Two-Way Malaysian Sign Language Communication System for Inclusive Education

Veron Zhen Liang Hii[†*], Aaron Ken Kiat Lo[†*], Ida Pei Xin Lee[†*],
Alec Vince Gonzales Abuan[†*], Sue Han Lee[†], Patrick Hang Hui Then[†]
[†]Swinburne University of Technology Sarawak Campus, Malaysia
E-mail: 102768326@students.swinburne.edu.my, 102768452@students.swinburne.edu.my, 102766016@students.swinburne.edu.my,
105456145@swinburne.edu.my, shlee@swinburne.edu.my, pthen@swinburne.edu.my

***Abstract*—Inclusive education aims to create equal learning opportunities for all, but significant gaps still exist, particularly for the deaf community. This paper addresses these gaps by proposing a new educational platform that integrates cutting-edge technology to improve accessibility and engagement for deaf learners. Our solution introduces an AI-powered two-way sign language communication system specifically designed for integration into classrooms. With an avatar-based approach, the system focuses on simple technology transfer for Sign Language Production (SLP) and utilises advanced deep learning methods for Sign Language Recognition (SLR). This enables seamless and effective communication between deaf students and educators. To our knowledge, this is the first comprehensive approach to digital inclusion in inclusive learning that primarily addresses the specific needs of the deaf community. As part of our initiative, we have created the first Malaysian Sign Language (BIM) education dataset to serve as a benchmark in this area. A new user testing framework has also been developed to quantify the effectiveness of the system in an educational context. The results of the survey emphasise the critical importance and necessity of such an educational platform.**

I. Introduction

Inclusive education aims to remove barriers for all children and ensure access to education, active engagement and optimal academic and social outcomes [1]. Malaysia has been working towards this goal with initiatives such as the Zero Reject Policy [2]. However, the Malaysian deaf community still faces major challenges in accessing equal education. People who are deaf or hard of hearing communicate primarily through the use of the Malaysian Sign Language (BIM), which relies on body gestures such as hands, arms and facial expressions. Studies emphasise that the main barrier for deaf students in mainstream classrooms is linguistic rather than physical [3]. As a result, they often find it difficult to participate fully in a society where spoken language is predominantly used. Inclusive education policies that aim to enrol all children in mainstream schools cannot be applied directly to deaf learners. One promising approach is to integrate sign language communication into schools and to provide sign language translation tools facilitated by widely available assistive technologies.

Assistive technologies related to sign language translation aids have been investigated in recent studies [4]. These projects have proven to be successful as they have received positive feedback from the deaf community on the improvements in communication and engagement. However, they face several challenges that prevent them from being utilised in educational or communication settings. One major challenge is the high cost associated with these technologies, making them prohibitively expensive for classroom use, especially where budgets are limited. Additionally, many existing tools are still at an early stage of development and not yet publicly available. Importantly, these tools often do not support two-way communication, relying instead on translators to facilitate interaction, which limits their effectiveness in real-time communication scenarios.

To overcome these challenges, a robust two-way communication system is essential. As not all deaf people rely on spoken language, it is crucial to provide a method of converting spoken language into sign language. Conversely, it is equally important to translate sign language into spoken or written language for hearing people. As the existing systems currently do not provide support for BIM, our goal is to create an inclusive education system that utilises BIM for all deaf learners and educators in Malaysia.

To achieve this, we propose an AI-powered two-way communication system for BIM, which includes two main approaches: 1) An avatar-based approach for Sign Language Production (SLP) and 2) A deep learning approach for Sign Language Recognition (SLR).

The main contributions of our work are as follows:

1) We proposed a novel educational platform to improve inclusive education for the deaf community. This platform focuses on BIM and can be extended to cover different sign languages.
2) We introduced new digital inclusion strategies that integrate the two-way communication system in sign language. This innovative system uses an avatar-based approach for SLP and utilises deep learning methods for SLR.
3) We set a new benchmark by introducing a BIM dataset for the education sector together with a novel usability testing framework. This sets a new standard for evaluating the usability of educational tools in this field.

[*]These authors contributed equally to this work

The rest of the paper is organised as follows: II provides a literature review of related works on SLP, SLR and AI in education. In III the educational BIM dataset and a self-collected text dataset are presented. The sections IV and V deal with the methodology of SLP and SLR respectively. Section VI presents the analysis of the experiments and the results obtained. Finally, section VII concludes our study.

