A Preliminary Study on Khmer Sign Language Recognition Using Neural Networks

Ponleur Veng **Nab Mat** **Kimhuoy Yann** **Vichhika Sina**
Sokleap Som **Rottana Ly**
Cambodia Academy of Digital Technology, Phnom Penh, Cambodia
ponleur.veng@cadt.edu.kh

Abstract

People with hearing impairments often face challenges in both social interactions and their language development, especially when they communicate with individuals who have little or no knowledge about sign language. Although sign language research has rapidly progressed we still need to working on for different linguistic across countries. This paper provides the preliminary study of Khmer Sign Language (KSL) recognition with different neural networks approaches. We also compare the performance where we train with RGB video frame and pose keypoint extraction to find accuracy yet computing efficient approach. We collected a dataset consisting of 6 deaf participants, featuring 20 signs across 580 videos, with each video recorded at a frame rate of 30 frames per second. The dataset was collected using digital cameras and smartphones in different environments to ensure the model's robustness across different devices and conditions. We trained several architectures and show that SlowFast achieved the best performance with 92.81% accuracy, 92.50% F1, 92.81% recall, and 93.57% precision. The keypoint pipeline with R3D-18 also performed competitively (92.05% accuracy, 92.56% F1, 92.05% recall, 95.21% precision), suggesting a promising trade-off for scenarios with tighter computed budgets. For this research has shown that SlowFast with RGB frame provide higher accuracy while pose keypoint show the better scalability for training. For the future work will expand the dataset and class number improve accuracy and generalizability, especially in real-world scenarios.

Keywords: *Sign Language Recognition, Transformer Model, Video Vision Transformer Model, SlowFast, R3D-18, LSTM, Bi-LSTM, GRU, Velocity features, Low-resource dataset*

1 Introduction

Sign language serves as an essential means of communication for people who are deaf or hard of hearing, enabling interaction and social integration both within their communities and with hearing individuals who understand sign language. Globally, it is estimated that more than 70 million people experience deafness, with more than 80% residing in developing countries, where access to resources and support is often limited [1, 2]. In 2020, approximately 1,700 newborns were born deaf, further highlighting the global prevalence of hearing loss [3]. In Cambodia, of a total population of 16 million people, there are 1.5 million deaf and hearing impaired. Approximately 3.5% of 1.5 million deaf people are profoundly deaf [4]. These figures highlight the critical importance of research aimed at improving communication and accessibility for deaf and hard-of-hearing individuals worldwide.

However, in Cambodia, support for the deaf for accessibility and communication in education and employment is limited. Although Krousar Thmey provides education for hearing-impaired children and the Maryknoll Deaf Development Program offers training for adults, most deaf individuals lack access to essential services [5]. In addition, another main challenge is when students start to learn new words at school and their family members and relatives cannot understand that new sign language vocabulary. In 2019, children with disabilities were three times less likely to attend school than other children [6], highlighting the challenges facing the deaf community and the urgent need for technologies to improve communication and a better education approach.

We aim to contribute to the community by providing (i) better communication and support for self-learning, we propose sign language

Table 1. Dataset Overview

Attribute	Value
Number of Videos	580
Participants	6 deaf participants
Location	NISE
Number of Signs	20
Frame Rate (fps)	30
Resolution (pixels)	1920×1080
Recording Devices	Digital cameras and smartphones

recognition by providing accurate recognition, which students can use to acquire new sign language independently to help this community with a better education experience and social integration, and (ii) compare different techniques for effective training and results by choosing two different approaches. Model training with RGB frame video with several architectures such as the SlowFast network [7], Channel Separated Convolutional Networks (CSN) [8], ViViT [9], and another approach pose keypoint extraction with mediapipe with velocity techniques and then apply the deep learning model such as LSTM [10], Bi-LSTM [11], and GRU [12] for better comparison of effective training applied to the larger dataset.

We collected our own data set that starts with a small data set from 6 deaf participants consisting of 20 signs with 580 videos, captured in different environments and lighting conditions to reflect real-world scenarios.

