# BACHELOR THESIS

## RESEARCH AND DEVELOPMENT OF A SIGN LANGUAGE RECOGNITION SYSTEM TO SUPPORT HEARING IMPAIRED COMMUNICATION

Nguyễn Hoàng Minh – 22BI13291

External supervisor: Assoc. Prof. Nguyen Duc Dung - IOIT

Internal supervisor: Prof. Nghiem Thi Phuong

# Outline

## I: Introduction

- Context and motivation
- Objectives

## II. Theoretical Background

- Sign Language Recognition
- Spatial-Temporal Graph Convolutional Network (ST-GCN)

## III. Materials and Methods

- Dataset
- Model Architecture

## IV. Results and Discussion

- Model configuration and training
- Experimental Result and Discussion
- Models Limitations

## V. Conclusion and Future Work

- Conclusion
- Future Work

# I. Introduction

01. Context and motivation

02. Objectives

# I: Introduction

## 1. Context and motivation

- Sign language is a language that uses hand expressions to express to others
- Used by deaf people
- Real-life challenges: Few people know
- Sign Language Recognition is the process of recognizing and converting sign language into words

## 2. Objectives

- Sign language model that can recognize signs
- High accuracy
- Suitable for all devices

# II. Theoretical Background

01. Sign Language Recognition

02. Spatial-Temporal Graph Convolutional Network (ST-GCN)

# II. Theoretical Background

## 1. Sign Language Recognition

### Traditional Method

- Hidden Markov Model: Simple sequential model, weak in long and complex sequences.
- Dynamic Time Warping: Time alignment for sequence comparison, does not learn from data, only for simple actions.
- Support Vector Machine : Good for single characters, but difficult with long time series and complex transformations.

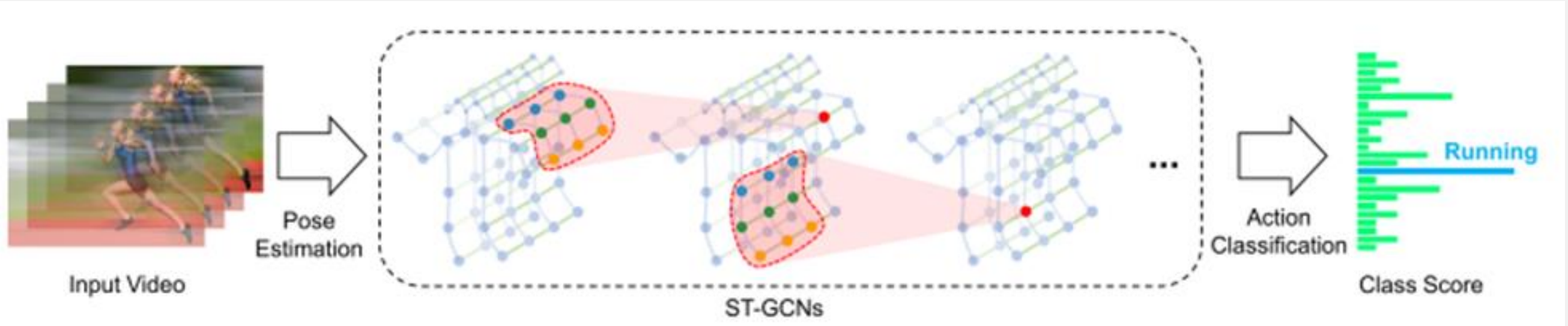
### Sign Language Recognition with Deep Learning

- Automatically learn spatio-temporal features from videos or keypoints
- Includes 2 main steps: Feature extraction

  Symbol classification

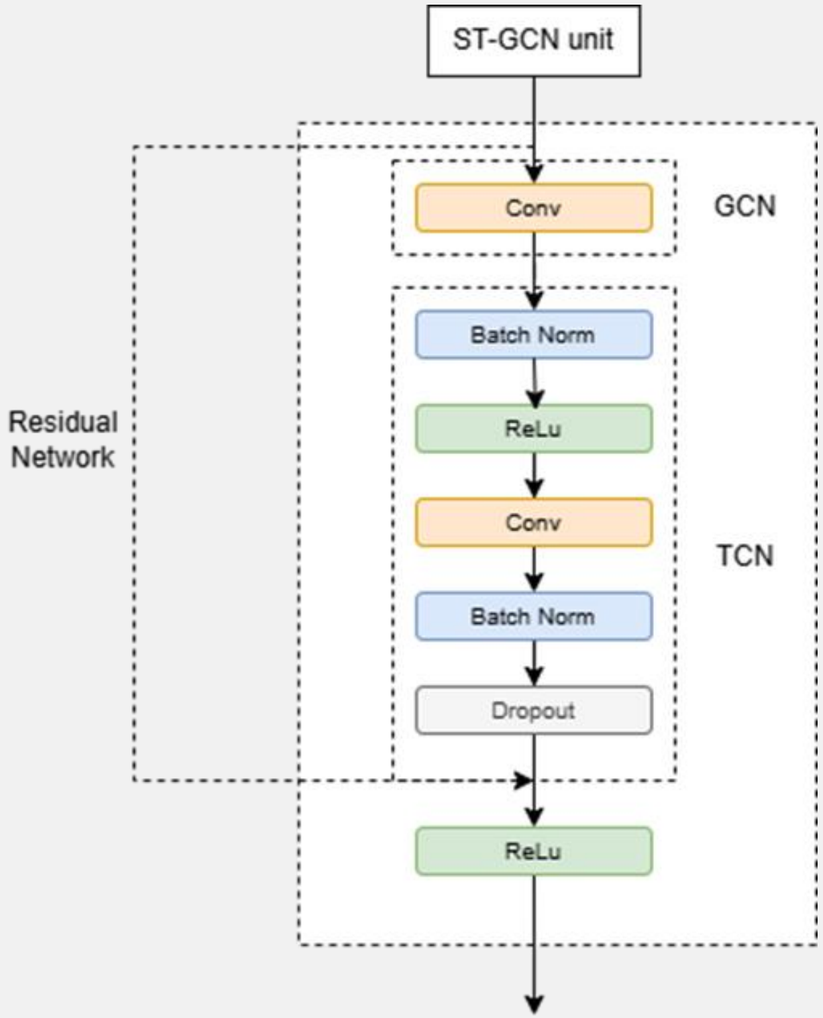
# II. Theoretical Background

## 2. Spatial-Temporal Graph Convolutional Network (ST-GCN)



# II. Theoretical Background

## 2. Spatial-Temporal Graph Convolutional Network (ST-GCN)




GCN is used to learn spatial features from the skeleton

TCN helps learn movement over time

# III. Materials and Methods

01. Dataset

02. Model Architecture

# III. Materials and Methods

## 1. Dataset

### 1.1. Data Collection

#### MS-ASL, WLASL100

| | Label | Video | Mean per label | Training | Validation | Testing |
|---|---|---|---|---|---|---|
| MS-ASL | 1000 | 25513 | 25.5 | 16054 | 5287 | 4172 |
| WLASL100 | 100 | 2038 | 20.4 | 2038 | 1442 | 338 |

