A Comparative Analysis of Unimodal and Multi-Modal Architectures for the Robust Recognition of Khmer Consonant Hand Signs

Sereyratanak Heng **Sethisak San**
School of Digital Technology, American University of Phnom Penh, Phnom Penh, Cambodia
2023472heng@aupp.edu.kh

Abstract

Automated systems for under-resourced languages like Khmer Sign Language require real-world robustness, a factor often overlooked in favor of accuracy on clean data. This study evaluates deep learning models on both a large, curated dataset of 33 Khmer consonant signs and a manually created "Challenge Set" featuring realistic degradations. We systematically compared unimodal (Vision-Only, Skeleton-Only) and various multi-modal fusion architectures (MLP, LSTM, Attention). Our findings were decisive and counter-intuitive. A unimodal Skeleton-Only (LSTM) model was the most robust, achieving 81% accuracy on the Challenge Set. In stark contrast, all multi-modal fusion models, which combined skeletal data with features from a pretrained EfficientNetB0, underperformed significantly, with an advanced attention model collapsing to just 11% accuracy. We identify this failure as a critical case of "Modality Mismatch," where the brittle vision model produces erroneous, high-confidence features ("Confident Garbage") that degrade the fusion process. This work proves that for applications with a significant domain shift, a simpler, more robust unimodal model can be decisively superior to a complex multi-modal system, challenging the assumption that more data is always better.

Keywords: *Deep Learning, Machine Learning (ML), Convolutional Neural Network (CNN), Mediapipe, Sign Language, Classification*

1 Introduction

A person who is not able to hear as well as someone with normal hearing is said to have hearing loss, usually called deaf or hard-of-hearing. Those people use Sign language as the primary means of communication, yet the lack of widespread understanding often create significant barriers in education, healthcare and daily life. This communication gap can lead to social isolation, reduced opportunities, and limited access to essential services. Despite the serious of this issue, it is often overlooked and the public awareness is remaining low. However, only a small percentage of people are capable of using sign language, most of whom are either members of the deaf community or professionals working in the interpreting field. Hard-of-hearing refers to people with hearing loss ranging from mild to severe. It can start from one ear or both, leading to difficulty in hearing for conversation, speech, or loud sounds. The use of sign language by deaf and hard-of-hearing people plays a vital role in the interaction of communities around the world. According to the World Federation of the Deaf, there are approximately 70 million Deaf individuals worldwide (2024) [1] . There are more than 300 distinct sign languages around the world, as each county has its own grammar and lexicon that is inspired by their spoken language (UN, 2024) [2].

Figure 1. On the International Day of the Deaf at the Ministry of Social Affairs [3]

The Institute of Statistics of the Ministry of Planning has stated that there are 19,993 deaf or hard-of-hearing people, which is equivalent to 2.9% of the total population in Cambodia reported by the census [3].

According to the World Health Organization

(WHO, 2025) [4], the causes of hearing loss and deafness can be classified based on different stages of life. **Prenatally,** they may be due to genetic factors or infections such as rubella. **Perinatal** risks include birth asphyxia, jaundice, low birth weight, and related complications. In **childhood and adolescence**, chronic ear infections and meningitis are key contributors. In **adulthood and older age**, hearing loss may stem from chronic diseases, lifestyle factors, or age-related conditions. Across all stages, factors such as ear trauma, noise exposure, ototoxic drugs or chemicals, poor nutrition, and progressive genetic conditions can also lead to hearing impairment.

The deaf and hard-of-hearing (DHH) community share various experience of discrimination, stigma and prejudice from hearing people with other linguistically and culturally minority hearing groups in the United States and organizational networks [5]. For example, a DHH person may have trouble communicating with their hearing family members, suffer from bullying at school, or encounter a conflict between their own values as a DHH individual and the values and expectations of others in their environment [6].

Hearing loss impacts multiple dimensions of life, including individual, social, and societal levels. At the individual level, it leads to limitation in communication and speech development, along with adverse effects on cognitive functioning. On the social level, DHH individual frequently experience isolation, loneliness and stigma. For the societal level, DHH people encounter barriers in accessing education and employment, resulting in reduced workforce participation and increase economic burden. From a public health perspective, hearing loss substantially increases years lived with disability (YLDs) and contributes to disability-adjusted life years (DALYs), highlighting its serious global impact (World Health Organization, 2021) [7].