The paper is structured as follows. Section 2 provides a review of related work in sign language recognition, Section 3 outlines the methodology used for KSL recognition, Section 4 presents the experimental setup, Section 5 presents the results, and Section 6 concludes with a discussion of the findings and potential future directions for research.

2 Related Work

Sign language recognition has gained significant attention due to its potential to bridge the communication gap for deaf and hard-of-hearing individuals. The Word-Level American Sign Language Dataset (WLASL), introduced by Li et al. [13], stands as the largest word-level ASL dataset, featuring over 21,000 video samples across 2,000 glosses performed by 119 signers. This dataset's inclusion of significant inter-signer variability and annotations for dialectal variations has positioned it as a benchmark for large-scale word-level sign recognition.

In recent years, advancements in machine learning and deep learning, particularly with transformer-based architectures like ViViT [9] and spatiotemporal models like the SlowFast Network, have enabled significant progress in the computer vision field [7, 14, 15].

The SlowFast network has emerged as a powerful architecture for advancing sign language recognition tasks. Ahn [16] leveraged a two-pathway SlowFast network for Continuous Sign Language Recognition (CSLR), introducing Bi-directional Feature Fusion (BFF) and Pathway Feature Enhancement (PFE) to achieve state-of-the-art performance on datasets such as PHOENIX14 [17] and CSL-Daily [18]. Hassan [19] extended the application of SlowFast Networks to dynamic sign language recognition on the WLASL dataset and achieved a 79.34% in top-1 accuracy. Similarly, Radhakrishnan [20] demonstrated the effectiveness of the SlowFast model in word-level sign language detection on the MSASL dataset [21], achieving a 92.35% increase in top-1 accuracy. These studies highlight the versatility and robustness of SlowFast architectures across diverse sign language recognition tasks.

In 2018, Tran et al. proposed another approach for video classification called Channel-Separated Convolutionals (CSN) [8]. A new method was presented to add factorizing on the channel, while factorizing simply on the spatial and temporal dimensions. This approach was evaluated on Kinetics-400 dataset [22], achieving 76.6% in top-1 accuracy without any pre-trained dataset applied. By applying pre-trained dataset with Sport1M [23] and evaluate on the same data, the CSN model increased the top-1

accuracy by 1.8%.

In 2020, Camgoz et al. [24] proposed a transformer-based model joint end-to-end sign language recognition and translation, using Connectionist Temporal Classification (CTC) loss. Their approach achieved state-of-the-art results on the RWTH-PHOENIX-Weather-2014T dataset [25], improving both recognition and translation accuracy, with a significant boost in BLEU scores for translation tasks. In the same year, De Coster et al. [26] proposed a method for Sign Language Recognition using Transformer Networks. They combine OpenPose-based feature extraction with end-to-end feature learning using CNNs for sign language recognition [27]. Applying the multi-head attention mechanism [28] from transformers, they recognize isolated signs in the Flemish Sign Language corpus, achieving 74.7% accuracy on a 100-class vocabulary, significantly outperforming previous methods. Three years later, Kothadiya et al. [29] proposed SIGNFORMER, a Vision Transformer model for static Indian sign language recognition. It divides signs into positional embedding patches processed by a transformer with self-attention layers. The model achieved 99.29% accuracy with minimal training epochs, outperforming convolution-based architectures and showing effectiveness under various augmentations.

Keypoint-based sign language recognition has been developing by utilizing 2D/3D landmarks (hand, face, and body) that emphasize motion dynamics and are essential to the recognition task. For Indian SLR, Subramanian et al. [30] presented a MediaPipe-optimized GRU (MOPGRU) in 2022. After being normalized and denoised, holistic landmarks are modeled using a modified GRU cell whose update gate is conditioned by the reset gate. This results in significant gains over the video corpus's vanilla LSTM/GRU baselines. This study has demonstrated that recurrent design on posture sequence can compete with more complex vision backbones while maintaining real-time compatibility.

Building on similar pose-first pipelines, Bhadouria et al. present an LSTM-based SLR system that ingests MediaPipe landmarks extracted from videos and reports strong recognition with a compact temporal model—underscoring that recurrent architectures remain competitive when features are cleanly factorized into trajectories of keypoints [31].