# III. Materials and Methods

## 1. Dataset

### 1.2. Data Preprocessing

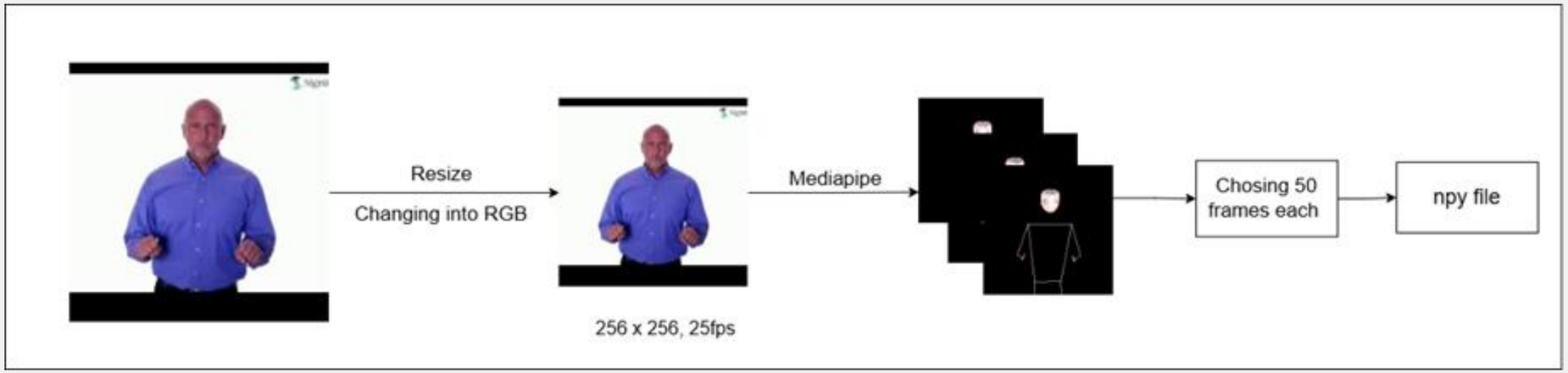
#### My trainning dataset

| | Label | Video | Mean per label | Training | Validation | Testing |
|---|---|---|---|---|---|---|
| My dataset | 100 | 4000 | 40 | 2800 | 800 | 400 |
| WLASL100 | 100 | 2038 | 20.4 | 2038 | 1442 | 338 |

# III. Materials and Methods

## 1. Dataset

### 1.2. Data Preprocessing



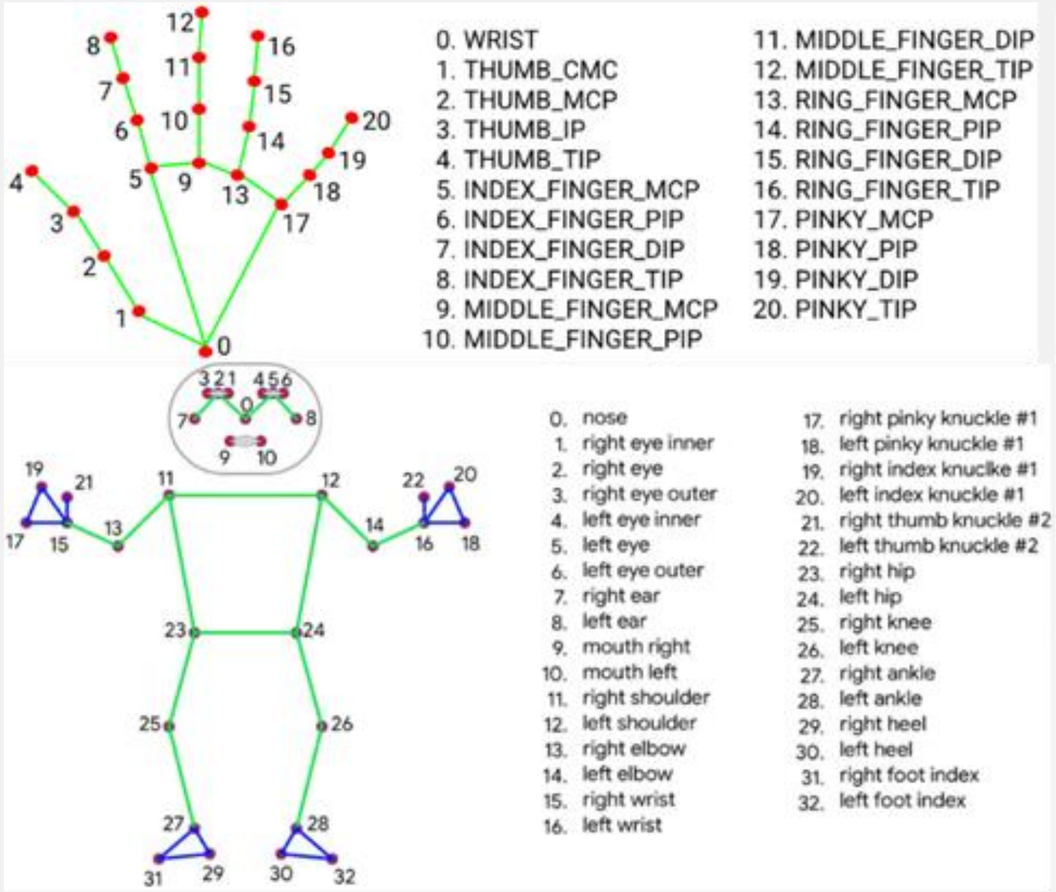

- Resize into 256x 256, 25 fps
- Changing color into RGB
- Load extracted video with 50 frames
- With videos more than 50 frames: Select the frame in the middle of the video
- For videos less than 50 frames: Add a black frame at the end of the video

