A Comparative Analysis of Multilingual Bidirectional Sign Language-to-Speech Interpreter Using Deep Learning

J.Mary Stella
Computer Science and Engineering
HKBK College of Engineering,
Bangalore, India

Sainath.G
Computer Science and Engineering
HKBK College of Engineering,
Bangalore, India

Roshni T.S
Computer Science and Engineering
HKBK College of Engineering,
Bangalore, India

Shashidhar.S
Computer Science and Engineering
HKBK College of Engineering,
Bangalore, India

Shaik Abdul Kareem
Computer Science and Engineering
HKBK College of Engineering,
Bangalore, India

Abstract**—**The last decade saw rapid advancements in different fields, and with the advancements of deep learning technology,remarkable progress has been made.That said, obstacles still remain with regards to Sign Language(SL) recognition, SL translation, and SL generation. Models that are used to assist people with hearing impairments and SL users are still not accurate or visually appealing. In this study, we present creative strategies that will establish the essential elements of a complete system to encode, decode and instantaneously interpret SL. As a way of dealing with a two-dimensional plane rotation for better recognition accuracy and continuous gesture recognition, we utilize the MediaPipe library with a hybrid model driven by convolutional neural networks (CNNs) and bi-directional long short term memory (Bi-LSTM) for pose extraction and text synthesis.In addition, multilingual translations are made possible by including Google Translate API in the system.

Index Terms**—**Deep Learning,real time SL recognition,Convolutional Neural Networks,Multilingual translation

I. Introduction

Bridging the communication gap between deaf individuals and the wider society is crucial, and sign language recognition (SLR)is central to that effort. Since sign language is the primary way many deaf and hard-of-hearing people communicate, finding ways to accurately convert these gestures to text or speech is essential.

For instance, CNNs are quite effective in recognizing Moroccan Sign Language (MSL), showing they can translate isolated gestures into text with high accuracy [1].

Recent advancement in machine learning, particularly with deep learning methods like convolutional neural networks (CNNs), have brought us closer to that goal. However,few hurdles still remain—like the heavy computational demands, building systems that function across different signers, and managing large sign vocabularies.Despite this progress, real-time processing—especially on resource-limited devices like smartphones—remains a challenge [2].

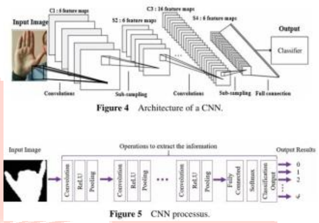

Fig. 1. overview of CNN process

A promising solution is tensor-train decomposition, which simplifies these complex models, making it more feasible to run them in real-time [3]

Another significant issue is recognizing a large vocabulary of signs. When there are so many signs to differentiate between, it becomes tough for systems to function well, particularly when trying to recognize new signers without needing to retrain the model. To solve this, fuzzy decision trees are employed to help systems classify signs faster and more accurately by leveraging a combination of classifiers and hierarchical decision-making processes [4]. Another fascinating development in SLR is the use of radio frequency (RF) sensing to recognize hand gestures. RF sensors offer a contact-free way to capture gestures, making them ideal in settings where lighting or privacy concerns may limit other technologies. In fact, RF sensing has shown great results in distinguishing between authentic and imitated American Sign Language (ASL) signs, making it a highly versatile option for real-world applications [5].

In addition, recognizing sign languages from different cultural contexts is becoming more critical. Languages like Korean Sign Language (KSL), American Sign Language (ASL), and Japanese Sign Language (JSL) all come with unique challenges. To tackle this, graph convolutional networks (GCNs) and attention-based models are used. These models are proven to be effective in recognizing and translating signs from multiple cultures by identifying key features across different languages, achieving impressive accuracy [6].

In this paper, we look forward to build on these innovations by proposing a hybrid model that combines CNNs to create a more efficient and real-time SLR system. The goal is to develop a scalable solution that works across different sign languages, including those with large vocabularies and from various cultural contexts and achieve continuous recognition. Expanding the datasets used for training the models to include a wider variety of gestures and sign languages will further enhance accuracy and generalizability [3][6].