3 Methodology

3.1 Dataset Collection

To work on the Khmer Sign Language Recognition task, a small corpus was developed for the training and evaluation. The dataset consists of 580 videos recorded by 6 deaf participants at the National Institute for Special Education (NISE). The signs represent everyday concepts and actions, including directional terms, locations, and objects commonly encountered in daily life. For better organization, the terms can be grouped into categories and listed as follows. Under Locations and Actions, we begin with "Where", "When", "Market", "Buy", and "Location". In the Directions category, we list "Left", "Right", "North", and "South". Next, under Objects, we have "Pen", "Blue Pen", "Red Pen", "Pencil", "Book", "Line", and "Eraser". Lastly, in the People category, we list "Teacher", "Director", "Female Director", and "Deputy Director". This structure emphasizes the most important terms first, with clear grouping and Khmer translations in parentheses for better understanding.

The data was recorded at 30 frames per second with a resolution of 1080 x 1920 pixels, using both cameras and smartphones across various environments and lighting conditions. This diverse dataset is well-suited for training and evaluating sign language recognition models in real-world scenarios. A sample of the proposed dataset is shown Table 1.

3.2 RGB video frame with Neural Network Approaches

Figure 2 illustrates the Khmer Sign Language Recognition System workflow from video data using the Sign Recognition models (CSN, SlowFast and ViViT). The process begins with the input of raw video data, which undergoes preprocessing steps to prepare it for model input. These steps include frame extraction, resizing, and normalization to ensure that the video frames are consistent in size and values before they are passed to the model. Once the preprocessing was done, it fed processed data into the recognition models. The extracted features are then used for Classification, where the model assigns a specific label (sign) based on the learned

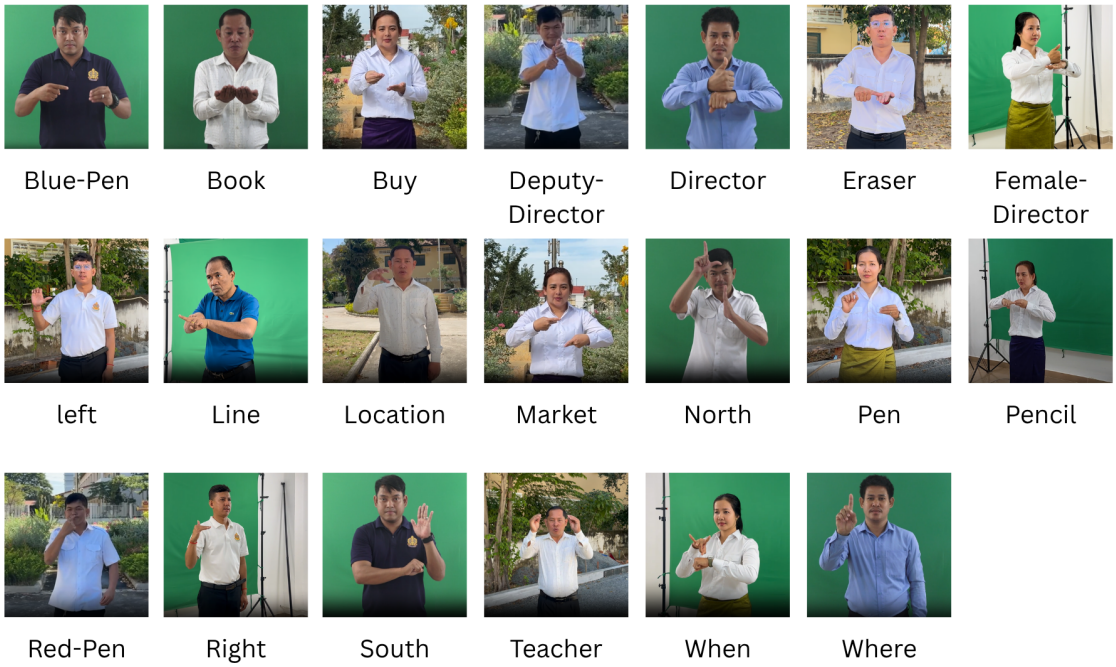

Figure 1. Sample Video Khmer Sign Language Dataset

patterns from the KSL dataset. This stage generates the prediction for the input video, which is the recognized sign. Finally, the output from the classification stage is presented as the recognized sign, corresponding to the KSL sign, providing valuable feedback for the user. For KSL recognition tasks, 4 models were selected for the experiment: Video Vision Transformers (ViViT), SlowFast Network, and Channel-Separated Convolutional Networks (CSN), and R3D-18.

