# Baseline Deep Learning Model for Khmer Sign Language Recognition with a Small Dataset

**Sokleap Som** **Rottana Ly** **Nab Mat**
Cambodia Academy of Digital Technology, Phnom Penh, Cambodia
sokleap.som@cadt.edu.kh

## Abstract

People with hearing impairments in Cambodia often face challenges in daily communication and learning, especially when others cannot understand Khmer Sign Language. These difficulties can limit social interaction and access to education. To address this problem, this research proposes a Khmer Sign Language recognition approach using deep learning to support better communication and self-learning tools for the Deaf community. A dataset containing 100 sign classes was collected, with 20 fully annotated classes used for training, each including 28 to 31 samples recorded in varied environments at the National Institute for Special Educat ion (NISE). The data were processed into 30 fps and 1920 × 1080 resolution to ensure temporal smoothness and clear motion capture. Two models based on 3D ResNet (R3D-18) were trained and compared: one using raw RGB frames and another using keypoints extracted by MediaPipe. The RGB-based model achieved 88.35% precision, 84.09% recall, 83.72% F1-score, and 84.09% accuracy. The keypoint-based model achieved 95.21% precision, 92.05% recall, 92.56% F1-score, and 92.05% accuracy, showing that focusing on body and hand landmarks improves robustness and generalization on small datasets. This research provides a foundation for Khmer Sign Language recognition using limited data. Future work will expand the dataset, explore more classes, and improve inference for real-time applications while applying Early-Stopping to prevent overfitting.

***Keywords:*** *Khmer Sign Language Recognition, Deep Learning, 3D ResNet-18, MediaPipe, Keypoint-based learning*

## 1 Introduction

The Khmer Sign Language is the main mode of communication for people who are deaf or hard of hearing in Cambodia. However, limited technological support makes it difficult for deaf people to interact with hearing people unfamiliar with sign language, leading to social isolation and reduced educational opportunities. Although organizations such as Krousar Thmey and the Deaf Development Programme have provided support, accessible digital tools remain scarce.

Recent advances in deep learning have led to significant progress in Sign Language Recognition (SLR), especially through 3D Convolutional Neural Networks (3D CNNs) capable of learning spatio-temporal features from video sequences [1], [2]. Advanced architectures such as SlowFast [3] and X3D [4] improve motion understanding, while Transformer-based models such as ViViT [5] and Sign Language Transformers [6] achieve strong results through attention mechanisms. However, these models are based on large-scale datasets, which are unavailable for underrepresented languages such as Khmer Sign Language [7], [8].

To address data scarcity, keypoint-based methods use pose estimation frameworks such as MediaPipe to extract skeletal landmarks that capture sign movement while filtering out background noise [9], [10]. Combining MediaPipe with CNN-LSTM or Transformer models has achieved promising results on small datasets [11], [12], [13]. Building on these insights, this research presents a baseline R3D-18 model trained on RGB and keypoint data to establish a foundational benchmark for Khmer Sign Language recognition.

This research contributes by creating a new Khmer Sign Language dataset that provides one of the first structured collections for this language. It also establishes a baseline deep learning model using both RGB and keypoint input, offering a reference for future studies. The results and methods can help future researchers build larger models and improve technologies that support the Deaf community in Cambodia.

This paper is organized as follows. Section 2 reviews existing research and related work on Sign Language Recognition and Keypoint-Based approaches. Section 3 presents the proposed methodology, including dataset collection, preprocessing, and model architecture of R3D-18. Section 4 describes the experimental setup and evaluation metrics used in this research. Section 5 discusses the experimental results and performance comparisons between the RGB and keypoint-based models. Finally, Section 6 concludes the article and outlines directions for future work.

## 2 Related Work

Research in sign language recognition has advanced rapidly with deep learning models that capture spatio-temporal patterns from video data. Early studies using 3D CNNs such as 3D ResNet and R(2+1)D demonstrated strong capabilities to model motion dynamics across frames [1], [2], [14]. Later architectures like SlowFast [3] and X3D [4] enhanced temporal efficiency and accuracy, becoming standard baselines for video classification tasks, including SLR.

Transformer-based models such as ViViT [5] and Sign Language Transformers [6] further improved recognition through long-range attention, but their performance depends on extensive labeled data. This limits their use in low-resource settings, where annotated datasets are scarce [7]–[15]. Consequently, baseline models trained on small datasets remain essential to establish feasible starting points for further research and deployment.