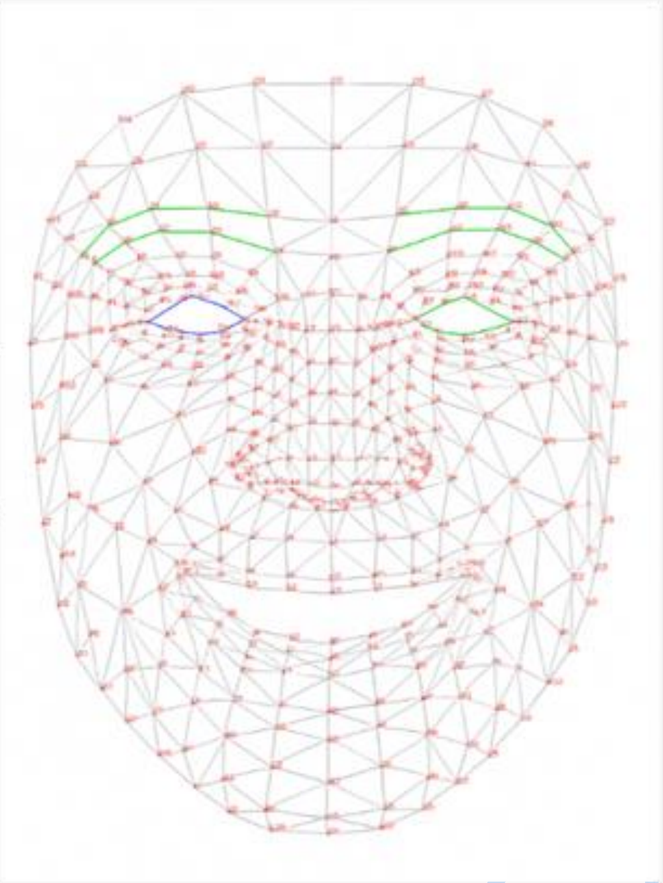
# III. Materials and Methods

## 1. Dataset

### 1.2. Data Preprocessing

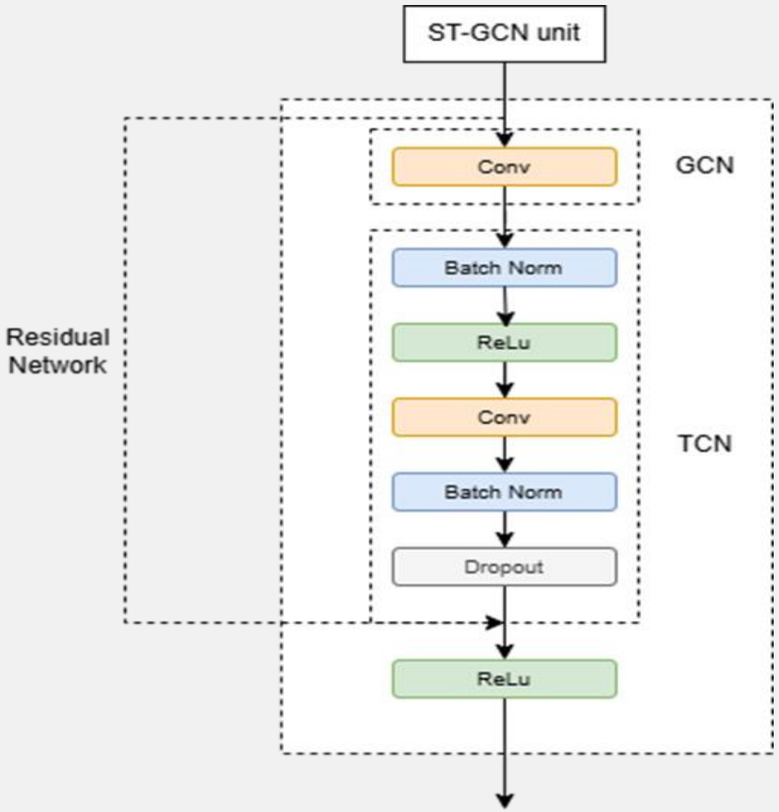
- Extract keypoint by Mediapipe Frameworks
- 21 hands, 33 pose and 468 face keypoints