II. Related works

A. AI for Inclusive Education

Several studies have investigated the integration of AI in inclusive education [5]. Microsoft Translator, for example, uses a headset worn by the speaker to translate speech into subtitles for deaf learners [6]. Moreover, AI tools such as LIFEisGAME help children with Autism Spectrum Disorders (ASD) to understand facial expressions [6].

However, these tools primarily support one-way communication by translating spoken or written content into subtitles or simplified text. This limitation hinders full engagement as deaf people are unable to express their thoughts, ask questions or participate effectively in discussions without a two-way communication system [7].

B. Sign Language Production (SLP)

In the process known as SLP, descriptions in spoken language are automatically translated into sign language. This is achieved by converting sign language glosses into sign language pose sequences. This section explores the latest techniques in SLP, including NMT, conditional image and video generation and avatar-based approaches.

1) Neural machine translation (NMT) approaches: In NMT, word sequences in whole sentences are modelled and predicted with the help of Artificial Neural Networks (ANNs). Pre-processing tasks such as chunking, recognising sentence boundaries and tokenising words are a particular challenge for sign language due to differences in word order and gloss level compared to spoken language. To overcome this challenge, researchers [8] have used Convolutional Neural Networks (CNNs) together with a seq2seq model to translate spoken language sentences from sign language videos.

2) Conditional image / video generation approaches: Conditional image generation, driven by deep learning, has made progress in creating images or videos based on specific inputs. The combination of VAEs and GANs is promising for SLP as it leverages the reliability of VAEs and the discriminative capabilities of GANs for more lifelike human motion synthesis in videos. Hybrid frameworks integrating these models have been proposed for sign language video generation. However, challenges remain, particularly in the effective control of GANs [9], where issues such as gradient disappearance and mode collapse during training limit pattern diversity and hinder meaningful feature learning.

3) Avatar approaches: Recent SLP methods use animated avatars to produce sign language content online, as the JASigning [1] project shows. JASigning uses transcription languages such as HamNoSys IV-C0a and SiGMLIV-C0b as input for its sign translation model. Kim [10] uses a named entity transformer with avatar layering. However, JASigning's software is outdated and existing methods use proprietary representations that are incompatible with unknown users. To bridge this gap, we have developed a unique JSON-based representation for data and animation scripts.

C. Sign Language Recognition (SLR)

SLR, known as automatically recognises sign languages and translates them into spoken language descriptions. This section examines the latest methods of CNN, RNN and other approaches.

1) CNN and RNN approaches: CNNs and RNNs are widely used in sign language recognition (SLR), with CNNs usually serving as feature extractors and RNNs performing temporal modelling. Hu, Gao, Liu and Feng [11] proposed a correlation network that uses 2D CNN for image-based feature detection, and one-dimensional CNN and BiLSTM for temporal modelling. The 2D-CNN extracts features from each image, and the 1D-CNN and BiLSTM capture temporal correlations between images, improving the network's ability to recognise characters over time.

2) Other approaches: More recent research has explored other approaches to SLR. One study [12] developed a CNN- and transformer-based multi-branch network that combines the transformer's ability to compute long-range dependencies and the CNN's ability to extract local features. This hybrid approach leverages the strengths of CNNs and transformers and improves the network's ability to recognise symbols at different temporal and spatial scales. Another study [13] introduced a novel approach to context-aware continuous sign language recognition using the GAN architecture.

III. Dataset

In this study, we collected a novel educational dataset that contains BIM and focuses on medical contexts [2]. The dataset consists of 1040 videos with 40 unique glosses and 10 sets of text annotations in Malay or Bahasa Malaysia (BM) including English (ENG) translation.

a) Privacy, bias and ethical considerations: In order to protect the privacy and rights of the project participants, we have sent out a declaration of consent with the required information, which was to be read and signed.

b) Signers: There are a total of seven sign language signers in this project, three of whom are experts from Sarawak Society for the Deaf (SSD) and four of whom are non-experts. The three experts claimed to be Deaf, either as fluent or professional sign language speakers. The other non-experts belonged to the university community and were counselled by the sign language experts and received feedback for the recording.