To overcome small-data constraints, keypoint-based SLR has emerged as an effective approach, where skeletal landmarks replace full RGB frames to emphasize motion cues and reduce noise. MediaPipe has gained popularity for its lightweight and real-time extraction of hand, face, and body landmarks [10]. Recent studies integrating MediaPipe with CNN-LSTM [12], Transformer [11], and hybrid networks [16], [13] show that models driven by landmarks generalize better and train faster with limited data, aligning with the goal of this research.

## 3 Methodology

This section describes three main components: data collection, data annotation, and model selection. First, the data collection process outlines how Khmer Sign Language videos were recorded and organized. The data annotation process explains how gestures were labeled and validated to ensure linguistic accuracy. Finally, the model selection subsection discusses the 3D ResNet-18 (R3D-18) [1] architecture used for RGB and keypoint-based inputs [17], [12], [10], including its configuration and training parameters.

### 3.1 Data Collection

A Khmer Sign Language dataset was a task that was cooperated between Cambodia Academy of Digital Technology (CADT) and the National Institute of Special Education (NISE) in Phnom Penh to record the data as videos for this research. The complete dataset contains 100 classes, but only 20 fully annotated classes were used to ensure consistency and reliable labels. The remaining 80 were excluded due to missing annotations, as unlabeled data could introduce ambiguity and hinder supervised learning.

The videos were recorded in 1920 × 1080 resolution as Full HD at 30 fps with the digital camera. Although sometimes NISE helps record videos using the smartphone which might have slightly different resolution but we used the tool that helps all videos resolution to be consistency. This resolution was chosen to ensure clear visualization of finger and hand movements while keeping a balance between detail and computational cost. A higher frame rate (30 fps) helps capture smooth temporal motion without excessive storage usage. Although both cameras and smartphones were used, quality differences were minimal after preprocessing since all videos were standardized in resolution and frame rate.

The 20 annotated signs span four categories: *Locations and Actions* (“Where”, “When”, “Market”, “Buy”, “Location”), *Directions* (“Left”, “Right”, “North”, “South”), *Objects* (“Pen”, “Blue Pen”, “Red Pen”, “Pencil”, “Book”, “Ruler”, “Eraser”), and *People* (“Teacher”, “Director”, “Female Director”, “Deputy Director”). Six deaf participants recorded the signs using cameras and smartphones at 1920 × 1080 resolution and 30 fps in varied environments. A sample of the dataset is illustrated in Figure 1, and a summary is shown in Table 1.