3.2.1 Video Vision Transformers (ViViT)

This model proposed by Arnab et al. [9], apply transformer architectures to video data by dividing video frames into spatial patches and learning long-term temporal dependencies. This method surpasses convolutional approaches in datasets like Kinetics-400 due to its attention-based mechanism. ViViT's ability to model complex spatial-temporal interactions makes it suitable for capturing the nuances of sign gestures.

3.2.2 The SlowFast network

Introduced by Feichtenhofer et al. [7], processes video inputs at two different frame rates—one fast pathway for motion-sensitive features and one slow pathway for fine-grained spatial details. This dual-pathway architecture excels in capturing both dynamic and static elements of actions, making it particularly effective for sign language gestures that combine subtle hand movements with facial expressions. Recent work demonstrates its strong performance on datasets such as Charades and AVA, highlighting its adaptability to video-based tasks.

3.2.3 Channel-Separated Convolutional Networks (CSN)

This neural network has proposed by Tran et al. [8] which designed for tasks like video analysis. In these networks, the layers that process information are split into two types: one type focuses on mixing information across different channels (like colors or features) without looking at the surrounding space, and the other type focuses on analyzing patterns in the local space (like shapes or movements) without mixing channel information. This separation makes the network more efficient and specialized. Traditional networks combine both tasks in one step, but CSNs handle them separately. This idea has been proposed in Xception [32] and MobileNet [33] for image classification and R(2+1)D [34] video classification, but CSNs specifically aim to separate channel-related processing from spatial and temporal processing (movement-related).

3.2.4 ResNet 3D (R3D)

3D Residual Networks (R3D), proposed by Hara et al. [35] are another version of the 2D ResNet design [36], are CNNs for video that replace 2D spatial operations with 3D spatio-temporal kernels (time × height × width). In R3D, each residual block jointly models appearance and motion in a single 3D convolution while preserving skip connections, allowing deeper networks to train stably. This lets the model learn direct spatio-temporal features from short clips rather than single frames, which is effective for action and sign-language recognition. Unlike factorized designs such as R(2+1)D [37], which split spatial and temporal processing into separate steps, R3D performs them together in one operation—yielding a strong, simple baseline. A common lightweight instance is R3D-18, widely used as a backbone in this research.

3.3 Pose Key Point with Deep Learning Approach

3.3.1 Training Process

Similarly, from the previous training to training with the keypoint process with deep learning models, we start from the video frame. There are some slight differences in the preprocessing dataset step: we do all frame extraction from the video and the sampling from thoes frames. Then we will apply the keypoint extraction with mediapipe [38]. In order to capture the movement of the keypoint for each word, we apply the velocity technique [39]. The next step is the training process with deep learning models (LSTM, Bi-LSTM, and GRU) and end with the evaluation method.

3.3.2 Long Short-Term Memory (LSTM)

The model achitecture was introduced by Hochreiter and Schmidhuber in 1997 [10], the canonical origin of LSTM. It is a gated RNN that combats vanishing/exploding gradients by adding a persistent cell state and three gates (input, forget, and output). With this can help the model carry information over long time spans—crucial for signs whose meaning depends on motion over many frames.

3.3.3 Bidirectional Long short-Term Memory (Bi-LSTM

A bidirectional RNN [11] processes the sequence in forward and backward time and concatenates the two hidden states, giving access to past and future context at each frame. The principle was also introduced by Schuster and Paliwal in 1997; swapping the vanilla recurrent cell for LSTM yields Bi-LSTM, now standard for sequence labeling and SLR.