Assistive devices include hearing aids, and cochlear implants provide important support for people with DHH. However, these solutions are not universally accessible or effective for everyone because the World Health Organization estimates that only about 3% of the need for hearing aids has been provide, leaving a large portion of people without adequate support [8]. Olkin

Figure 2. UN providing hearing aid [8]

(2002) highlighted that many professional training sites lack adequate support for deaf individuals, noting that about 80% were missing essential tools like teletypewriter (TTY) systems, which hinder equal participation for trainees with hearing loss. [9].

In order to improve sign language learning for all groups, Dr. Naa claims that machine learning and related technologies can be used to translate sign language into words and vice versa [10].

Despite the growing advancement of sign language recognition worldwide, Khmer Sign Language (KSL) remains a low-resource language, constrained by limited datasets and insufficient technological support. Currently, no robust KSL recognition system exists in Cambodia that can effectively translate gestures into text. This scarcity creates substantial barriers to accessibility and inclusion for the deaf community.

This research aims to address this gap by developing machine learning models capable of translating KSL into text. The proposed system seeks to advance sign language recognition in low-resource contexts, such a system is essential for promoting more equitable communication and participation in Cambodian society, with particular potential to enhance accessibility in education and other critical domains.

2 Literature review

2.1 Object detection

Sign language recognition has seen significant advancements in recent years with the integration of deep learning and real-time object detection techniques. Buttar et al. (2023) developed a hybrid approach for American Sign Language (ASL) recognition that effectively com-

bines YOLOv6 [11]—a state-of-the-art real-time object detection model—for static hand gesture detection, with Long Short-Term Memory (LSTM) [12] networks and MediaPipe [13] to recognize dynamic gestures [14]. This comprehensive solution bridges the gap between static and dynamic sign recognition, enhancing the overall system's flexibility and performance. In the domain of few-shot object detection, several innovative approaches have emerged to address the challenge of limited labeled data. The Meta-DETR framework proposed by Zhang et al. (2022) eliminates the need for traditional region proposal methods by leveraging image-level few-shot learning [15]. It is capable of detecting a wide variety of object categories such as animals, vehicles, household items, and tools, even with minimal training examples. This model emphasizes recognizing novel or unseen classes by learning semantic relationships between base and new categories. Similarly, Zhang et al. (2021) introduced a method for accurate few-shot object detection by incorporating support-query mutual guidance and hybrid loss functions. Their approach enhances detection performance for rare object classes by effectively exploiting the relationship between support and query images [16].

2.2 Sign Language Recognition

Focusing on sign language recognition in different linguistic contexts, Shenoy et al. (2021) developed a real-time Indian Sign Language (ISL) recognition system that uses a smartphone camera to identify 33 static hand poses and 12 dynamic gestures [17]. By applying grid-based feature extraction and a k-Nearest Neighbors (k-NN) classifier, their system achieved high accuracies of 99.7% for static poses and 97.23% for dynamic gestures, significantly aiding communication for individuals with hearing and speech impairments. In a comparative study, Kondo et al. (2024) analyzed the performance of Vision Transformers (ViT) [18] versus Convolutional Neural Networks (CNN) [19] for Japanese Sign Language recognition [20]. Their findings revealed that ViT models, particularly those using angular features, outperform traditional CNNs, offering valuable insights into the potential of transformer-based architectures in sign language applications. Daniels et al. (2021) proposed a recognition system for Indonesian Sign Language (BISINDO) that utilizes the YOLOv3 object detection algorithm [21]. The system demonstrated excellent performance, achieving 100% accuracy on static image data and 72.97% accuracy on dynamic video data, underscoring the promise of real-time object detection models for sign language recognition. Additionally, Dong et al. (2022) introduced Incremental-DETR, an extension of the DETR [22] architecture that integrates fine-tuning with self-supervised learning [23]. This method allows the model to learn new object classes with minimal labeled data while maintaining performance on previously learned base classes. It addresses the issue of catastrophic forgetting and overfitting, making it a robust solution for few-shot learning scenarios. Recent studies have shown that MediaPipe Holistic provides a reliable framework for extracting multimodal landmarks of the hands, face, and body, which significantly improves the performance of continuous sign language recognition models [24].

These studies collectively illustrate the rapid evolution of sign language and object detection technologies. They highlight the importance of integrating real-time detection, few-shot learning, and advanced deep learning models to build robust, accurate, and adaptable recognition systems.