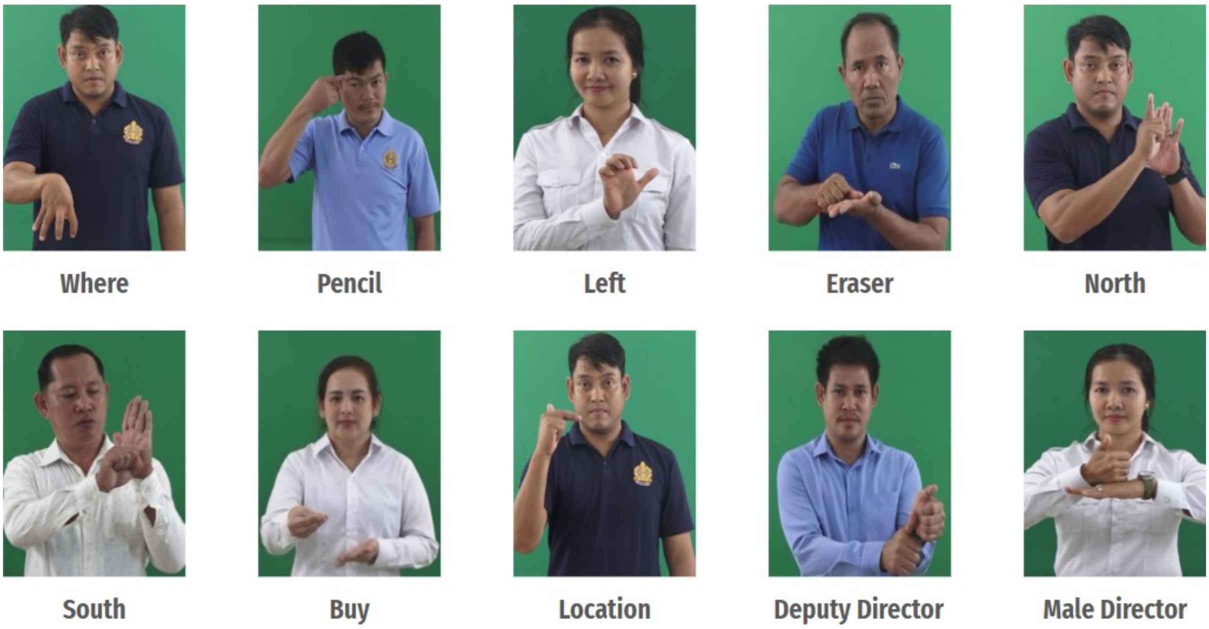


Figure 1. Sample of the proposed Khmer Sign Language dataset.

Table 1. Dataset Overview

| | |
|---|---|
| **Total Classes:** | 100 |
| **Classes Used:** | 20 Annotated |
| **Videos Used:** | 591 Videos |
| **Participants:** | 6 Deaf Participants |
| **Location:** | NISE |
| **Frame Rate:** | 30 fps |
| **Resolution:** | 1920 × 1080 |
| **Record Device:** | Camera and Smartphone |

### 3.2 Data Annotation

Data Annotation was conducted manually by trained experts, with additional validation steps to ensure the highest level of linguistic accuracy and consistency. Each video in the dataset was carefully reviewed and labeled according to its corresponding sign class, ensuring that it contained a single complete gesture without ambiguity or overlap. By including only verified and fully validated samples, we preserved the integrity and uniformity of the dataset, which is particularly crucial when evaluating the performance of the baseline model on small datasets. This meticulous process not only guarantees reliable annotations but also provides a solid foundation for reproducible research and meaningful comparisons across different computational models.

### 3.3 3D ResNet-18 (R3D-18)

The R3D-18 model is an extended version of ResNet-18 that replaces 2D convolutions with 3D ones, allowing it to learn spatial and temporal features of video frames [1]. This ability is important for Sign Language Recognition, where hand motion and movement over time carry meaning. Compared to larger models such as I3D [2], SlowFast [3], or Transformer-based networks [5], R3D-18 provides a good balance between accuracy and efficiency, making it suitable for small datasets.

With this lightweight model, it is powerful enough to learn motion features without overfitting the 591-video dataset around 27–31 samples per class. Its 3D filters capture motion patterns across 30-frame sequences while keeping the spatial structure clear. Using pre-trained R3D-18 weights from large video datasets also improves generalization through transfer learning.

For the keypoint-based version, the MediaPipe framework was used to extract 543 keypoints from the body, hands, and face, giving the model simplified inputs that are less sensitive to lighting or background changes. Previous studies have shown that such landmark-based features work better in small datasets and improve robustness.

Both the RGB and keypoint models were

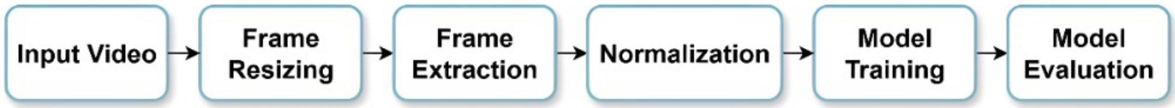


Figure 2. Preprocessing workflow for RGB input.

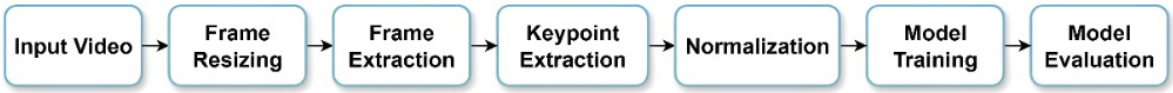


Figure 3. Preprocessing workflow for MediaPipe keypoint input.

trained with the same settings: Adam optimizer (learning rate $1e^{-4}$), batch size 16, and 5 epochs. Five epochs were chosen because overfitting appeared beyond that point, giving a good trade-off between accuracy and training time. The RGB model achieved 84.09% accuracy, while the keypoint model reached 92.05%, demonstrating that focusing on landmarks gives better results on small data. After applying early stopping and model checkpointing (Figure 6), the keypoint-based model improved further to 94.32% accuracy, 95.43% precision, and 94.28% F1-score.