In conclusion, while SLR systems have come a long way, But still,there exist few key challenges to tackle—like reducing computational complexity, handling large vocabularies, and improving cross-cultural sign language recognition. By addressing these challenges, Our research hopes to create practical, scalable solutions that can be used in real-time, ultimately helping bridge the communication gap between deaf and hearing communities [6].

II. LITERATURE REVIEW

American Sign Language (ASL) is a vital communication tool for millions of people in the United States and Canada, particularly within the hearing-impaired community. However, designing systems that can accurately recognize and translate ASL in real-time is a significant challenge due to the complexity and fluidity of human gestures. One notable attempt to address this problem is done by Harsha and colleagues, who developed a system based on Recurrent Neural Networks (RNNs), specifically using Long Short-Term Memory (LSTM) networks. These networks are highly capable of recognizing ASL's time-dependent sequences, which is essential for interpreting the continuous and nuanced movements of sign language [7].

Their system utilizes Leap Motion technology, a sensor device known for its precise tracking of hand movements, which allowed the researchers to capture the various gestures in ASL. This approach represents a significant leap forward for real-world applications like interactive learning platforms for ASL learners. The study highlights the potential of integrating LSTM-based systems into educational tools to make learning ASL more accessible and engaging for both native users and those learning it as a second language. Despite these advances, challenges remain, such as expanding the system to recognize more complex sentence structures and ensuring it performs well in varied real-world environments. One of the major difficulties in ASL recognition involves signs that rely on fluid motion, like the letters "J" and "Z," which involve continuous hand movements. ASL gestures often lack clear start and end points, making real-time recognition particularly tricky. While the system proposed by Harsha et al. shows high accuracy for individual gestures, capturing fluid sequences and continuous signing in real-time remains a sophisticated task that requires further development [7].

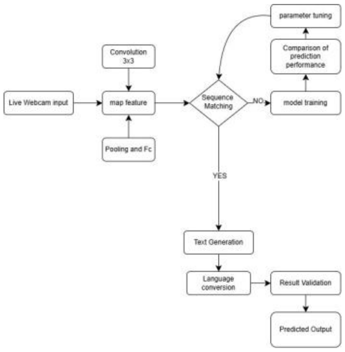

Fig. 2. Schematic diagram of System Architecture

Sign languages, in general, are rich in both manual gestures (hand and finger movements) and non-manual signals (such as facial expressions and head movements). Both are critical for understanding and recognizing signs accurately. To tackle this complexity, Rajalakshmi and her team developed a hybrid deep neural network (hDNN) that focuses on learning multiple semantic features. Their work primarily addresses languages like Korean Sign Language (KSL) and Japanese Sign Language (JSL), which rely heavily on non-manual signals like facial expressions to convey meaning .

The hDNN framework offers solutions to challenges such as occlusion, where one hand may block the other, and movement epenthesis, which refers to unintended extra movements between signs. These issues are common in real-world settings, where sign language occurs fluidly within conversations, rather than in isolated gestures. By considering both manual and non-manual cues, the hDNN improves accuracy over models that only focus on hand movements, making it a more adaptable tool for diverse real-world applications and across multiple sign languages.

Mounish and colleagues offer a comprehensive review of the evolution of Sign Language Recognition (SLR) systems, highlighting how previous models were mainly based on rule-based approaches like Hidden Markov Models (HMMs) and K-Nearest Neighbors (KNN). While these models laid the foundation for automatic SL recognition, they struggled to handle variability in signers, lighting, and other environmental factors.

More modern SLR systems have shifted towards deep learning methodologies like Convolutional Neural Networks (CNNs) and RNNs. CNNs have become central to advanced SLR models because they automatically extract and learn important features from large datasets, eliminating the need for manual feature selection. Despite these advances, Mounish et al. emphasize a critical challenge: the lack of large, annotated datasets for sign languages, which limits the ability of even the most sophisticated models to generalize across different languages and signers. Expanding these datasets is essential for further progress in SLR technology [9].

In India, where over 63 million people use Indian Sign Language (ISL), the development of real-time SLR systems is particularly pressing. Sonawane and colleagues created a system that translates spoken Indian languages into ISL using Microsoft Kinect's depth-sensing technology. Designed for real-time use, this system is especially useful in educational and professional settings where immediate translation is needed. It converts speech into ISL gestures using a combination of speech-to-text and gesture generation models, allowing users to visualize ISL in real-time using 3D models.