3 Methodology

Our methodology is designed to rigorously test our model performance in not only the ideal scenarios but also real-world scenarios as well. In order to achieve this, we created two distinct datasets, a dual-stream data preprocessing pipeline, and a collection of model architectures for comparative analysis.

3.1 Dataset and Evaluation Strategy

Dataset classes

Since the Khmer alphabet consists of 33 letters, we created 33 separate classes, each class representing a different alphabet. The majority of hand shapes in the Khmer Sign Alphabet are characterized more by the position of the fingers than by their specific orientation. For instance, even when the hand is slightly rotated, the particular finger arrangement for "kor" keeps the hand shape visually different from other letters. For references to take images of each alphabet sign we used an app call 'Khmer

Figure 3. Simple Classes of Khmer Alphabet Sign Language from Krousar Thmey [25]

Sign Language' that is developed through collaboration between Research Triangle Institute (RTI) and Ministry of Education, Youth and Sport (MoEYS) under leadership and technical support from Department of Information and Technology (DIT), Special Education Department (SED), and National Institute for Special Education (NISE) [26]. In addition, the Khmer Sign Language chart developed by Krousar Thmey was used as a reference for standardizing the representations of 33 consonants (figure 3) [25]. These resources together served as the foundation for building a structured dataset, which can be applied to machine learning models for sign-to-text translation.

3.1.1 Primary Training Dataset

As no large-scale public dataset is available for Khmer Consonant Hand Signs, a primary dataset was manually collected and curated for this study. To facillitate the rapid collection of a large volume of data, we captured frames from a video stream and convert them into individual images. This method allowed us to assemble a comprehensive dataset of over 17,000 images distributed across 33 classes each representing a unique Khmer consonant. The data was collected under controlled and uniform condition, characterized by consistent lighting, a single person, and recorded at the same background. Standard visual references were use to ensure the accuracy of the performed signs. This dataset served as the foundation for training and validating our models in a controlled environment.

3.1.2 The "Challenge Set": A Real-World Robustness Benchmark

To evaluate model robustness beyond the controlled environment of the primary dataset, we also curated a "Challenge Set". This smaller but more difficult dataset was specifically collected and designed to simulate common real-world data degradation and represent a domain shift from the training data. The data consists of: (1) Images from people not present in the training set, introducing more variations in hand shape, size, and skin tone. (2) Image were captured in

Figure 4. Example of some classes

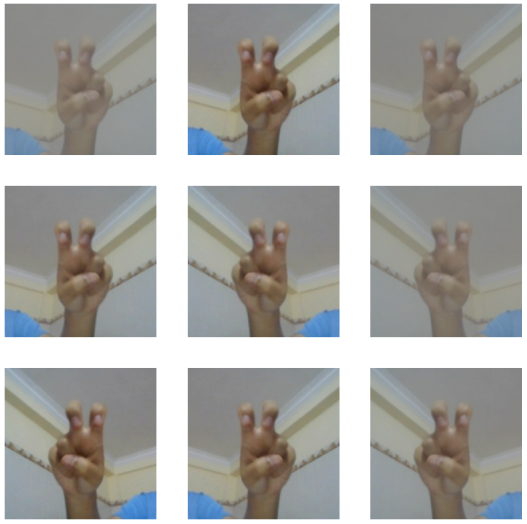

Figure 5. Example of Augmented Image

an environment with different lighting and background (3) Signs performed against busy, real-world backgrounds

By evaluating models on this challenge set as the central focus of our robustness analysis, it will reveal the true generalization capability of each architecture.

3.2 Multi-Modal Data Preprocessing

We implemented a two-stream preprocessing pipeline utilizing Google's MediaPipe HandLandmarker to extract the visual and skeletal features from each raw image.

Image Augmentation

3.2.1 Vision Stream Pipeline

The objective of the vision pipeline is to generate a focused, normalized image of the hand for the Convolutional Neural Network. (1) MediaPipe's HandLandmarker is used to detect the 21 keypoints of the hand in the full resolution image. (2) Bounding box is then calculated from the detected landmark and expanded with a 20-pixel padding to ensure the entire hand is caputred while minimizing the background noise. (3) The original image is then cropped based on the bounding box and resized to a uniform 224x224 pixels. (4) The resized image is processed using the specific preprocess_input function corresponding to its CNN backbone(EfficientNetB0) to scale the pixel values to the expected range. (5) During training, Image augmentation will be applied to the dataset to increase the variation in the Dataset. The image augmentation techniques used for this current model are Horizontal Flipping, Random Contrast, and Random Brightness as shown in figure 5.