3.3.4 Gated Recurrent Unit (GRU)

This model is a lighter gated RNN [12] that merges the input/forget logic into update and reset gates—fewer parameters than LSTM, often similar accuracy. It was introduced in the RNN Encoder–Decoder work by Cho et al. (2014) and Chung et al. (2014) and was competitive with the LSTM model as well.

3.4 Evaluation Matrice

After the classification task to show the accuracy of each classification model, this task should be evaluated with evaluation matrices. In this paper, we choose the evaluation suitable for classification tasks such as , Precision, Recall, and F1-Score, Accuracy [40].

Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

3.4.1 Precision

The precision is the ratio $TP/(TP + FP)$ where TP is the number of true positives and FP the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{1}$$

3.4.2 Recall

The recall is the ratio $TP/(TP+FN))$ where TP is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2}$$

3.4.3 F1-Score

The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.

$$F1 == \frac{2TP}{2TP + FP + FN}. \tag{3}$$

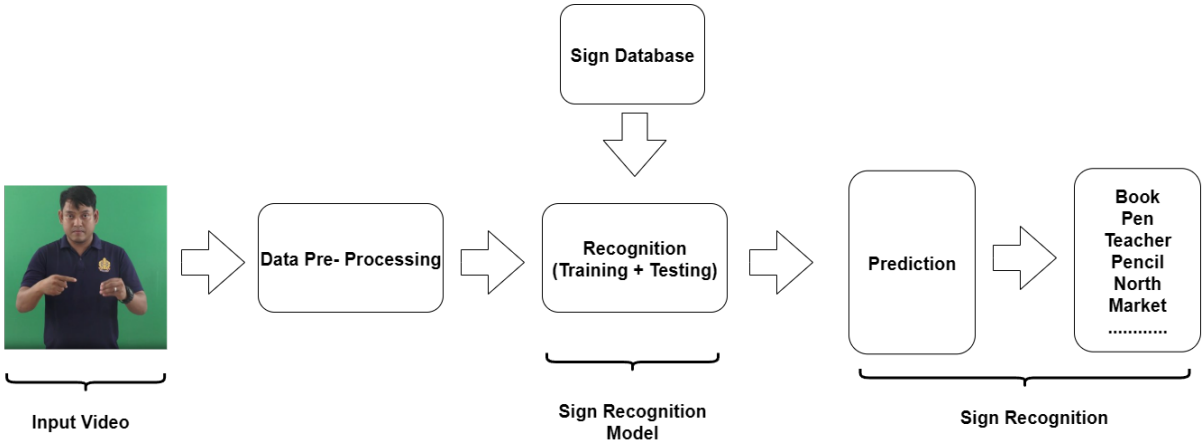

Figure 2. Flow Chart Frame Video Training Approach

3.4.4 Accuracy

In multilabel classification, accuracy function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true.

$$\text{Subset Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{\mathbf{y}}_i = \mathbf{y}_i)\,, \quad (4)$$

where $\hat{\mathbf{y}}_i$ and $\mathbf{y}_i$ are the predicted and true *label sets* for sample i, and $\mathbb{1}(\cdot)$ is the indicator function.

4 Experimental Setup

4.1 Data Pre-processing and Augmentation

We resize the resolution of all original video frames such that the diagonal size is 224 pixels for ViViT and 256 pixels for SlowFast, R3D-18, and CSN. For SlowFast training, R3D-18, and CSN, we center crop and apply normalization with the mean value of [0.45, 0.45, 0.45] and standard derivation value of [0.225, 0.225, 0.225]. For ViViT training, we randomly crop 224x224 pixels and apply normalization and random horizontal flipping. Note that, we randomly selected 30 frames.

4.2 Implementations Detail

All the experiments are implemented in Pytorch and pre-trained weight with Kinetics-400 datasets [22] are used for all models. We train all the models with Adam Optimizer [41]. Cross-Entropy loss function [42] is used in our experiments. All the models are trained with 50 epochs on each subset.

We split the sample of our dataset into the training and testing following the ratio of 70% and 30% of the total sample, respectively. To make sure the dataset is generalize for both training and testing set by manually for each class.