[1] https://vh.cmp.uea.ac.uk/index.php/JASigning

[2] https://arekku21.github.io/MSL-Medical/

c) Recording Pipeline: Prior to recording, a set of 40 unique glosses and 10 sentences relating to medical contexts were agreed in advance and selected by SSD. On the day of recording, signers and participants were informed about the glosses and sentences to ensure consistency. The signers were instructed to begin and end with their hands positioned at the side of their body. Once filming began, we recorded each vocabulary word and phrase three times for each signer. For two of the three recordings of the sentences, a formal approach was to be followed, in which the corpus takes into account all connectives, nouns, pronouns, etc. In the third recording, the informal, practical and more intuitive sign language was used, as recommended by the sign language experts. Although there are differences, the meaning of the sentences remains the same according to the annotations.

d) Green Screen Studio: The *Green Screen Studio* at Swinburne University of Technology Sarawak (SUTS) was the strategically chosen location to shoot the videos. We had the intention that the background can be edited in possible future works. Modern high-definition (HD) webcams were used for the recordings, which were aimed frontally at the participants and shot from the upper body. Both cameras used for the videos had a resolution of 1920x1080 at 30 frames per second (FPS).

e) Dataset Modalities: The dataset consists of dual modality glosses and sentences with corpus in texts and videos.
Gloss. Gloss is a text form that can be used as an intermediate representation for the transcription of signs with spoken language words [14].
Sentence. Sentences that have been transcribed in BM/ENG are constructed in the data set using glosses so that they correspond to the approximate meanings of the sign language.
Video corpus. Video corpus is the recorded videos that contain BIM and focus on medical contexts. The video corpus is pre-processed manually using video editing software with minimal changes. For easier processing, we cut the video replays into their respective video clips. We then organise the clips and add the BM/ENG gloss annotations/rewritten sentences as class labels. This corpus contains the CSV files that contain the file names of the training and test data along with the class labels. We also prepare the video corpus by splitting it into training and test data in a ratio of 80:20 for model training.
Text corpus. Text corpus refers to a self-constructed corpus based on the vocabulary of the education BIM dataset. It is prepared for the text-to-gloss system that has been integrated as part of SLP. It is divided into two parts: Source and Target. The source is written in BM/ENG and serves as input text for the user, while the target consists of a combination of glosses written in BM. It consists of 1348 pairs of source and target data, which differ in length and structure. We have also done some pre-processing on the text corpus by splitting it into training and test sentences in a ratio of 80:20.

IV. SLP METHODOLOGY

The goal of SLP is to generate sign language from the language used in the spoken language community, e.g. producing BIM from BM. In this study, the input would be the user text presented in sentences, while the output would be the sign language presented in an avatar animation. Our SLP framework consists of three stages: Text-to-Gloss, Gloss-to-Animation Mapping and Avatar Animation.

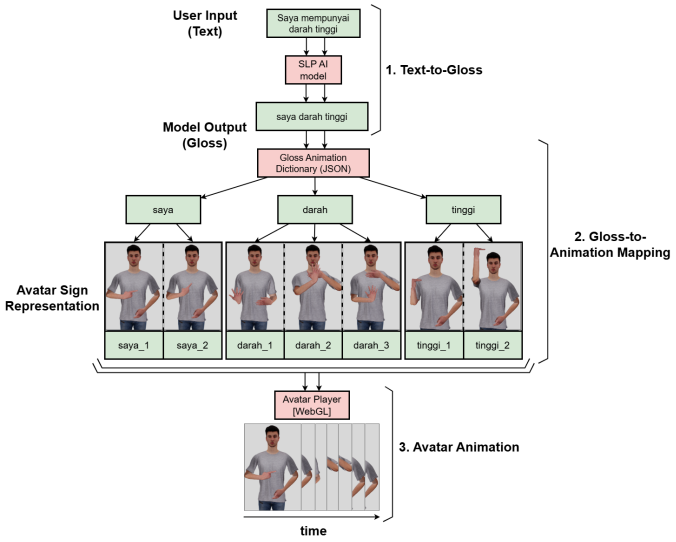

Fig. 1. SLP Methodology

A. Text-to-Gloss

Text-to-Gloss aims to achieve the translation of user text into glosses. In our approach, we use Transformer [15], a deep learning architecture as our model architecture. Transformer consists of two parts, the encoder and the decoder.
Encoder. We chose a pre-trained BERT model trained on a public BM dataset available on GitHub [3].
Decoder. We use three different configurations for the decoder. *Config A* adopts the standard Transformer configuration with 768 hidden sizes, 8 attention heads and 6 hidden layers. *Config B* takes over the configuration of the encoder with a hidden size of 336, a number of 12 attention heads and a number of 4 hidden layers. In addition, we used the pre-trained BERT model as *Config C* of the decoder.