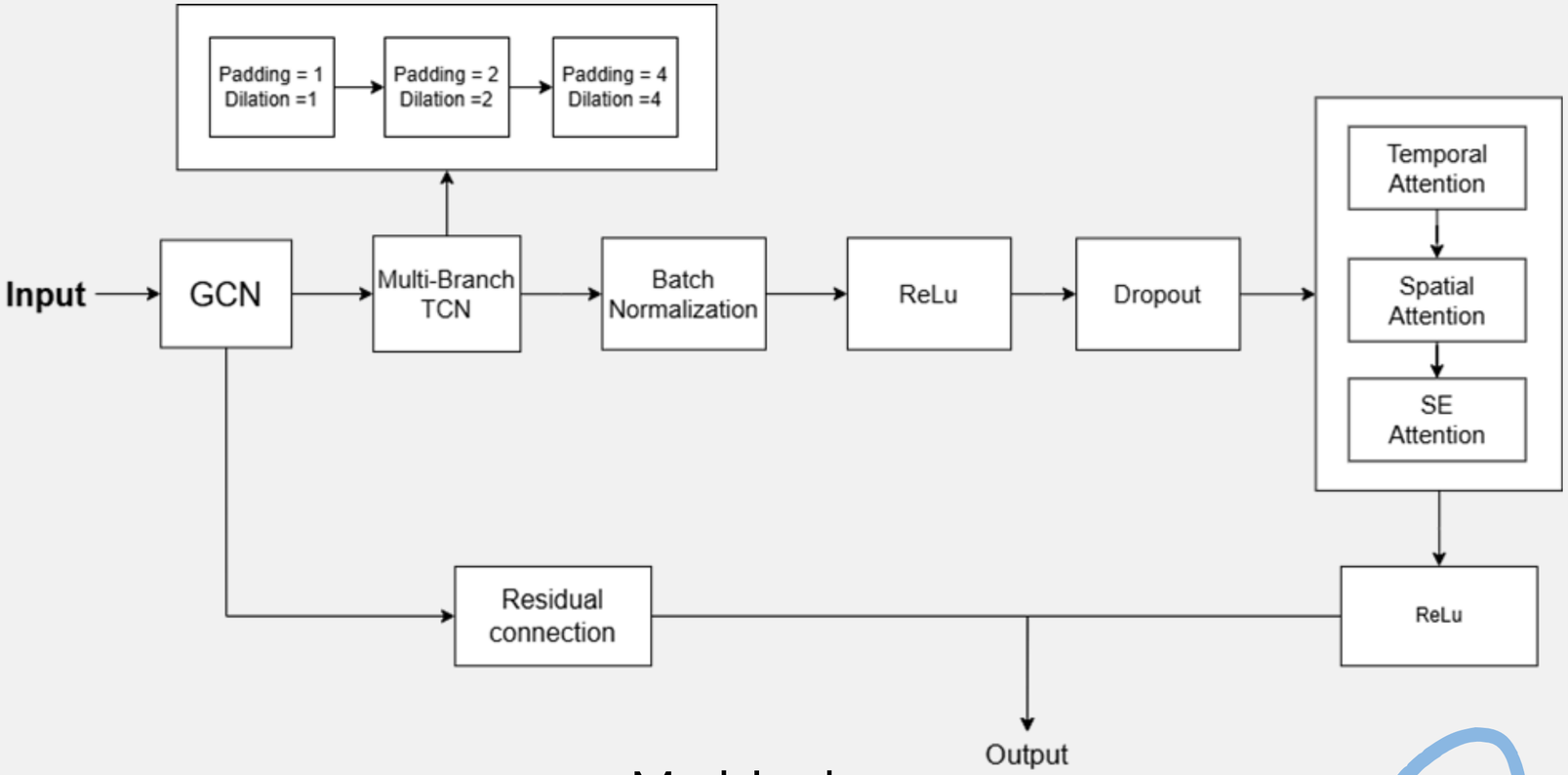



# III. Materials and Methods

## 2. Model Architecture



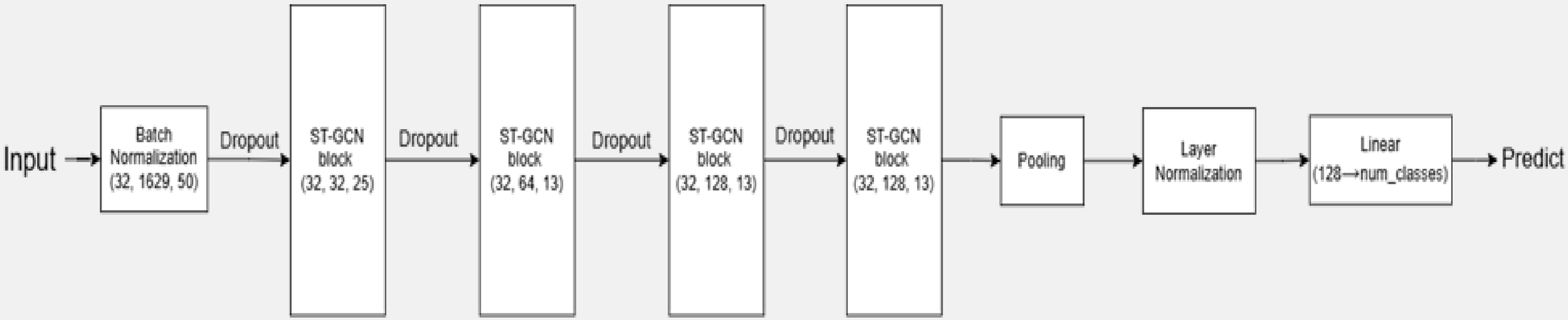

Original blocks




My blocks

# III. Materials and Methods

## 2. Model Architecture



# IV. Results and Discussion

01. Model Configuration and Training

02. Experimental Result and Discussion

03. Model Limitations

# IV. Results and Discussion

## 1. Model Configuration and Training

| Parameters | Value |
|---|---|
| Epoch | 200 |
| Optimizer | Adam |
| $\beta 1$ | 0.9 |
| $\beta 2$ | 0.99 |
| Batch size | 32 |
| Weight Decay | 0.05 |

| Parameters | Value |
|---|---|
| Initial Learining rate | 1e-3 |
| Minimum Learining rate | 1e-6 |
| Scheduler | ReduceLROnPlateau |
| Factor | 0.05 |
| Scheduler epoch | 15 |
| Total of parameters | 343714 |

# IV. Results and Discussion

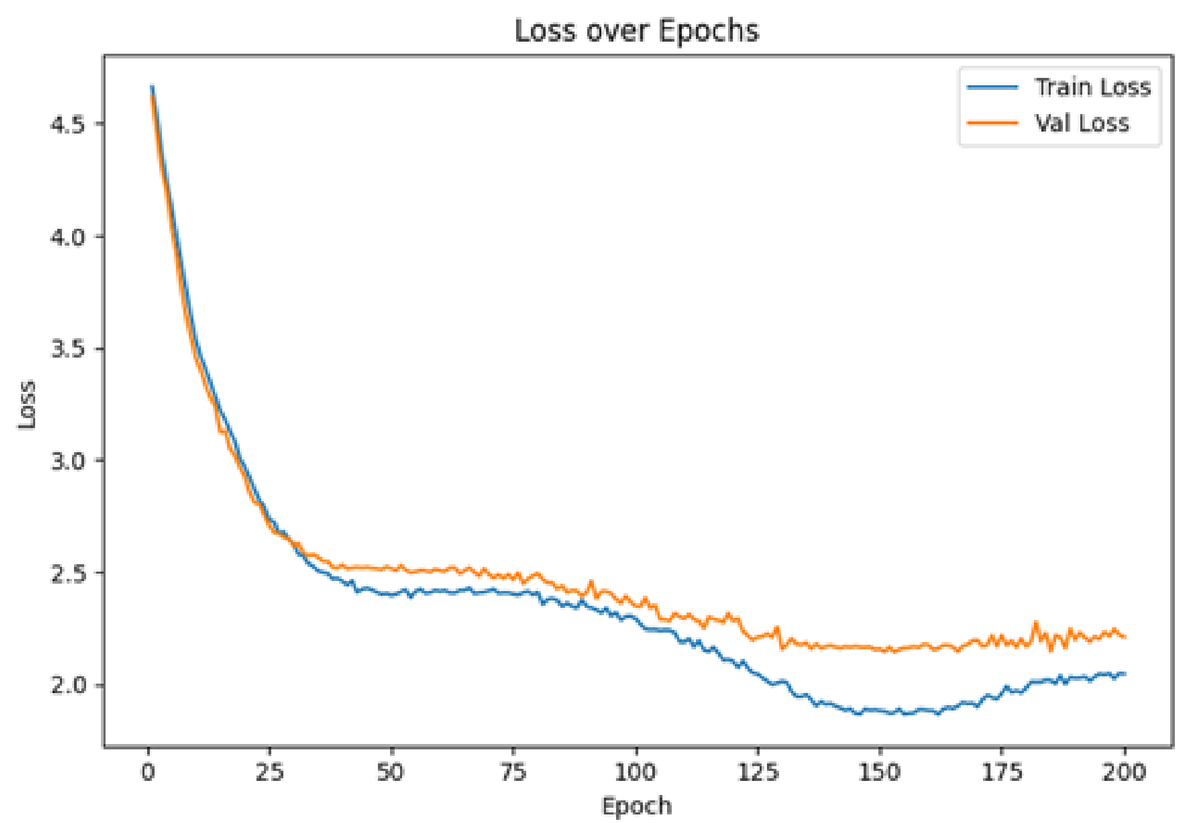
## 2. Experimental Result and Discussion