3.2.2 Skeleton Stream Pipeline

The skeleton pipeline extracts a quantitativem geometric representation of the hand's pose. (1) The normalized 3D world coordinate (x, y, z) of the 21 keypoints are extracted from MediaPipe output. (2) The landmarks are structured into a tensor with the shape of (21, 3) in order to make it suitable for models designed to process sequential data.

3.3 Model Architectures

To systematically evaluate the different learning strategies for Khmer Consonant Hand Sign Recognition, We compared a collection of unimodal and multi-modal architectures. All multimodal models share a common two-stream structure consisting of a Vision Branch to process image data and a Skeleton Branch to process landmark data, which are then combined by a fusion mechanism.

3.3.1 Baseline Multi-Modal Architecture (CNN + MLP)

Our baseline fusion architecture, illustrated in Figure 6, serves as the foundation for our com-

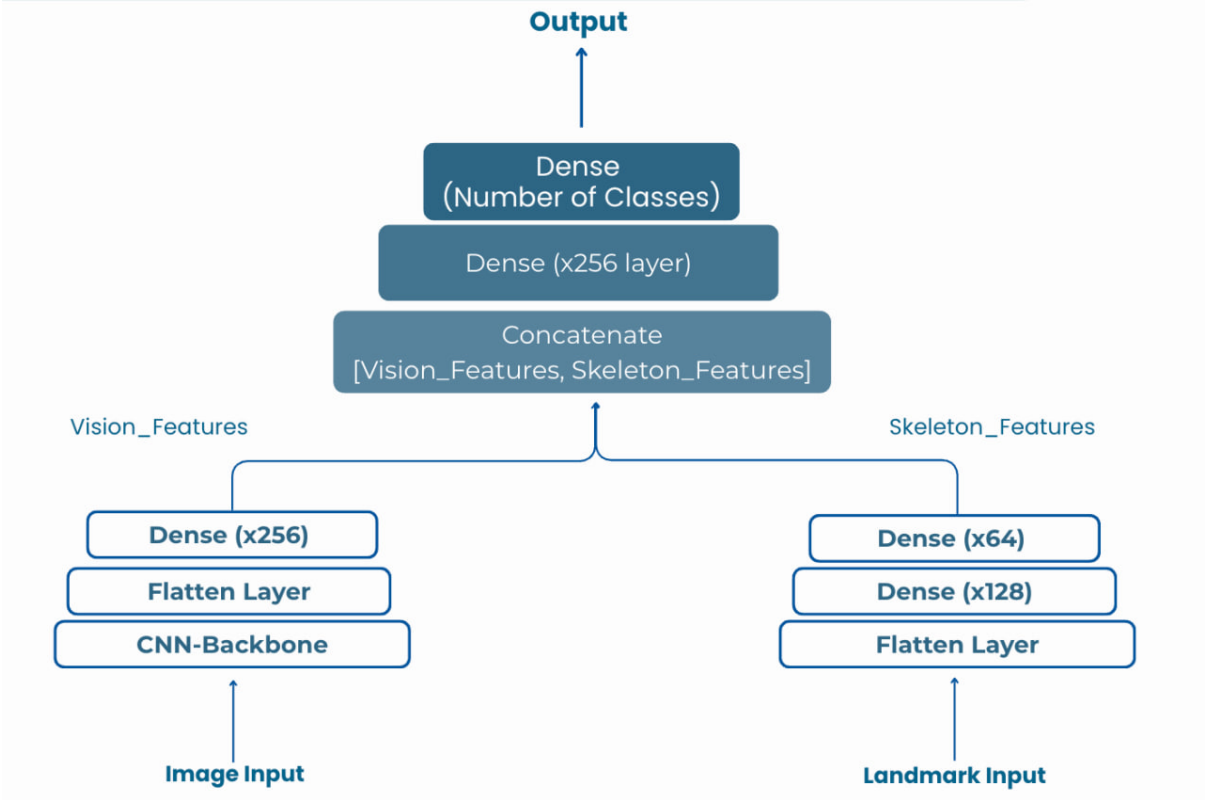

Figure 6. The baseline two-stream multi-modal fusion architecture.

parative analysis.