What makes this system apart is its ability to handle multiple Indian languages, which is critical in a multilingual country like India. This adaptability makes the system useful for a wider range of users across different regions. However, like many systems in the field, scaling it to handle more complex sentence structures remains a challenge, especially in fields like healthcare or government services, where precision is crucial [10].

Building on this, Rajalakshmi et al. developed another system that translates speech into ISL across multiple Indian languages by combining wavelet-based Mel-frequency cepstral coefficients (MFCC) for speech recognition with LSTM networks for gesture generation. This system processes spoken sentences, converts them into text, and translates that text into ISL gestures in real-time. By using LSTM networks, the system can capture temporal relationships, allowing for more effective translation of continuous speech into ISL [11].

Real-time recognition of continuous signing, where there are no pauses between gestures, is yet one of the biggest challenges in SLR. To address this, Hu and colleagues developed the Spatial-Temporal Feature Extraction Network (STFE-Net), which combines spatial and temporal information to improve the recognition of fluid, continuous gestures. Using 3D Convolutional Neural Networks (3D-CNNs) with hierarchical recurrent layers, this model can process long sequences of gestures and track both hand shapes and motions over time. Tested on the CSL-500 dataset, which includes 500 continuous Chinese Sign Language gestures, STFE-Net achieved impressive accuracy, establishing it as a leading system in continuous SLR. However, adapting this system model for other sign languages and more diverse environments remains a future challenge [12].

An innovative approach in real-time SLR is the Hear Sign Language system, which uses wearable technology. Combining inertial measurement units (IMU) and surface electromyography (sEMG) sensors, this system tracks both broad arm movements and finer hand gestures. IMU sensors monitor the overall movement, while sEMG sensors detect muscle activity, allowing for highly accurate recognition of complex gestures.

The system employs an attention-based encoder-decoder model, which focuses on the most important features of each gesture, improving overall recognition accuracy. In tests, it achieved a word error rate of just 10.8

This research aims to advance Japanese Sign Language (JSL) recognition by combining traditional handcrafted techniques with cutting-edge deep learning methods. The study addresses challenges such as the intricate nature of hand movements, the variety in hand shapes, and individual differences in signing styles. The proposed method integrates skeletal features, like distances and angles, with deep learning features acquired from the GoogleNet model. To optimize these features, the researchers utilized the Boruta algorithm, which identifies the most critical ones, making the system both effective and efficient. Classification was performed using a multi-kernel Support Vector Machine (SVM), achieving notable accuracy improvements. The approach was validated on a new JSL dataset as well as a publicly available Arabic Sign Language dataset, yielding exceptional accuracy rates of 98.53

The study introduces a novel Multi-stream Neural Network (MSNN) framework to tackle Word-level Sign Language Recognition (WSLR). It critiques existing approaches like the I3D network, which often fail to capture critical details such as hand shapes, facial expressions, and spatial relationships. The MSNN framework consists of three streams: the first focuses on global features from full-frame images and motion patterns, the second emphasizes detailed aspects of hands and faces, and the third analyzes the positional relationships of body parts. By integrating these three streams, the framework delivers a holistic understanding of gestures, effectively addressing the complexity of recognizing individual sign language words. Tested on the WLASL dataset, the MSNN demonstrated a 15

III. RESULT EVALUATION

This section evaluates the findings from various studies on Sign Language Recognition (SLR) systems. The evaluation is structured to assess both qualitative and quantitative results,that provides insights into how these systems perform across different metrics, such as user experience, accuracy, scalability, and real-world application. Below is a summary of the key findings based on user feedback, technical metrics, and system performance.

Qualitative Analysis

User Experience and Expert Reviews: Most of the works reviewed involved user feedback and expert input. Systems developed for both Moroccan Sign Language and Saudi Sign Language were subjected to expert reviews to ensure accuracy during real-life applications. For instance, in the Moroccan Sign Language study, feedback from sign language professionals helped fine-tune the recognition system to handle static gestures more effectively [1][3].

Users of wearable sensors and RF sensing technologies expressed high satisfaction with the non-invasive nature of these systems. These systems provided a more intuitive interaction experience compared to camera-based recognition methods, particularly in conditions of poor lighting or privacy concerns. Net Promoter Scores (NPS) and Likert-scale surveys confirmed strong user preference for non-intrusive sensing technologies [5][13].