5 Result and Discussion

Table 2 indicates the performance of the two different approaches working on KSL 20 classes of the dataset: (i) RGB frame video architecture (SlowFast, CSN, ViViT, R3D-18), (ii) the pose keypoint with the sequence model (LSTM, Bi-LSTM, GRU, ResNet-3D-18). The evaluated methods are precision, recall, F1-score, and accuracy.

On the RGB side, SlowFast emerged as the top performer, achieving an accuracy of 92.81% and an F1-score of 92.50%. Its success can be credited to its unique dual-pathway architecture, which processes both fast and slow temporal features simultaneously. This design allows it to capture fine-grained motion details while also understanding broader temporal patterns, making it particularly well-suited for the complexities of sign language recognition. CSN came in second, with an accuracy of 82.87% and an F1-score of 81.20%. While its channel-separated convolution approach effectively reduces redundancy and improves feature extraction, it falls short of SlowFast's performance because it lacks the explicit temporal modeling capabilities that make SlowFast so effective. ViViT, a transformer-based model, achieved an accuracy of 75.13%, showing that while transformers are powerful for processing sequences,

Table 2. The results of KSL recognition using different neural network approaches.

Architecture	Precision	Recall	F1-Score	Accuracy
RGB Video–Based Models				
CSN	86.00%	82.87%	81.20%	82.87%
SlowFast	**93.57%**	**92.81%**	**92.50%**	**92.81%**
ViViT	82.68%	75.13%	75.31%	75.13%
R3D-18	88.35%	84.09%	83.72%	84.09%
Pose Keypoint–Based Models (and Fusion)				
Pose Keypoint + LSTM	90.00%	88.00%	87.00%	88.00%
Pose Keypoint + Bi-LSTM	87.00%	86.00%	86.00%	86.00%
Pose Keypoint + GRU	84.00%	80.00%	80.00%	81.00%
Pose Keypoint + R3D-18	**95.21%**	**92.05%**	**92.56%**	**92.05%**

they may not be as effective for recognizing fine-grained motion in sign language. This could be because transformers typically require large amounts of data to perform well and lack the built-in spatial and temporal processing advantages of CNN-based models.

For the pose-keypoint approach, models ingest normalized joint trajectories (with velocity) and focus on explicit motion cues while being robust to background and lighting. LSTM out perform the accuracy result amount of RNNs with an accuracy of 88% and 87% of F1 score. followed by Bi-LSTM (accuracy of 86% and 86% with F1 score) and GRU accuracy of 81% and 80% of F1 score. This result suggests that making motion explicit via velocity helps, and that a well-regularized Bi-LSTM can be competitive when clips provide sufficient context.

The best overall for the performance is pose keypoint with ResNet-3D-18, which come up with an accuracy of 92.05% and an F1 score of 92.56% althought it is slighly beaten SLowFast on F1 score, with the precision of 95.21% showing that it has the fewest false positives.

These results highlight that models explicitly designed for motion modeling, such as **SlowFast** with RGB video, outperform others by efficiently capturing both spatial and temporal features in sign language recognition tasks. However, pose keypoint with **ResNet-3D-18** also performs strongly by combining motion with limited appearance cues; in our summary (Table 3), this configuration attains the best F1 with the highest precision while keeping medium training and inference cost. In resource-constrained or cluttered backgrounds, the **Pose Keypoint + LSTM/Bi-LSTM/GRU** family is attractive: it models *motion only*, is *low/low* for training and inference compute, and is notably robust to background, though it may miss appearance-dependent distinctions (e.g., subtle hand texture or mouthing). **RGB + R3D-18** remains a solid, balanced baseline (*medium* train/infer) but lacks the dual-rate temporal pathway that gives **SlowFast** its edge. Finally, **CSN/ViViT** capture both motion and appearance but typically require medium/high compute and careful data/tuning; when such resources are available, they can be highly competitive. Overall, the choice hinges on data scale, hardware budget, and how much appearance information is needed; a pragmatic path is to start with keypoints (for efficiency and robustness) and scale to RGB+3D CNNs as resources permit.

6 Limitation

This experiment is still working on the small dataset with the limitation of the signer and variety of the environment. For the first experiment, we start working on the splitting, which follows the 70/30 train-test protocol rather than a signer-independent schema, which can lead to overestimating performance on an unseen signer.