- **Vision Branch**: The visual stream take a preprocessed 224 x 224 pixel image as input. The image is fed into a pretrained CNN Backbone(EfficientNetB0), with its layers frozen to act as a powerful feature extractor. The resulting feature map is flattened into a vector and passed through a Dense layer with 256 units to produce the final Vision_Features
- **Skeleton Branch**: The skeletal stream takes a (21, 3) tensor of landmark coordinates as input. This tensor is flattened into a 63-dimensional vector and processed by a Multi-Layer Perceptron, consisting of a Dense layer with 128 units followed by Dense layer with 64 units. This produces the final Skeleton_Features
- **Fusion and Classification Head**: The Vision_Features and Skeleton_Features are then combined via a Concatenate layer. This unified feature vector is passed through a final classification head, consisting of a Dense layer with 256 units (ReLU) and a Dropout layer, before the final softmax output layer predicts one of the 33 classes.

3.4 Architectural Variations for Comparative Analysis

To investigate the effectiveness and robustness of the system, we evaluated several key architectural variations.

Unimodal Baselines:

- **Vision-Only Model**: Consists solely of the EfficientNetB0 Vision Branch and the final classification head.
- **Skeleton-Only Model**: Consists solely of a skeleton processing branch (LSTM or MLP) and the final classification head. This is crucial for establishing the standalone robustness of the geometric data.

Fusion Model Variations:

- **EfficientNetB0 + LSTM**: Replaces the MLP in the skeleton branch with a Long Short-Term Memory (LSTM) layer to test

if processing landmarks as a sequence improves performance.

- **Attention-Based Model (EfficientNetB0 + 1D-CNN + Attention)**: Our most sophisticated architecture. It replaces the MLP with a 1D-CNN to detect local geometric motifs in the landmarks. Crucially, it replaces simple concatenation with a gating-based Attention mechanism, designed to allow the model to learn the relative importance of the vision and skeleton streams during fusion.

This suite of models allows for a direct comparison of unimodal vs. multi-modal performance, as well as an analysis of the efficacy of different skeleton processing techniques and fusion strategies.

3.5 Experiments

Our experimental setup was designed to rigorously test our core hypotheses regarding model robustness under realistic conditions. All models were trained and evaluated using the TensorFlow/Keras framework.

- **Training Protocol**: All models were trained on the "clean" training set for a maximum of 50 epochs, using a BATCH_SIZE of 32 and the Adam optimizer. A memory-safe Data Generator was implemented to process images on-the-fly, preventing RAM exhaustion. An EarlyStopping callback with a patience of 7 epochs, monitoring val_loss, was used to ensure each model was trained to its optimal point before overfitting.

- **Evaluation Protocol**: Each optimally trained model was evaluated on two distinct datasets: The held-out Clean Test Set, to measure performance under ideal conditions. The manually curated Challenge Set, to measure true real-world robustness against data degradation.

4 Result

The experiments yielded a clear, consistent, and highly informative set of results. While all models performed exceptionally well on the clean, academic dataset, their performance diverged dramatically on the difficult real-world Challenge Set. The quantitative findings are summarized in table 1.

As shown in table 1, all evaluated models achieved near-perfect accuracy (92-100%) on the clean test set. However, on the Challenge Set, a stark performance hierarchy emerged. The Skeleton Model (LSTM) established itself as the most robust architecture, achieving the highest accuracy of 81.00%. In contrast, all multi-modal fusion models underperformed this unimodal baseline. The EfficientNetB0 + MLP model achieved 60.00%, while the EfficientNetB0 + LSTM and Attention-Based Model experienced a catastrophic performance collapse, dropping to 22.00% and 11.00% respectively.

5 Discussion and Findings

Our comprehensive evaluation led to a decisive and counter-intuitive primary finding: for the task of robustly recognizing Khmer Consonant Hand Signs under real-world conditions, the unimodal Skeleton-Only model is definitively the superior architecture. The failure of all multi-modal fusion attempts, particularly the catastrophic collapse of the advanced Attention-Based Model, provides powerful evidence for the "Confident Garbage" phenomenon. We diagnose this as follows:

- **Vision Model**: The EfficientNetB0 vision model, while powerful, demonstrates extreme brittleness when faced with the domain shift of the Challenge Set. Its standalone accuracy of 56% shows it struggles significantly with visual degradation.