We use NVIDIA RTX3060 to train our models until convergence. The model is trained using the cross-entropy loss function and optimised with Adam optimiser [16]. The initial learning rates are set to 1×10^{-5} and 1×10^{-6}, with weight decay of 1×10^{-7} and a batch size of 32. Specifically, for *Config C*, the learning rate of 1×10^{-6} is used for fine-tuning purposes. During testing, the model is evaluated on the test dataset, which containing unseen sentences from the text corpus. We compare the generated sentences with the ground-truth sentences using metrics such as BLEU, ROUGE-L and accuracy.

B. Gloss-to-animation mapping

Gloss-to-animation mapping aims to map the glosses into their respective animation. We use a JSON file as dictionary to store the mapping between the glosses and their animation.

[3]https://github.com/mesolitica/malaya

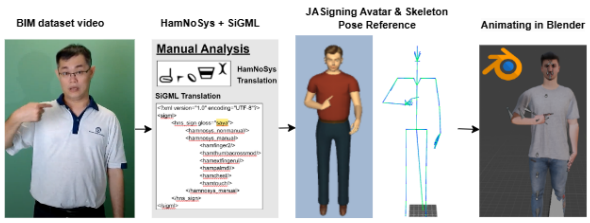

Fig. 2. Flowchart of Avatar Animation Creation: The diagram shows how video data is analyzed to create sign language animations for the avatar.

C. Avatar animation

Avatar animation indicates the movement of 3D avatar to represent the sign language. This is one of the crucial component in our methodology. The avatar animation emphasizes the importance of accurate symbol generation, automated translation, and verification to ensure the fidelity of the resulting sign language animation. To achieve the avatar animation, there are several stages has been researched and processed, which is visualised on figure 2, as follows:

a) HamNoSys dataset analysis: Initially, each sign language video is subjected to careful manual analysis to identify the movements and positions of the hands as well as non-manual features such as facial expressions. The HamNoSys generation tool [4], is employed to transcribe these identified movements into HamNoSys symbols.

b) SiGML translation and verification: The next stage involves converting HamNoSys symbols into SiGML using an automated translation system [5] available on GitHub. This tool maps HamNoSys components to their corresponding SiGML representations, streamlining the conversion process. Subsequently, we perform skeletal model referencing and pose extraction using SiGML in a distinct avatar player to animate our 3D avatar.

c) JSON-based representation: In the final stages, our workflow involves animating the avatar using Blender. The animations are then imported into our WebGL engine within a web-based environment. To facilitate the playback of the avatar animations based on text input, we have developed a custom JSON-based representation for mapping gloss-to-animation.

V. SLR METHODOLOGY

The aim of SLR is to translate sign language into languages used by the spoken language community, e.g. translating BIM into BM. In our methodology, the input consists of sign language represented by BIM videos, while the output is a sequence of text presented in sentences. Our SLR framework consists of two stages: feature extraction, and encoding-decoding.

A. Feature extraction

For feature extraction from the video frames, we use a pre-trained CNN model. For each video, 32 frames are selected at evenly spaced intervals to ensure that the selected frames throughout the videos. Specifically, we use VGG-16 model

[4]https://www.sign-lang.uni-hamburg.de/hamnosys/input/

[5]https://github.com/carolNeves/HamNoSys2SiGML

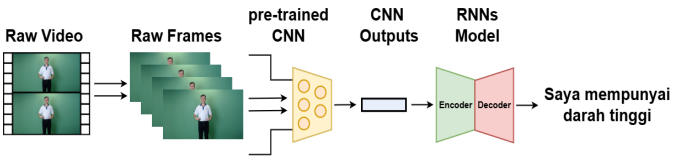

Fig. 3. SLR Methodology

[17] with prior knowledge from ImageNet weights as the pre-trained CNN model. In addition, we exclude the top three layers and freeze all other layers.