We did not include a separate validation set or K-fold cross-validation, which could lead to suboptimal hyperparameters and early stopping. For the model we rely on the kinetic-400 pretraining and a single keypoint encoding (MediaPipe landmarks), and for evaluation we only apply accuracy, precision, recall, and F1 score, without per-class evaluation analysis and other robustness techniques.

7 Conclusion

To conclude, among all the experiments that we have conducted, the pose keypoint with

Table 3. Model families, compute, and behavior.

Approach	Captures	Train	Infer	Notes
RGB Video + SlowFast	Motion + appearance	High	High	Strong RGB-only
RGB Video + R3D-18	Motion + appearance	Med	Med	Solid baseline
CSN / ViViT	Motion + appearance	Med/High	Med/High	Needs data/tuning
Pose Keypoint + LSTM/Bi-LSTM/GRU	Motion only	Low	**Low**	Robust to background
Pose Keypoint + R3D-18	Motion + appearance	Med/High	**Med/High**	Best F1; highest precision

the ResNet-3D-18 model performed better for Khmer Sign Language recognition, with effective evidence, even if it was slightly beaten by the SlowFast model with RGB video. Additionally, when it comes to increasing the number of datasets, it will be beneficial for computing resource limitations. Our results demonstrate the effectiveness of the pose keypoint with the ResNet-3D-18 in sign language recognition, achieving an accuracy of 92.05%, an F1 score of 92.56%, a recall of 92.05%, and a precision of 95.21%. This work addresses the gap in KSL recognition and contributes to improving accessibility and a better learning approach for the deaf community in Cambodia. To further improve the system's accuracy and adaptability, especially in real-world situations, future research will investigate expanding the dataset, adding more sign modifications, and experimenting with different model architectures. For deaf people in Cambodia, the creation of a KSL recognition system will be very helpful, enabling improved communication and enhancing educational opportunities and inclusivity.

8 Acknowledgment

This research was initiated and supervised by Dr. Ye Kyaw Thu and was made possible through the collaboration between the National Institute for Special Education (NISE) and the Cambodia Academy of Digital Technology (CADT). We would like to thank the National Institute for Special Education (NISE) for providing essential resources and contributing to the data collection, as well as for manually verifying the dataset. We also extend our gratitude to the Cambodia Academy of Digital Technology (CADT) for offering the necessary support, including research materials and opportunities.

References

[1] A. Firth, *Deafness and Hard of Hearing*. Berkeley, CA: Apress, 2024, pp. 147–182. [Online]. Available: https://doi.org/10.1007/979-8-8688-0152-5_5

[2] J. S. Izquierdo-Condoy, L. E. Sánchez Abadiano, W. Sánchez, I. Rodríguez, K. De La Cruz Matías, C. Paz, and E. Ortiz-Prado, "Exploring healthcare barriers and satisfaction levels among deaf individuals in ecuador: A video-based survey approach," *Disability and Health Journal*, vol. 17, no. 3, p. 101622, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1936657424000530

[3] M. A. De Rosa, M. T. Bernardi, S. Kleppe, and K. Walz, "Hearing loss: Genetic testing, current advances and the situation in latin america," *Genes*, vol. 15, no. 2, 2024. [Online]. Available: https://www.mdpi.com/2073-4425/15/2/178

[4] A. Derrington, "The (Lack of) Deaf Culture in Cambodia — Ashley Derrington — ashleyderrington.com," https://www.ashleyderrington.com/blog/post-5#:~:text=Of%20those%2016%20million%2C%20roughly,deaf%20person%20in%20the%20world, [Accessed 29-01-2025].

[5] C. J. Waterworth, M. Marella, M. F. Bhutta, R. Dowell, K. Khim, and P. L. Annear, "Access to ear and hearing care services in cambodia: a qualitative enquiry into experiences of key informants," *The Journal of Laryngology 38; Otology*, vol. 138, no. 1, p. 22–32, 2024.