- **Skeleton Model**: The Skeleton-Only model, processing pure geometric data, is largely immune to visual noise. It maintains a high accuracy of 81%, establishing itself as a highly robust and reliable signal source.

- **The Failure of Fusion**:When these two signals are fused, the vision branch produces high-confidence but erroneous feature vectors on degraded images. This "confident garbage" poisons the fusion process. The final classifier, attempting to reconcile a strong, correct signal with a loud, incorrect one, makes a poor compromise. This is why

Table 1. Comprehensive Comparison of Model Performance Across Key Metrics

Model Category	Model Configuration	Inference Time	Test Set (%)	Challenge Set (%)
Baseline Models	Vision Model (EfficientNetB0)	6.3 ms	100	56.00
	Skeleton Model (LSTM)	3.0 ms	100	81.00
Fusion Models	EfficientNetB0 + MLP	7.0 ms	100	60.00
	EfficientNetB0 + LSTM	8.8 ms	92.00	22.00
Attention-Based Model	EfficientNetB0 + 1D CNN + Attention	11.0 ms	100	11.00

the EfficientNetB0 + MLP fusion (60%) is worse than the skeleton model alone (81%).

- **The Attention Mechanism**: The Attention model's collapse to 11% is the most critical piece of evidence. This indicates that during training on the clean dataset, the model learned a fatal policy: to heavily trust the vision branch. When faced with the Challenge Set, it continued to apply this policy, actively ignoring the correct skeleton data and amplifying the "confident garbage" from the vision branch, leading to a near-total failure.

6 Future Work

Our findings clearly indicate that the primary bottleneck for building a state-of-the-art fusion model is not the fusion architecture itself, but the brittleness of the vision branch. Therefore, future work should prioritize improving the vision model's robustness.

- **Data-Centric Approach**: The most promising path is to curate a larger and more diverse visual training dataset. Actively collecting and augmenting the training set with examples of poor lighting, motion blur, multiple signers, and complex backgrounds is essential to teach the vision model to generalize.

- **Self-Supervised Pre-training**: Before fine-tuning on sign language data, the vision backbone could be pre-trained on a large, unlabeled dataset of diverse hand images. This would help it learn more general and robust features specific to hands, rather than relying solely on ImageNet.

- **Uncertainty-Gated Fusion**: Future research could explore fusion mechanisms that explicitly model the uncertainty of each branch's prediction, allowing the model to learn to completely discard the vision input when its confidence is low.

7 Conclusion

This research set out to identify the most robust architecture for Khmer Consonant Hand Sign Recognition. Through a systematic comparison of six distinct unimodal and multi-modal models on both a clean dataset and a real-world "Challenge Set," we arrived at an unambiguous conclusion. A simple, unimodal Skeleton-Only (LSTM) model, which processes geometric hand landmarks, decisively outperformed all complex multi-modal fusion architectures in robust, real-world conditions. Our analysis revealed a critical "Confident Garbage" phenomenon, where a powerful but brittle pretrained vision model actively degrades the performance of the more robust skeleton stream during fusion. This work serves as a vital case study, demonstrating that for applications with a significant domain shift, a simpler, more robust unimodal model can be the superior choice for deployment, challenging the prevailing assumption that multi-modal systems are inherently better. However, despite these impressive results, several limitations still remains. Confusion still persists among groups of similar signs, and the present research focus only on static alphabet gesture rather than continuous gesture. Additionally, the current research does not contains vowel signs or common phrases signs which is widely used by deaf people. Lastly, Real-world robustness which includes varying lighting, different backgrounds, skin tones, varying hand shapes, still have not been thoroughly tested and evaluated yet.

Based on our research and evaluation, future work will focus on expanding Khmer Sign Language Recognition (KSLR) for educational purposes. This paper had proposed the initial stage that concentrate on consonant recognition. After

this, we will extended to include vowels, numbers, and additional sign. This step-by-step expansion will provide a structured learning tool for children with hearing impairments. Alongside this, more experimentation with attention-based architectures will be conducted to enhance class separation, and robustness will be evaluated using real-world datasets under diverse conditions.