During training, validation loss was monitored each epoch. If it did not improve for 5 epochs (*EARLY_STOPPING_PATIENCE* = 5), training was stopped automatically to prevent overfitting. Whenever the validation loss improved, the model weights (*state_dict*) were saved in *BEST_MODEL_PATH*. This process also works with *nn.DataParallel* models. Only the weights were saved, but the optimizer and epoch number can also be stored if training needs to be resumed later. The final performance after applying these methods is shown in Table 4.

## 4 Experimental Setup

### 4.1 Data Preprocessing

Each video was resized to 224 × 224 pixels and normalized to [0, 1]. Two modalities were explored: RGB, preserving the appearance features Figure 2; and MediaPipe, extracting 2D keypoint landmarks for the hands, body, and face Figure 3. The natural variation in the recording conditions provided sufficient diversity for model training.

### 4.2 Implementation and Evaluation Setup

The experiments were conducted in PyTorch During early development, models were trained for 5 epoch using cross-entropy loss, Adam Optimizer ($1e^{-4}$), and a batch size of 16. The dataset was split into 70%, 15%, and 15% for training, validation, and testing under a signer-dependent scheme. This ratio was chosen to maintain a balanced trade-off between learning and generalization, providing enough data for model fitting while keeping unseen samples for fair evaluation. The RGB inputs were normalized pixel frames, while the keypoint inputs were coordinate-normalized landmarks.

When applying early stopping and model checkpointing during extended 14 epoch experiments, training was conducted in Kaggle's GPU environment using *nn.DataParallel* for multi-GPU processing. Note that *nn.DataParallel* operates in a single-node, single-process configuration and may incur CPU-to-GPU scatter overhead. When scaling to multiple GPUs, it is advisable to proportionally increase the batch size and save both the optimizer state and the current epoch to allow resuming interrupted training. The effect of applying early-stopping and model checkpointing during 14-epoch runs is to ensure the robustness of the model, and whenever the validate does not improve, save the model and select the best model at the best epoch in Figure 6.

## 5 Results and Discussion

The results are summarized in Table 2. The RGB-based R3D-18 achieved an accuracy of 84.09%, while the keypoint-based model reached 92.05%. The keypoint model also demonstrated higher precision and F1-score, indicating stronger consistency between classes.

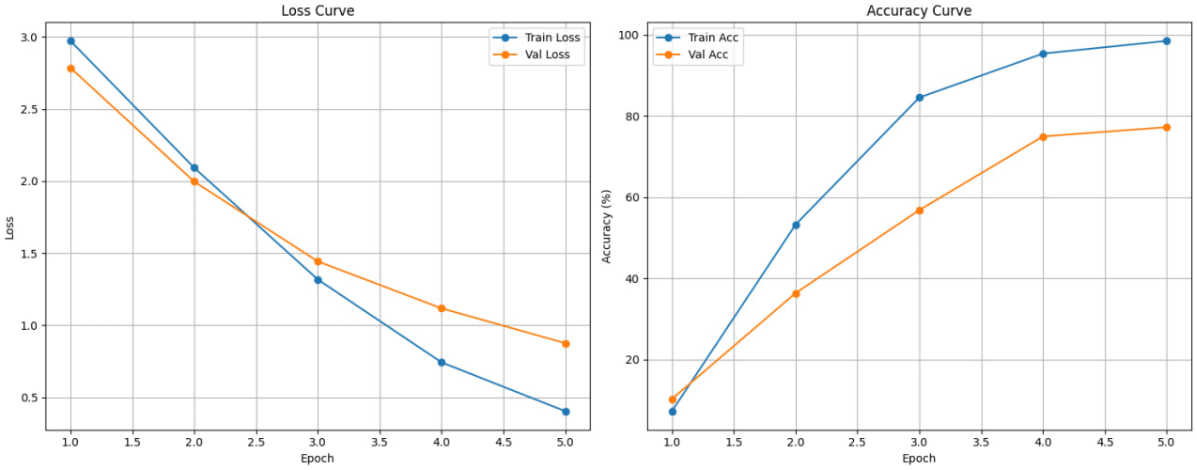
To further enhance performance, early stopping and model checkpointing were applied dur-