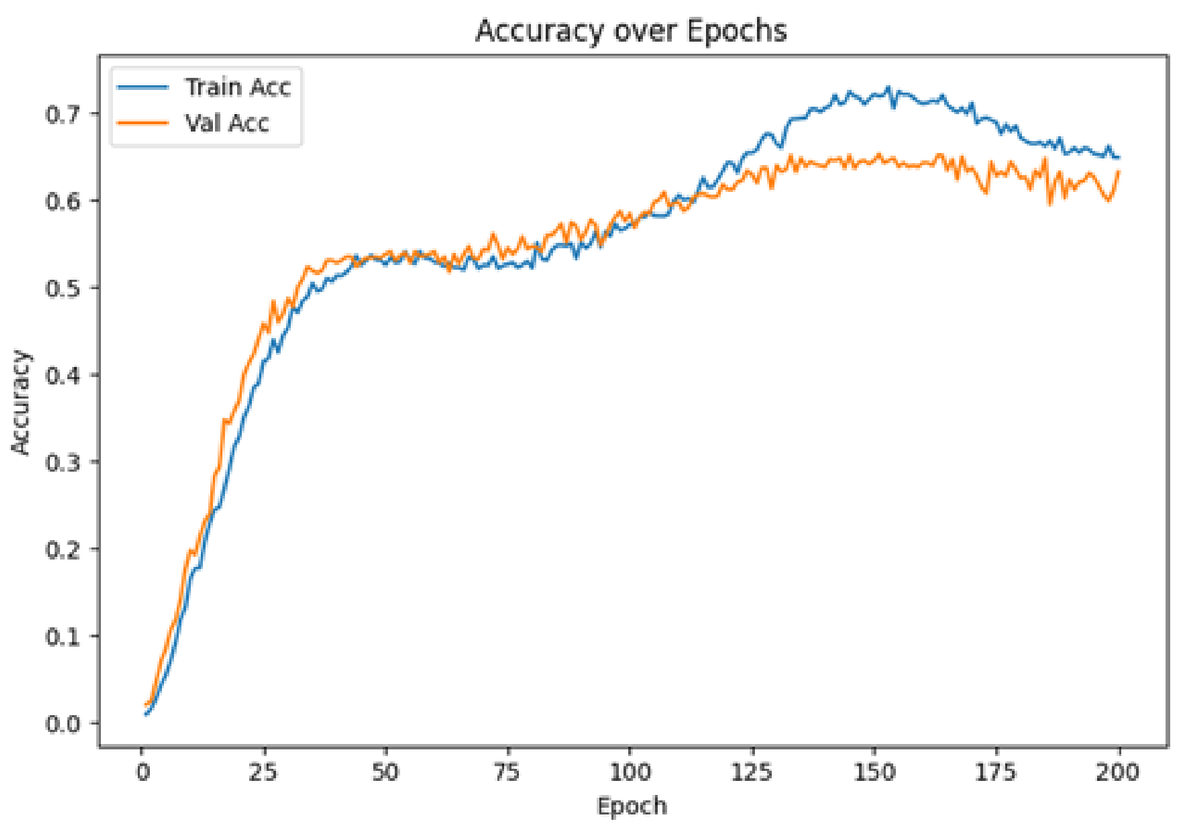
| Model | Dataset | Top-1 Accuracy | Top-5 Accuracy |
|---|---|---|---|
| Pose-GRU | WLASL100 | 46.51 | 76.74 |
| Pose-TGCN | WLASL100 | 55.43 | 78.68 |
| I3D | WLASL100 | 65.89 | 84.11 |
| Mine | WLASL100 | 65.20 | 85.29 |
| Mine | Custom | 74.49 | 90.37 |