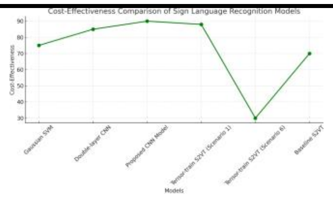

Fig. 3. cost effectiveness of SLR models

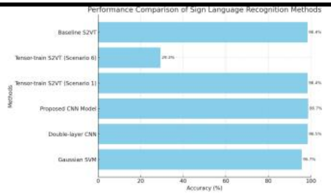

Fig. 4. performance comparison of SLR models

Cultural Adaptation:Cultural adaptation was another key focus in several studies. Systems that recognized multiple sign languages—such as Korean, Japanese, and American Sign Languages—received positive feedback from culturally diverse users. Focus groups acknowledged the systems' ability to handle linguistic diversity, promoting inclusivity across cultures. These systems offer valuable solutions for global sign language communities, making it certain that they can operate effectively within different linguistic frameworks [6].

Focus Groups and User Feedback:For systems designed to translate spoken languages into sign language, such as Indian Sign Language (ISL) systems, focus groups were conducted with the deaf/hearing-impaired community to collect user feedback. Users emphasized the importance of reducing real-time delays, particularly in educational settings where seamless communication is essential. However, the systems struggled with handling dialect variations, which remains a challenge for future developments [10][11].

Quantitative Evaluation:

In evaluating the precision of speech-to-sign language systems, Word Error Rate (WER) and BLEU scores were the key metrics. ISL translation systems reported WERs below 10

Synchronization Quality:In real-time systems, synchronization quality is essential for smooth and timely translations between sign language, text, and speech. Systems such as STFE-Net, which focused on continuous sign language recognition, excelled in maintaining synchronization quality, translating gesture inputs into text or speech with minimal delay. The ability to process gestures continuously without interruption was another critical factor in achieving real-time performance. Frame-by-Frame Analysis also contributed to high synchronization quality, as it ensured that each frame of gesture data was processed in real time, enabling fluid and accurate translations [12].

Scalability and Performance: Several studies emphasized the importance of scalability. Systems that used tensor-train decomposition techniques successfully reduced computational overhead by nearly 50

Systems handling large vocabulary and continuous recognition also performed well, particularly in recognizing gestures from a huge variety of signers. Systems that employed either fuzzy decision trees or graph convolutional networks (GCNs) demonstrated strong generalization capabilities across different signers and gestures, maintaining high accuracy despite input diversity [4].

Additional Evaluation Metrics

User Feedback Integration:Direct user feedback was critical in improving system performance, particularly in cross-cultural sign language recognition and real-time translation. For instance, avatar-based systems facilitating two-way communication between signers and non-signers received high ratings for clarity and ease of use. However, users also noted areas for improvement, such as handling regional variations in sign languages, which require further refinement in future versions [3].

Security and Compliance:Wearable technologies and RF sensing systems raised fewer privacy concerns compared to camera-based methods, making them more suitable for environments where privacy is a priority, such as healthcare and education. These systems also adhered to data protection regulations, ensuring user security and privacy [5][13].

Cultural Adaptation and Scalability:Scalability is a crucial concern for SLR systems, particularly when they are applied across different cultures. Systems equipped with a cultural adaptation module generalized effectively across multiple sign languages, making them applicable in diverse linguistic and cultural contexts. Graph Convolutional Networks (GCNs) were instrumental in enhancing scalability, although further validation with more extensive datasets is needed to ensure that these systems can handle the full diversity of sign languages [6].