Table 2. Performance Comparison Between RGB and MediaPipe Models

| Architecture | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| R3D-18 (RGB) | 88.35% | 84.09% | 83.72% | 84.09% |
| **R3D-18 (Keypoint)** | **95.21%** | **92.05%** | **92.56%** | **92.05%** |

Table 3. Sample of Most Confused Class Pairs

| True Label | Predicted Label | Count |
|---|---|---|
| Ruler | Eraser | 2 |
| Director | Female Director | 2 |



**(a)** **(b)**

Figure 4. Training and validation curves for the Keypoint-based model. (**a**) Loss Curve: Consistent, steep reduction in both Training and Validation Loss, confirming successful convergence. (**b**) Accuracy Curve: Substantial increase in both Training and Validation Accuracy, demonstrating significant predictive improvement.

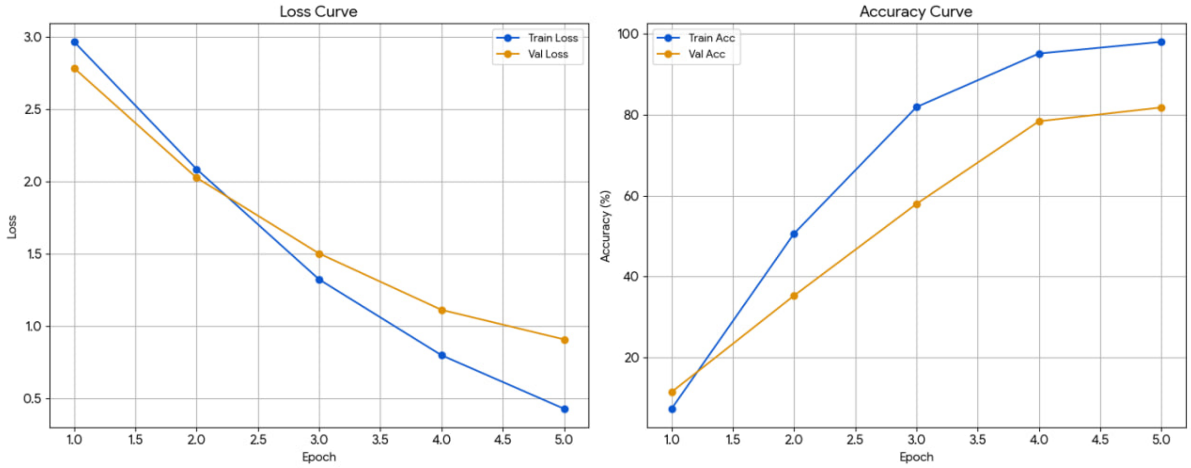
Table 4. Early-Stopping and Model Checkpointing Performance - 14 Epoch

| Architecture | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **R3D-18 (Keypoint)** | **95.43%** | **94.32%** | **94.28%** | **94.32%** |

ing a 14-epoch training process. This approach helped prevent overfitting and automatically saved the best-performing model based on validation loss. As shown in Table 4, the keypoint-based model improved to 94.32% accuracy, with 95.43% precision and 94.28% F1-score. The training curves in Figure 6 also demonstrate smoother convergence and stable validation trends, confirming that early stopping effectively reduced unnecessary epochs while preserving generalization.

