.10796746.

[7] D. LI et al., 'TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation', in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 12034–12045.

[8] 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', presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1459–1469.

[9] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, 'Transferring Cross-Domain Knowledge for Video Sign Language Recognition', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6205–6214.

[10] R. Zuo, F. Wei, and B. Mak, 'Natural Language-Assisted Sign Language Recognition', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14890–14900.

[11] M. K. Fallah, M. Najafi, S. Gorgin, and J.-A. Lee, 'An ultra-low-computation model for understanding sign languages', *Expert Systems with Applications,* vol. 249, p. 123782, Sept. 2024, doi: 10.1016 / j. swa. 2024.123782.

[12] Y. Zhang and X. Jiang, 'Recent Advances on Deep Learning for Sign Language Recognition', *CMES,* vol. 139, no. 3, Pp. 2399-2450, 2024, doi: 10.32604 / cmes. 2023.045731.

[13] A. Howard et al., 'Searching for MobileNetV3', presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.

[14] Y. S. Nugroho, C.-K. Yang, and Y.-C. Lai, 'Lightweight American Sign Language Recognition using a Deep Learning Approach', in Natural Language Processing and Machine Learning, Academy and Industry Research Collaboration Center (AIRCC), May 2023, pp. 75-90. Doi: 10.5121 / csit. 2023.130807.

[15] N. Hassan, A. S. M. Miah, and J. Shin, 'A Deep Bidirectional LSTM Model Enhanced by Transfer-Learning-Based Feature Extraction for Dynamic Human Activity Recognition, *Applied Sciences,* vol. 14, no. 2, Art. no. 2, Jan. 2024, doi: 10.3390/app14020603.

[16] 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, doi: 10.1109/ACCESS.2023.3247761.