[6] B. Baghdasaryan, G. Ghawi, T. Godfrey-Faussett, and U. H. Castillo, "Paving the pathway: Inclusive education for children with disabilities in cambodia," *UNICEF Innocenti-Global Office of Research and*

Foresight,, 2024.

[7] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," 2019. [Online]. Available: https://arxiv.org/abs/1812.03982

[8] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," 2019. [Online]. Available: https://arxiv.org/abs/1904.02811

[9] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," 2021. [Online]. Available: https://arxiv.org/abs/2103.15691

[10] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks," 9 2019. [Online]. Available: https://arxiv.org/abs/1909.09586

[11] R. Mayya, V. Venkataraman, A. P. R, and N. Darapaneni, "A Novel Bi-LSTM And Transformer Architecture For Generating Tabla Music," 4 2024. [Online]. Available: https://arxiv.org/abs/2404.05765

[12] R. Dey and F. M. Salem, "Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks," 1 2017. [Online]. Available: https://arxiv.org/abs/1701.05923

[13] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," 2020. [Online]. Available: https://arxiv.org/abs/1910.11006

[14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, Sep. 2022. [Online]. Available: https://doi.org/10.1145/3505244

[15] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.

[16] J. Ahn, Y. Jang, and J. S. Chung, "Slowfast network for continuous sign language recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 3920–3924.

[17] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney, "Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus," 05 2012.

[18] Q. Zhu, J. Li, F. Yuan, J. Fan, and Q. Gan, "A chinese continuous sign language dataset based on complex environments," 2024. [Online]. Available: https://arxiv.org/abs/2409.11960

[19] A. Hassan, A. Elgabry, and E. Hemayed, "Enhanced dynamic sign language recognition using slowfast networks," in *2021 17th International Computer Engineering Conference (ICENCO)*, 2021, pp. 124–128.

[20] S. Radhakrishnan, N. C. Mohan, M. Varma, J. Varma, and S. N. Pai, "Cross transferring activity recognition to word level sign language detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 2446–2453.

[21] 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.

[22] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.06950

[23] B. Varadarajan, G. Toderici, S. Vijayanarasimhan, and A. Natsev, "Efficient large scale video classification," 2015. [Online]. Available: https://arxiv.org/abs/1505.06250

[24] 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 (CVPR)*, June 2020.

[25] N. C. Camgoz, S. Hadfield, O. Koller,

H. Ney, and R. Bowden, "Neural sign language translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7784–7793.

[26] M. De Coster, M. Van Herreweghe, and J. Dambre, "Sign language recognition with transformer networks," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 6018–6024. [Online]. Available: https://aclanthology.org/2020.lrec-1.737/

[27] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," 2017. [Online]. Available: https://arxiv.org/abs/1611.08050

[28] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[29] D. R. Kothadiya, C. M. Bhatt, T. Saba, A. Rehman, and S. A. Bahaj, "Signformer: Deepvision transformer for sign language recognition," *IEEE Access*, vol. 11, pp. 4730–4739, 2023.

[30] 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, no. 1, 7 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-15998-7

[31] A. Bhadouria, P. Bindal, N. Khare, D. Singh, and A. Verma, "Lstm-based recognition of sign language," in *Proceedings of the 2024 International Conference on Contemporary Computing (IC3)*, ser. IC3 '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 508–514.

[32] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017. [Online]. Available: https://arxiv.org/abs/1610.02357

[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[34] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," 2018. [Online]. Available: https://arxiv.org/abs/1711.11248

[35] K. Hara, H. Kataoka, and Y. Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition," 8 2017. [Online]. Available: https://arxiv.org/abs/1708.07632?utm_source=chatgpt.com

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 12 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[37] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," 11 2017. [Online]. Available: https://arxiv.org/abs/1711.11248

[38] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," 6 2019. [Online]. Available: https://arxiv.org/abs/1906.08172

[39] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 104–111.

[40] scikit-learn developers, "Scikit-learn: Machine learning in python — performance metrics (precision_score, recall_score, accuracy_score, f1_score)," https://scikit-learn.org/stable/modules/model_evaluation.html, scikit-learn, accessed: 2025-10-01.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[42] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," 2023. [Online]. Available: https://arxiv.org/abs/2304.07288