Acknowledgment

We would like to express our sincere gratitude to our advisor and professor, Dr. Ly Rottana, for his constant encouragement, insightful guidance, and dedicated support throughout this research. We also thanks our two former teammates who contributed to the initial writing of this paper. Finally, we extend our appreciation to the scholar and researchers whose prior studies provided the foundation upon which this research was built.

References

[1] World Health Organization. Deafness and hearing loss. https://www.who.int/health-topics/hearing-loss, 2025. [Online; accessed 14-September-2025].

[2] United Nations. Sign languages day. https://www.un.org/en/observances/sign-languages-day/, 2025. [Online; accessed 14-September-2025].

[3] EAC News. Mosvy secretary of state encourages persons with disabilities not to despair as the government seeks to help. https://eacnews.asia/home/details/15414, September 2022. [Online; accessed 14-September-2025].

[4] World Health Organization. Deafness and hearing loss: Fact sheet. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, February 2025. [Online; accessed 14-September-2025].

[5] A. Aldalur et al. Acculturative stress, mental health, and well-being among deaf individuals. *International Journal of Environmental Research and Public Health*, 20(9), 2023. [Online; accessed 14-September-2025].

[6] Tal Lambez, Maayan Nagar, Anat Shoshani, and Ora Nakash. The association between deaf identity and emotional distress among adolescents. *School for Social Work: Faculty Publications*, May 2020. [Online; accessed 14-September-2025].

[7] World Health Organization. World report on hearing, March 2021. [Online; accessed 14-September-2025].

[8] World Health Organization. Deafness and hearing loss. https://www.who.int/health-topics/hearing-loss, February 2025. [Online; accessed 14-September-2025].

[9] Rhoda Olkin. Could you hold the door for me? including disability in diversity. *Cultural Diversity & Ethnic Minority Psychology*, 8(2):130–137, 2002. [Online; accessed 14-September-2025].

[10] Dr. Naa Adzoa Adzeley Boi-Dsane. Being understood: How to expand sign language access for the deaf community, September 2024. [Online; accessed 14-September-2025].

[11] Athulya Sundaresan Geetha. What is yolov6? a deep insight into the object detection model, 2024.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[13] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019.

[14] M. Buttar, A. Kumar, and K. Singh. Deep learning in sign language recognition: A hybrid approach for the recognition of static and dynamic signs. *Mathematics*, 11(17):3729, 2023.

[15] Z. Zhang, S. Liu, X. Wang, W. Wang, Z.-J. Zha, and J. Sun. Meta-detr: Few-shot object detection via unified image-level meta-learning, 2022.

[16] H. Zhang, Y. Wang, and Y. Gong. Accurate few-shot object detection with support-

query mutual guidance and hybrid loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12890–12899, 2021.

[17] P. Shenoy, R. Ganesan, and R. Varma. Real-time indian sign language recognition using deep learning and smartphone camera, 2021.

[18] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *CoRR*, abs/2102.12122, 2021.

[19] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[20] T. Kondo, Y. Tan, and K. Matsumoto. A performance comparison of japanese sign language recognition with vit and cnn using angular features. *Electronics*, 14(8):3228, 2024.

[21] R. D. A. Daniels, A. Mulyana, and A. Widodo. Real-time bisindo sign language recognition system using yolov3. In *IOP Conference Series: Materials Science and Engineering*, volume 1077, page 012029, 2021.

[22] Qiang Chen, Xiangbo Su, Xinyu Zhang, Jian Wang, Jiahui Chen, Yunpeng Shen, Chuchu Han, et al. Lw-detr: A transformer replacement to yolo for real-time detection. *arXiv preprint arXiv:2406.03459*, 2024.

[23] X. Dong, Y. Zhang, J. He, and W. Sun. Incremental-detr: Incremental few-shot object detection via self-supervised learning and fine-tuning, 2022. arXiv preprint.

[24] Sharvani Srivastava, Sudhakar Singh, Pooja, and Shiv Prakash. Continuous sign language recognition system using deep learning with mediapipe holistic. *Wireless Personal Communications*, 137:1455–1468, 2024.

[25] Khmer Post Asia. Khmer sign language learning made available online, March 2020. [Online; accessed 14-September-2025].

[26] Peak Team, Ministry of Education Youth, and Sport Cambodia. Ksl – khmer sign language. `https://play.google.com/store/apps/details?id=com.peak.moeys.rti.ksl&hl=en`, 2025. Accessed: 2025-09-14.