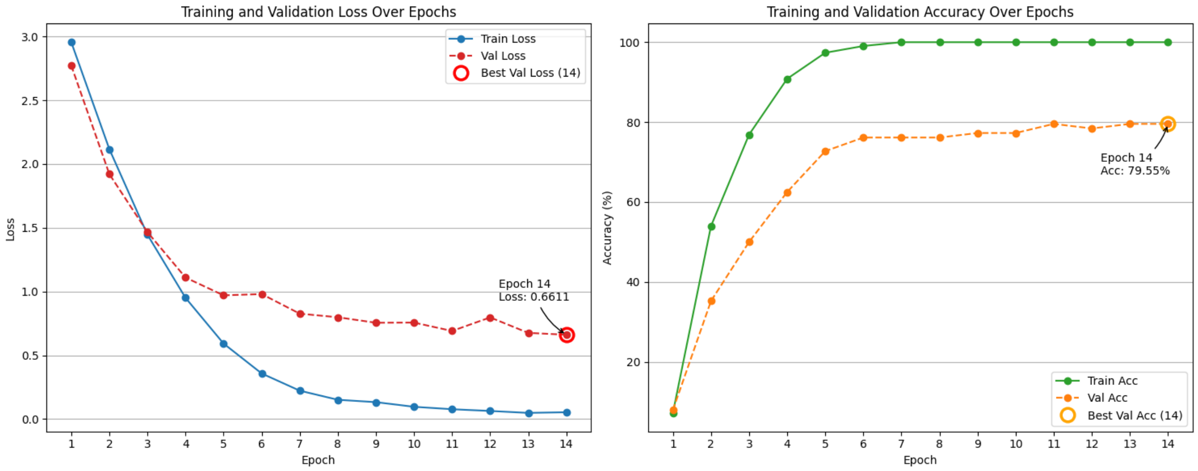
Some confusion remained, particularly between semantically and visually similar classes (Table 3). For example, “Director” and “Female Director” have nearly identical initial movements but differ subtly near the end of the gesture, making it difficult for the model to distinguish between them. Figure 9 provides a qualitative comparison showing similar hand trajectories that caused this confusion.

These findings confirm that, while RGB provides richer appearance information, it also introduces noise that hinders small-data learning. Keypoint-based features emphasize essen-



**(a)** **(b)**

Figure 5. Training and validation curves for the RGB-based model. (**a**) Loss Curve: Consistent, strong reduction in both Training and Validation Loss. (**b**) Accuracy Curve: Rapid, continuous increase in both Training and Validation Accuracy.



**(a)** **(b)**

Figure 6. Applied early stopping for the Keypoint-based model. (**a**) Loss Curve: Optimal checkpoint selected at Epoch 14 (Best Val Loss: 0.6611). (**b**) Accuracy Curve: Peak generalization reached at Epoch 14 (Best Val Acc: 79.55%).

tial motion cues and yield higher accuracy and generalization for Khmer Sign Language recognition.

## 6 Conclusion

To conclude, a baseline deep learning framework for Khmer Sign Language recognition was developed using a small annotated dataset. Two variants of the R3D-18 model were evaluated: RGB and MediaPipe keypoint-based. The keypoint model outperformed the RGB model, achieving an accuracy of 92.05%, as shown in Figures (4-7), while the RGB model reached an accuracy of 84.09% in Figures (5-8). This performance gap can be explained by the fact that the RGB model is more sensitive to lighting, background variations, and visual noise, whereas the keypoint model focuses on skeletal landmarks, capturing motion and structure more effectively under limited data conditions.

This study is limited by the small size dataset and the signer-dependent setup.

These results demonstrate that landmark-based representations enhance robustness and

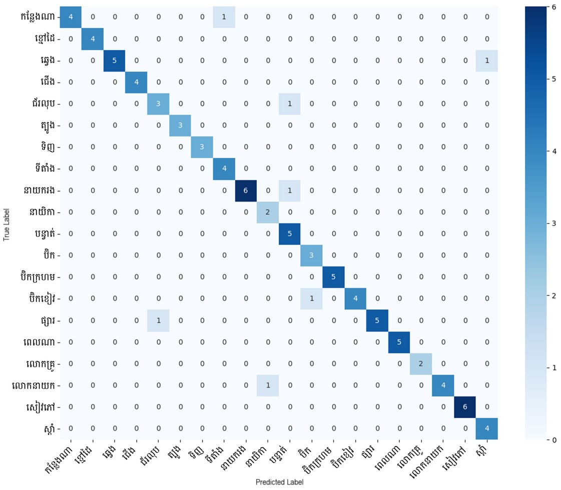


Figure 7. Confusion Matrix (Keypoint).