# IV. Results and Discussion

## 2. Experimental Result and Discussion

**WLASL100**

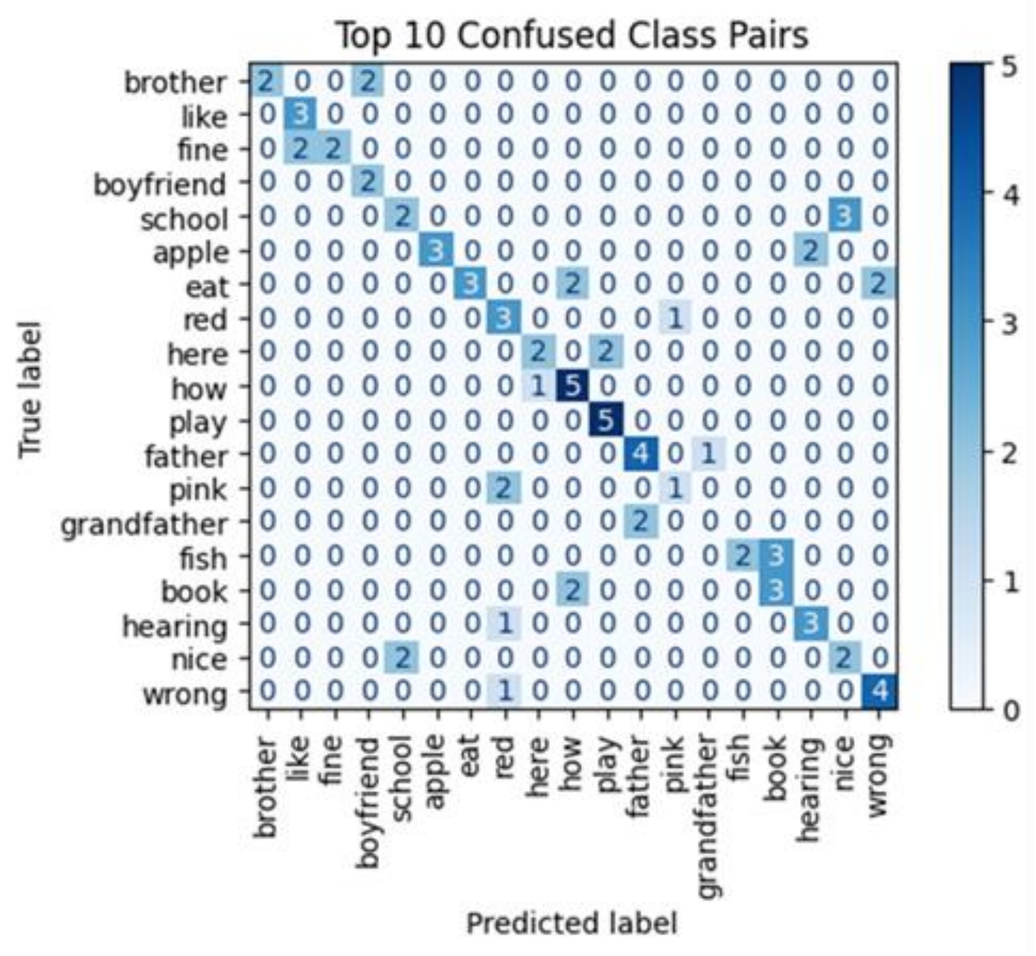


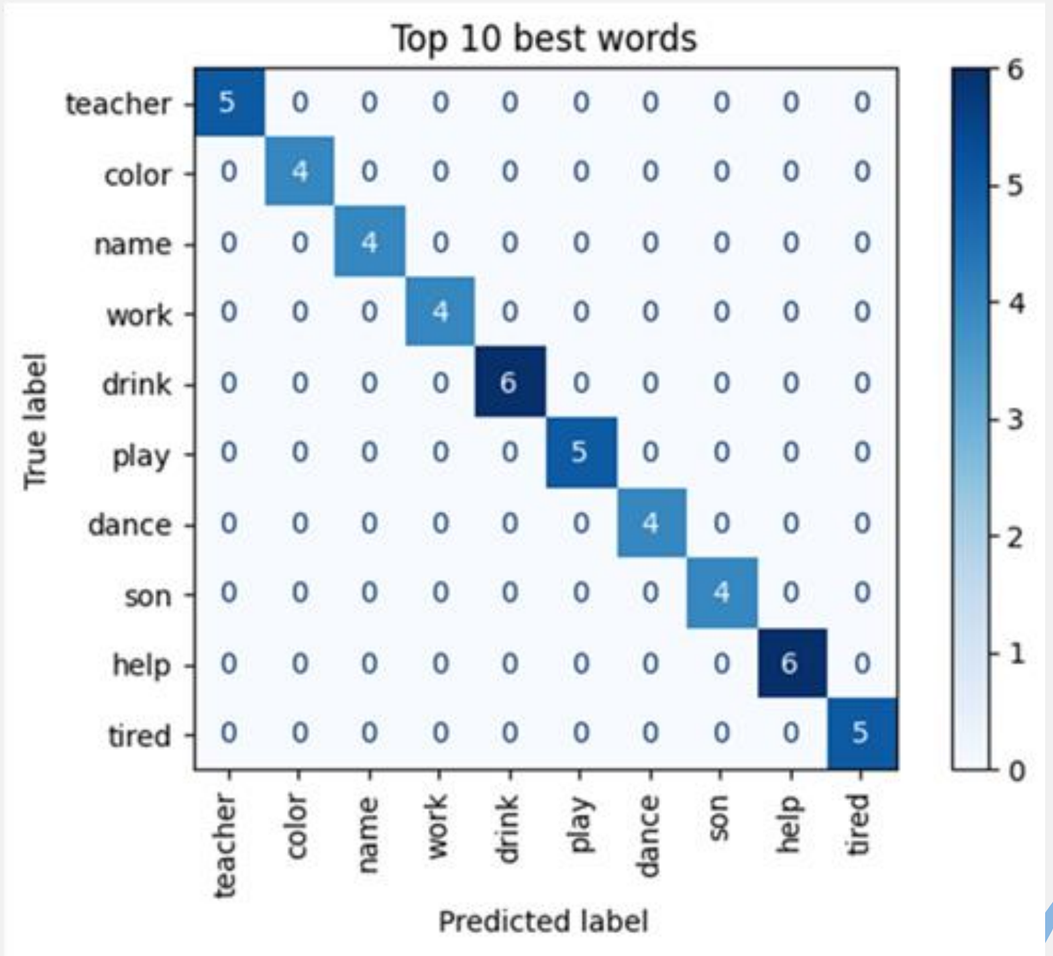



# IV. Results and Discussion

## 2. Experimental Result and Discussion

**WLASL100**

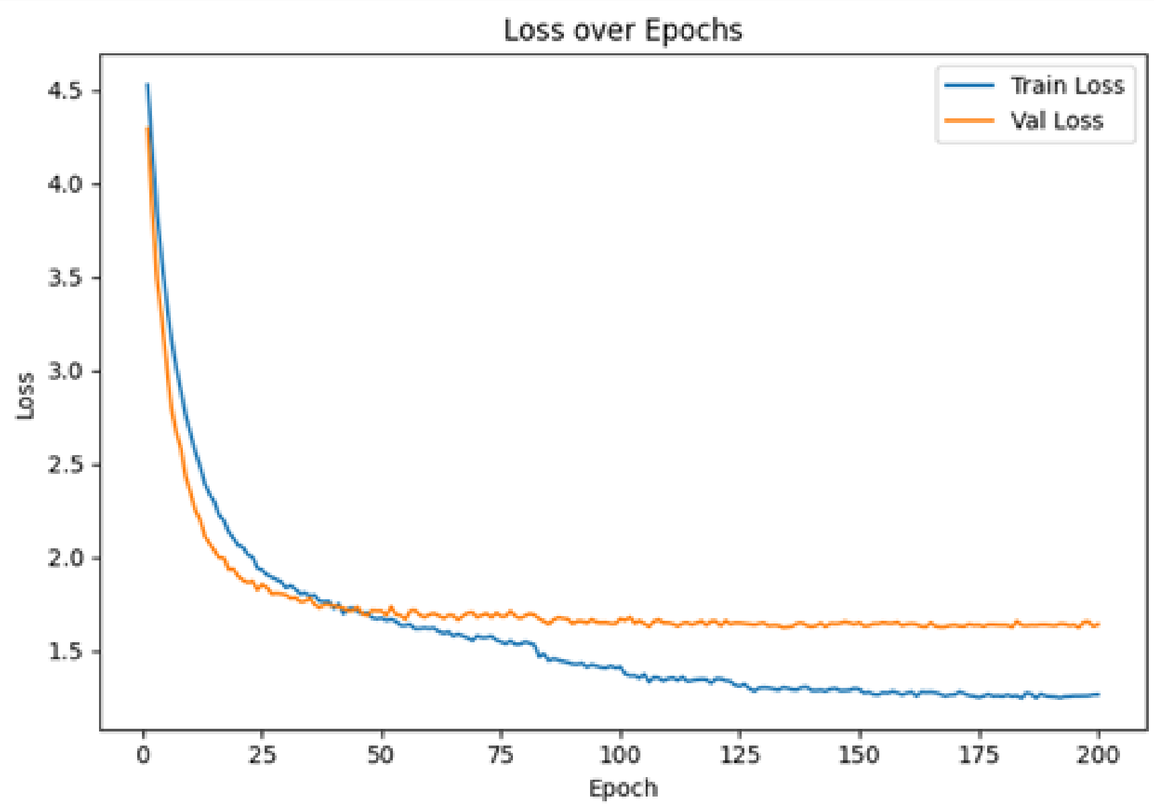


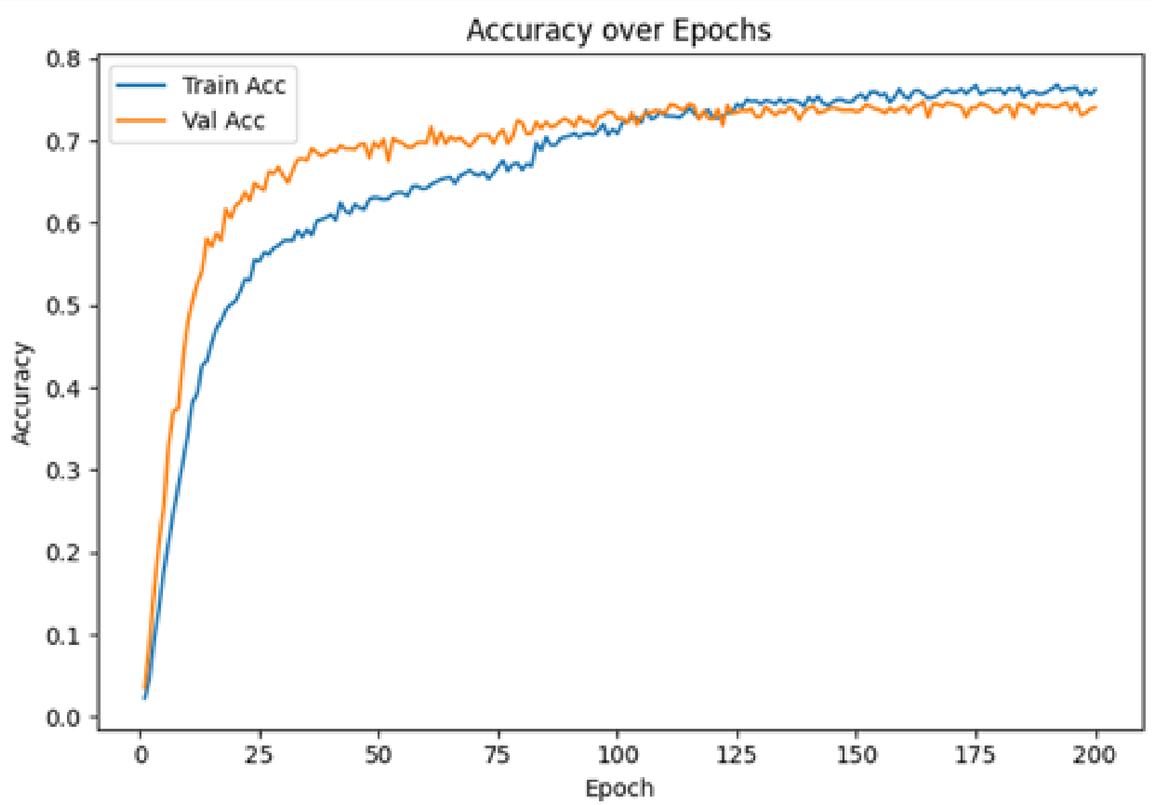



# IV. Results and Discussion

## 2. Experimental Result and Discussion

**Custom Dataset**

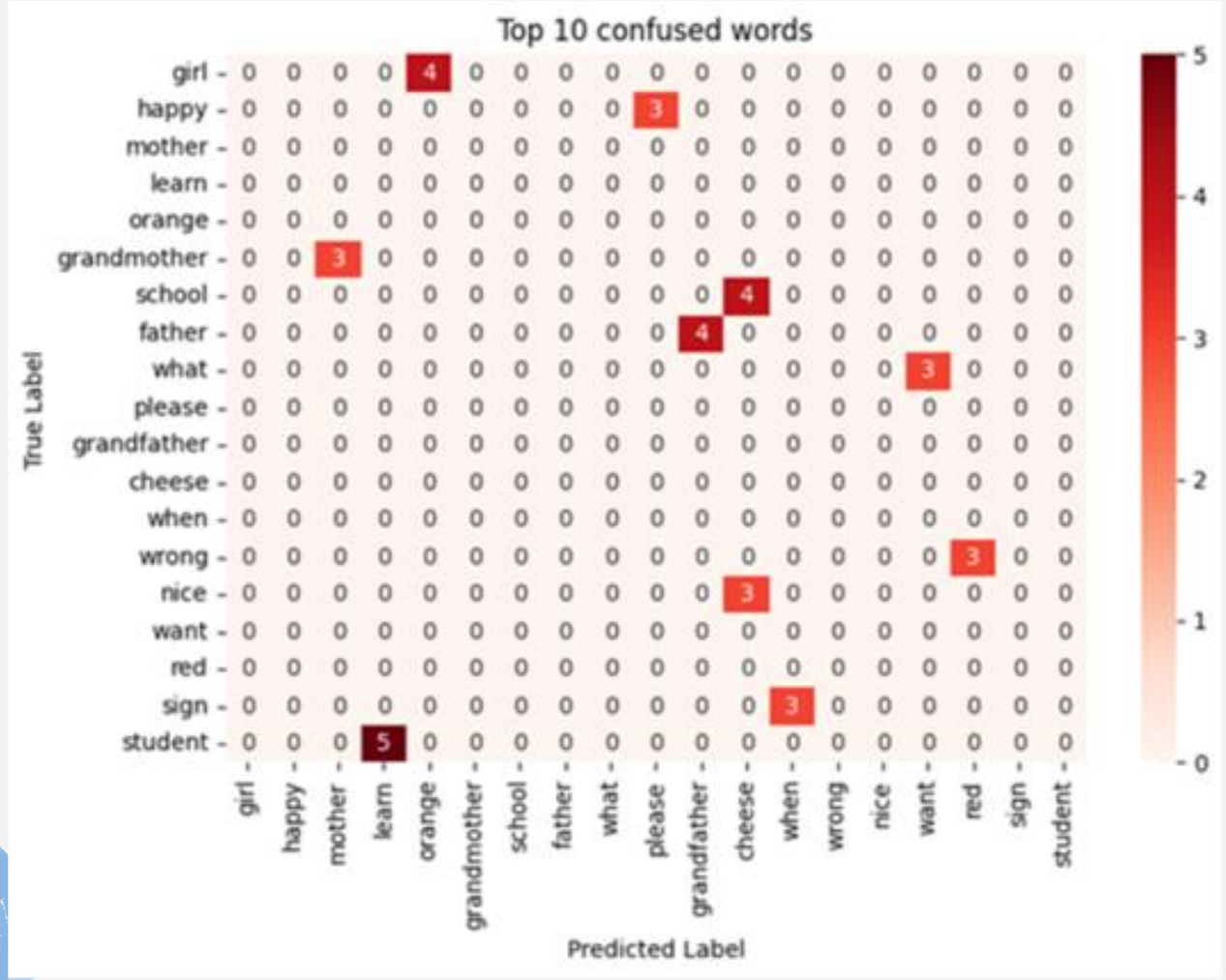


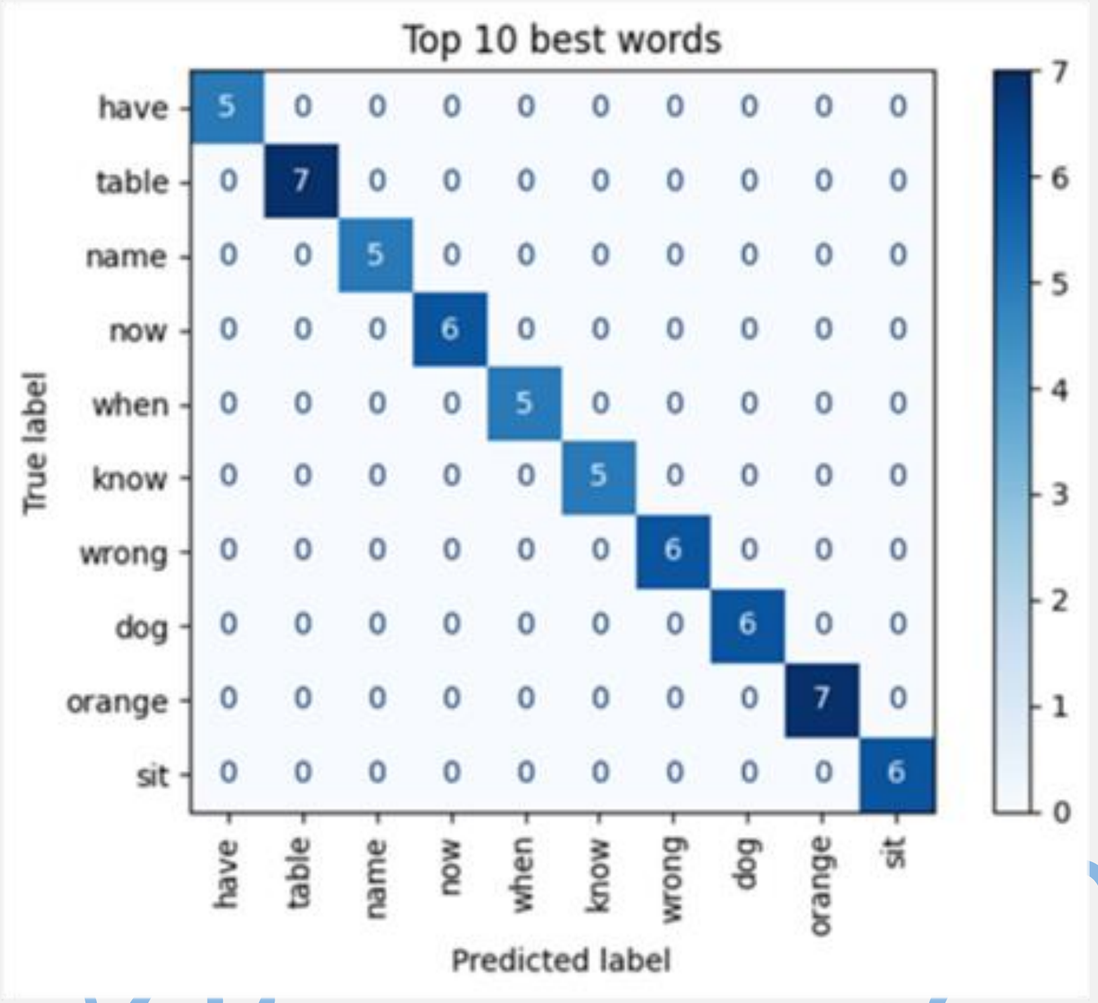



# IV. Results and Discussion

## 2. Experimental Result and Discussion

**Custom Dataset**



# IV. Results and Discussion

## 2. Experimental Result and Discussion

| | Dũng (22BI13103) | Me | Duy Anh (22BI13017) |
|---|---|---|---|
| **Dataset** | WLASL100 | WLASL100, custom | WLASL2000, MSASL1000 |