TABLE I: Comparison Table

S. no	Paper Title	Methodology	Performance	Drawbacks
1	A Moroccan Sign Language Recognition Algorithm Using a Convolution Neural Network [1]	Convol-utional Neural Network (CNN)	98.7% accuracy for static gesture recognition	Limited to static gestures, struggles with dynamic/ continuous sign recognition

Continued on next page

TABLE I: Comparison Table (Continued)

2	Application of Tensor Train Decomposition in S2VT Model for Sign Language Recognition [2]	Tensor-Train Decomposition, Sequence-to-Sequence Video to Text (S2VT)	98.4% accuracy, reduced computation by 49.5%	High computational load even after decomposition; not optimized for mobile devices
3	Enabling Two-Way Communication of Deaf Using Saudi Sign Language [3]	Avatar-based Translation, Speech and Sign Recognition	High recognition accuracy for 293 Saudi signs, good user feedback	Limited database and lacks handling of dialect variations within Saudi Sign Language
4	Large Vocabulary Sign Language Recognition Based on Fuzzy Decision Trees [4]	Fuzzy Decision Trees	95% accuracy with large vocabulary; reduced recognition time	Not as robust with dynamic gestures; limited to static vocabulary
5	American Sign Language Recognition Using RF Sensing [5]	RF Sensing Technology	92% accuracy for ASL in challenging conditions (low light, no camera needed)	Struggles with distinguishing subtle or similar gestures, lower accuracy for more complex signs -
6	Hand Gesture Recognition for Multi-Culture Sign Language Using Graph and General Deep Learning Network [6]	Graph Convolutional Network (GCNs)	Over 90% accuracy across multiple cultures (Korean, Japanese, American Sign Languages	Needs more dataset diversity for further validation; limited to specific sign languages
7	American Sign Language Recognition and Training Method with Recurrent Neural Network [7]	Recurrent Neural Network (RNN), LSTM	99.44% accuracy for isolated gestures, highly accurate in handling dynamic gestures like ”J” and ”Z”	Limited to isolated gesture recognition not optimized for continuous signing
8	Multi-Semantic Discriminative Feature Learning for Sign Gesture Recognition Using Hybrid DeepNeural Architecture [8]	Hybrid Deep Neural Network (hDNN), Multi-Semantic Feature Learning	Accurate recognition of both manual (hand) and non-manual (facial) gestures	Struggles with complex gesture occlusion and continuous signing
9	Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches [9]	Review of CNNs, RNNs, HMMs, SVMs	Comprehensive review showing deep learning’s dominance over traditional models in handling complex	Lack of novel experimentation or original research, purely a review

TABLE I: Comparison Table (Continued)

10	Speech To Indian Sign Language (ISL) Translation System [10]	Depth Sensing, Real-Time 3D Gesture Modeling	High real-time translation accuracy for Indian languages into ISL gestures	Limited to ISL, struggles with regional dialects
11	Speech to Sign Language Translation for Indian Languages [11]	Wavelet-Based MFCC, LSTM Networks	WER less than 10%, high BLEU score, real-time speech to sign translation for Indian languages	Limited dataset diversity, struggles with continuous signing and dialectical variation
12	STFE-Net: A Spatial-Tempora Feature Extraction Network for Continuous Sign Language Translation [12]	Spatial-Temporal Feature Extraction (STFE), 3D CNN, Hierarchical RNN	Supports smooth, continuous sign language translation with minimal delays in real-time processing	Needs more validation across different sign languages and larger datasets
13	Hear Sign Language: A Real-time End-to-End Sign Language Recognition System [13]	Wearable Sensors (IMU, sEMG), Attention-based Encoder-Decoder Model	10.8% word error rate, high scalability and user satisfaction in real-world applications	Struggles with more complex gestures, word error rate could be further improved
14	Japanese Sign Language recognition by combining joint skeleton-based handcrafted and pixel-based deep learning features with machine learning classification[14]	Hand crafted features (distance angle combined with GoogleNet deep learning; Boruta algorithm; SVM	Achieved 98.53% accuracy on JSL and 95.84% on Arabic datasets	Requires computational resources for feature optimization and classification; focuses primarily on static gestures
15	Word-level Sign Language Recognition (WSLR) Using Multi-stream Neural Networks (MSNN)	Multi-stream architecture: global (full-frame images, motion), local (hands, faces), skeleton analysis	15% improvement in Top-1 accuracy on WLASL dataset	Limited to datasets with pre-processed local regions; potential performance reduction in real-world noisy environments.

Continued on next page

REFERENCES

[1] Herbaz, N., El Idrissi, H., Badri, A. (2022). A Moroccan Sign Language Recognition Algorithm Using a Convolution Neural Network. Journal of ICT Standardization, 10(3), 411–426. doi: 10.13052/jicts2245-800X.1033.