B. Encoding-Decoding

We input extracted features into an encoder-decoder model built with stacked RNN layers, specifically using LSTMs and GRUs with three layers each. To prepare the features for encoding, we first apply a fully connected layer to reshape them into a 512-dimensional vector for each frame. These reshaped features are then fed into the encoder, which processes them through its layers to produce a final hidden state. This hidden state is passed to the decoder, along with an initial input vector of 512 zeros for the first decoding step. As the decoder processes this input, it produces outputs that are transformed into probabilities over the vocabulary by a fully connected layer with a LogSoftmax activation function. The word with the highest probability is selected, embedded, and used as the input for subsequent decoding steps, allowing the decoder to generate the sequence of words over time.

The output of the encoder-decoder model consists of sequences of one-hot encoded vectors, each representing a word or special token (such as <BOS>, <EOS>, <PAD>, or <UNK>) from the vocabulary, with each vector having a dimension equal to the vocabulary size.

C. Training & Testing

We train the model using an NVIDIA RTX3060 until convergence, employing the cross-entropy loss function to calculate losses at the sentence level. To emphasize meaningful tokens during training, we assign a weight of 0.5 to the <PAD>token. We use the SGD optimizer with a learning rate of 0.01, a momentum of 0.9, and a batch size of 1. For testing, we evaluate the model on a test dataset consisting of unseen videos from the video corpus. The generated sentences are compared to the ground-truth sentences using metrics such as BLEU, ROUGE-L, and accuracy.

Config	Metrics			
	BLEU-3	BLEU-4	ROUGE-L	Accuracy
A	0.8593	0.7464	0.9580	**0.9889**
B	**0.8694**	**0.7512**	**0.9659**	**0.9889**
C	0.8476	0.7316	0.9483	0.9855

TABLE I
SLP MODEL PERFORMANCE COMPARISON.

VI. EXPERIMENTS

In this section, we present the experiments and results obtained. The goal is to identify the most suitable model and configuration for SLP and SLR by evaluating their performance using various metrics.

A. SLP and SLR Model Performance Analysis

For the evaluation metrics, we utilise the BLEU score [18], ROUGE score [19] and the accuracy rates. In specific, we focused on the performance of model in BLEU-3, BLEU-4, ROUGE-L and accuracy scores on test sets in both video corpus and text corpus.

SLP. We evaluate the text-to-gloss model IV-A. The table I shows that *Config B* performs exceptionally well, achieving the highest score on the metrics BLEU-3, BLEU-4, ROUGE-L and Accuracy. *Config B* performs best in all possible metrics may due to its architectural alignment with the encoder, maintaining the same hidden size of 336. In contrast, *Config C*, despite utilizing a pre-trained model, performs lower than *Config B*, which it might be due to the misalignment between the nature of the target data and the features learned in the pre-trained models.

Model	Metrics			
	BLEU-3	BLEU-4	ROUGE-L	Accuracy
LSTM	0.4645	0.3378	0.8350	0.7010
GRU	**0.5050**	**0.3643**	**0.8692**	**0.7580**

TABLE II
SLR MODEL PERFORMANCE COMPARISON

SLR. We evaluate the encoder-decoder model V-B. As shown in the table II, the *GRU* model performs exceptionally well by achieving the highest score in the metrics BLEU-3, BLEU-4, ROUGE-L and Accuracy Score. This shows that the *GRU* model performs better than the *LSTM* model in all evaluated metrics for SLR, so we used the *GRU* model as the SLR model in our BIM two-way communication system.

B. User Testing

We conducted user testing using a new framework designed to measure the effectiveness of SLP and SLR systems. Our participants, nine in total, were deaf and communicated mainly using BIM. Although their primary means of communication was sign language, most of them were proficient in written BM and some were also proficient in written English and Chinese. This diversity of language backgrounds was taken into account to provide comprehensive feedback on the systems. Participants rated their satisfaction with the systems on predetermined criteria using a scale of 1 to 5.

1) *Hand Accuracy*: How closely do the avatar's hand movements resemble those of real signers?
2) *Facial Expressions*: How realistic are the facial expressions portrayed by the avatar in videos?
3) *Speed*: How well does the speed of the translations affect user experience?
4) *Interactivity*: To what extent does the system facilitate interaction in an educational context?
5) *Complexity*: How complex is it to perform sign-to-text or text-to-sign translations without assistance?