| **Preprocess** | • Frame Extraction to 64 frames per video<br>• Keypoints per frames per video (Mediapipe Framework)<br>• Resize frames to 224x224 | • Frame Extraction to 50 frames per video<br>• Resize frames to 224x224<br>• Extract keypoints full body by Mediapipe | RTMPose (full body) |
| **Model** | • S3D for video feature extraction<br>• StepNet for keypoint features<br>• Early fuse | Many of ST-GCN for learning keypoint features | • Pretrain Tranformers and Mask Modeling to get weights<br>• Integrate on Unisign |

# IV. Results and Discussion

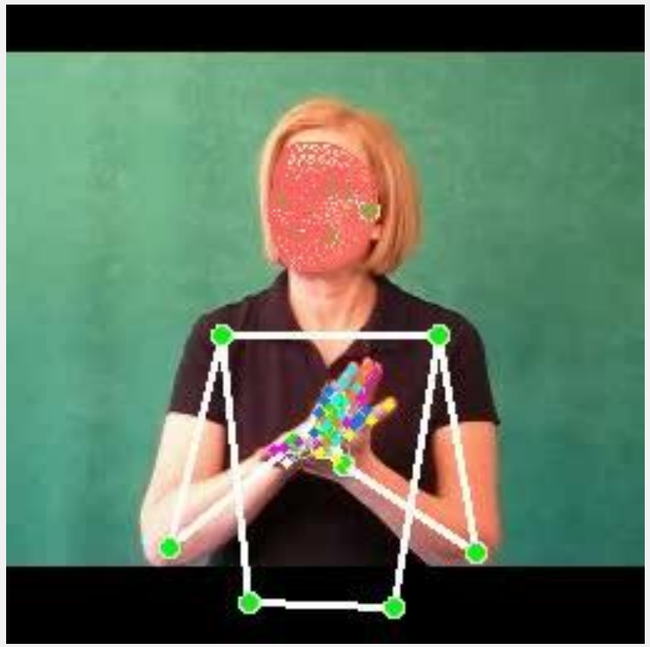
## 2. Experimental Result and Discussion

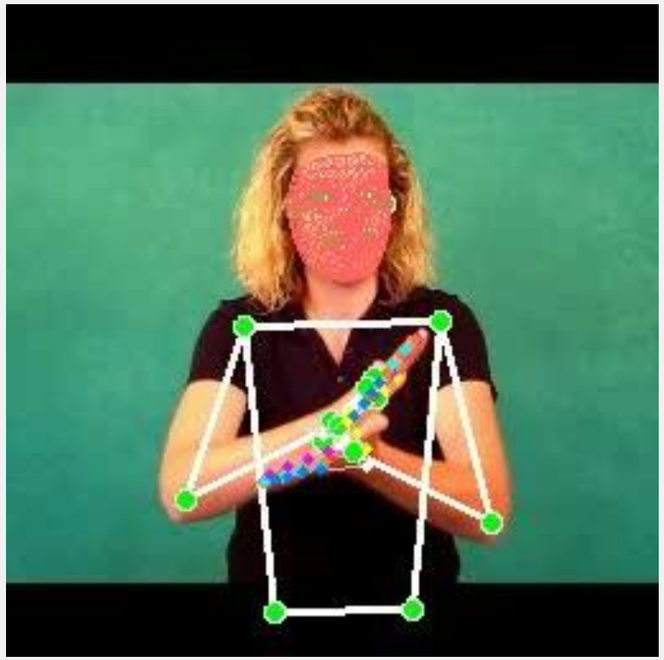
| | Dũng (22BI13103) | Me | Duy Anh (22BI13017) |
|---|---|---|---|
| **Total Parameters(M)** | 42.3 | 0.34 | 317.8 |
| **Execute Time** | 2.5h | 3h | 6h |
| **Result(Top 1 Acc)** | 92.49% | 65.20% | 58% |

# IV. Results and Discussion

## 3. Model Limitations

- Easy to confuse symbols with similar movements





Example of 2 word make most confused: ”School” (left) and “Paper”(right)

# V. Conclusion and Future Work

01. Conclusion

02. Future Work

# V. Conclusion and Future Work

## Conclusion

- Successful implementation of ST-GCN model using keypoints data from video
- The model learns well the characteristics of each sequence of signed movements
- Compatible on both CPU and GPU
- Fast training time, not too large number of parameters → suitable for common devices

## Future Work

- Further improve on other datasets to increase model evaluation
- Optimize architecture, fine-tune weights, better exploit time series
- Improve accuracy, increase practical application, expand future research

# Demo

# Thank you for listening!