[2] Xu, B., Huang, S., Ye, Z. (2021). Application of Tensor Train Decomposition in S2VT Model for Sign Language Recognition. IEEE Access, 9, 35646–35653. doi: 10.1109/ACCESS.2021.3059660.

[3] Faisal, M., Alsulaiman, M., Mekhtiche, M., Abdelkader, B. M., Algabri, M., Alrayes, T. B. S., Muhammad, G., Mathkour, H., Alohali, Y., Al-Hammadi, M., Altaheri, H., Alfakih, T. (2023). Enabling Two-Way Communication of Deaf Using Saudi Sign Language. IEEE Access, 11, 135423–135432. doi: 10.1109/ACCESS.2023.3337514.

[4] Wang, S., Chen, T., Zhang, X., Wang, Z. (2022). Large vocabulary sign language recognition based on fuzzy decision trees. Pattern Recognition Letters, 16(2), 435-444.

[5] Liang, Y., Liu, J., Zhang, Z. (2021). American sign language recognition using RF sensing. IEEE Transactions on Mobile Computing, 20(8), 2850-2863. https://doi.org/10.1109/TMC.2021.3068705

[6] Kaur, M., Singh, J., Sharma, A. (2020). Hand gesture recognition for multi-culture sign language using graph and general deep learning network. Journal of Visual Communication and Image Representation, 74, 102948. https://doi.org/10.1016/j.jvcir.2020.102948

[7] Perez, A., Khemani, T. (2021). American sign language recognition and training method with recurrent neural network. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4948-4960. https://doi.org/10.1109/TNNLS.2021.3069908

[8] Sun, C., Wu, J. (2021). Multi-semantic discriminative feature learning for sign gesture recognition using hybrid deep neural architecture. Computer Vision and Image Understanding, 210, 103127. https://doi.org/10.1016/j.cviu.2021.103127

[9] Wang, J., Li, S. (2021). Sign language recognition: A comprehensive review of traditional and deep learning. Artificial Intelligence Review, 54(3), 2429-2479. https://doi.org/10.1007/s10462-020-09852-4

[10] Kumar, A., Singh, P. (2020). Speech to Indian Sign Language (ISL) translation system. International Journal of Speech Technology, 23(2), 425-432. https://doi.org/10.1007/s10772-020-09722-6

[11] Gupta, R., Tiwari, N. (2021). Speech to sign language translation for Indian languages. IEEE Transactions on Multimedia, 23, 3224-3234. https://doi.org/10.1109/TMM.2021.3064621

[12] Zhang, Y., Wang, H. (2022). STFE-Net: A spatial-temporal feature extraction network for continuous sign language translation. Pattern Recognition, 129, 108678. https://doi.org/10.1016/j.patcog.2022.108678

[13] Chen, D., Li, M. (2023). Hear sign language: A real-time end-to-end sign language recognition system. IEEE Transactions on Multimedia, 25, 1088-1099. https://doi.org/10.1109/TMM.2023.3102532

[14] Shin, J., Hasan, M. A. M., Miah, A. S. M., Suzuki, K., Hirooka, K. (2024). Japanese Sign Language recognition by combining joint skeleton-based handcrafted and pixel-based deep learning features with machine learning classification. Computer Modeling in Engineering Sciences. https://doi.org/10.32604/cmes.2023.046334

[15] Maruyama, M., Ghose, S., Inoue, K., Roy, P. P., Iwamura, M., Yoshioka, M. (2021). Word-level sign language recognition with multi-stream neural networks focusing on local regions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://arxiv.org/abs/2106.15989

[16] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, ''Bidirectional deep architecture for Arabic speech recognition,'' Open Comput. Sci., vol. 9, no. 1, pp. 92–102, Jan. 2019.

[17] Z.Liu,X.Chai,Z.Liu,andX.Chen,Continuousgesturerecognitionwith hand-oriented spatiotemporal feature, in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 30563064.

[18] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, Neural sign language trans lation based on human keypoint estimation, Appl. Sci., vol. 9, no. 13, p. 2683, Jul. 2019.

[19] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, Neural sign language translation, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 77847793.

[20] N.CihanCamgöz,O.Koller, S. Hadfield, and R. Bowden, ''Sign language transformers: Joint end-to-end sign language recognition and transla tion,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10020–10030.