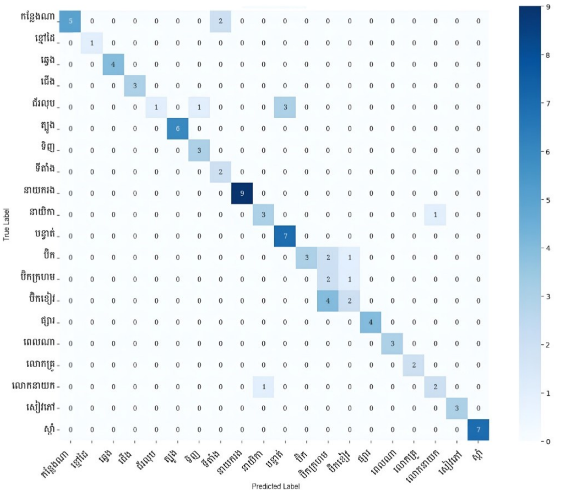


Figure 8. Confusion Matrix (RGB).

generalization for small datasets. Future work will add labels for all 100 classes, test the model with different signers, and use techniques like early stopping and saving model checkpoints with extensive testing to ensure robustness. We will also explore Transformer-based architectures and attention mechanisms over time to further improve accuracy. This research establishes a foundation for the recognition of Khmer Sign Language and contributes to the development of inclusive technologies that support communication, education, and accessibility for the Deaf community in Cambodia.

## Acknowledgment

We sincerely thank the teachers and staff of the National Institute of Special Education (NISE), Cambodia, for their valuable support in the data collection process. Their assistance in recording and verifying Khmer Sign Language videos was essential to this research, and we also appreciate all participants for their time and contribution to the dataset.

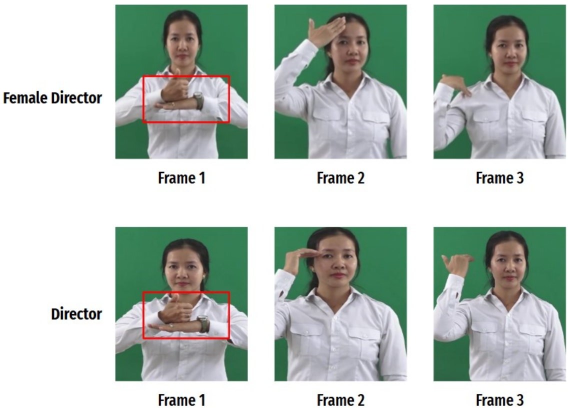


Figure 9. Similarity between "Male Director" and "Female Director" gestures, illustrating their visual resemblance.

## References

[1] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[2] J. a. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[3] 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 (ICCV)*, 2019, pp. 6202–6211.

[4] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 203–213.

[5] A. Arnab, M. Dehghani, G. Heigold, C. S. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Con-*

*ference on Computer Vision (ICCV)*, 2021, pp. 6836–6846.

[6] N. C. Camgöz, 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 (CVPR)*, 2020, pp. 10 023–10 033.

[7] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *arXiv preprint arXiv:2007.12530*, 2020. [Online]. Available: https://arxiv.org/abs/2007.12530

[8] M. Madhiarasan and P. P. Roy, "A comprehensive review of sign language recognition," *arXiv preprint arXiv:2203.02387*, 2022.

[9] O. Özdemir, I. Baytaş, and L. Akarun, "Multi-cue temporal modeling for skeleton-based sign language recognition," *Frontiers in Neuroscience*, vol. 17, 2023.

[10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.

[11] H.-H. Li and C.-C. Hsieh, "Dynamic hand gesture recognition using mediapipe and transformer," *Engineering Proceedings*, vol. 8, no. 1, pp. 1–6, 2025.

[12] A. Fitriani, M. P. Utama, and D. T. Iswanto, "Dynamic sign language recognition using mediapipe library and modified lstm method," *International Journal of Advanced Science Engineering Information Technology (IJASEIT)*, vol. 13, no. 5, pp. 19 401–19 408, 2023.

[13] M. H. Uddin, M. S. Islam, and M. R. Islam, "A deep learning-based bangla sign language recognition system using mediapipe and cnn-lstm architecture," *Sensors*, vol. 23, no. 21, pp. 1–18, 2023.

[14] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[15] N. Sarhan, S. Kammoun, and M. Hammami, "Unraveling a decade: A comprehensive survey on isolated sign language recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023, pp. 2312–2321.

[16] M. H. Ali, N. M. Noor, and H. A. Jalab, "Next-gen dynamic hand gesture recognition: Mediapipe, inception-v3 and lstm-based enhanced deep learning model," *Electronics*, vol. 13, no. 16, 2024.

[17] C. Lugaresi, J. Tang, H. Nash, C. McClanahan *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.