The results shown in Table III indicate slightly higher overall satisfaction for both SLP and SLR systems. Scores range from 1 to 5, where 1 indicates poor performance and 5 indicates excellent performance. The average satisfaction scores for the five key criteria of the SLP system are as follows: *Hand Accuracy* scored 3.67, indicating users found the hand movements accurate and useful; *Facial Expressions* scored 3.11, suggesting room for improvement in conveying emotional nuances; *Speed* received a 3.22, showing moderate satisfaction with responsiveness and real-time performance; and *Interactivity* scored 3.33, indicating satisfactory interaction but highlighting the need for a more seamless and intuitive experience.

Criteria	Systems	
	SLP	SLR
Hand Accuracy	3.67	-
Facial Expressions	3.11	-
Speed	3.22	3.22
Interactivity	3.33	3.33
Complexity	-	3.78

TABLE III
USABILITY SCORES FOR SLP & SLR SYSTEMS
(CRITERIA SCORES RANGING FROM 1 TO 5)

The average satisfaction scores for the three key criteria of the SLR model are as follows: *Complexity* scored the highest at 3.78, reflecting that users found the system relatively straightforward to use independently, without requiring additional guidance; *Interactivity* scored 3.33, reflecting reasonably intuitive and seamless user interaction, with potential for enhanced user-friendliness; and *Speed* received a score of 3.22, suggesting that the recognition speed is somewhat lacking and could benefit from improvements in real-time responsiveness to enhance the fluency and naturalness of communication.

To evaluate the impact of our two-way sign language communication system on promoting inclusive education, we administered a survey featuring binary (yes or no) questions after participants had the opportunity to test the systems themselves. The analysis of the collected survey data revealed that a significant majority of participants, 5 out of 9 (55.6%) acknowledged the potential of our system to facilitate equal educational opportunities for deaf individuals. Moreover, all respondents unanimously agreed that the system could effectively bridge the communication gap between deaf and non-deaf individuals. Such communication is essential for enhancing accessibility within inclusive education systems.

VII. CONCLUSION

The implementation of sign language recognition and production technologies marks a substantial step towards overcoming linguistic barriers that hinder the inclusion of the deaf in mainstream education. Utilizing deep learning models, particularly the Transformer-based architecture for text-to-gloss translation, ensures high accuracy and efficiency in processing sign language data. Additionally, the avatar-based SLP module enhances the system's accessibility and usability. The successful implementation and testing of this system highlight its practicality and potential for wider application, paving the way for more inclusive educational opportunities for deaf students.

However, current limitations include a reliance on a relatively small BIM dataset, which results in low model generalization power and affects performance across diverse sign languages and contexts, indicating room for improvement towards practical application.

Future Work. Future work will focus on expanding the BIM dataset and iteratively improving the deep learning models to enhance performance for both SLP and SLR.

ACKNOWLEDGMENT

We would like to extend special gratitude to SSD for their invaluable support, contributions, and feedback as signers. Appreciation is also given to Alison Dwyne G. Abuan, Zahin Masrur Rahman, Dr. Joel Than Chia Ming and Miss Choo Ai Ling from SUTS for their critical roles in data collection and processing. Additionally, we would like to especially thank Albert Wong Tuong Chui (current Chairman of Sarawak Deaf Community Services Association and former Chairman of SSD) for his deep insights into Deaf culture and BIM.

REFERENCES

[1] E.-J. Hoogerwerf, K. Mavrou, and I. Traina, *The role of assistive technology in fostering inclusive education : strategies and tools to support change*. Routledge, Taylor & Francis Group, 2021.

[2] M. Chin, "The zero reject policy: A way forward for inclusive education in malaysia?" *International Journal of Inclusive Education*, vol. 27, pp. 1–15, Nov. 2020. DOI: 10.1080/13603116.2020.1846800.

[3] T. Tedla and D. Negassa, "The inclusive education for deaf children in primary, secondary and preparatory schools in gondar, ethiopia," *Jurnal Humaniora*, vol. 31, p. 177, Dec. 2019. DOI: 10.22146/jh.44767.

[4] M. N. Osman, K. A. Sedek, Nur, Muhamad, and M. Maghribi, "Hearing assistive technology: Sign language translation application for hearing-impaired communication," *Springer eBooks*, pp. 1–11, Jan. 2020. DOI: 10.1007/978-981-15-3434-8_1. (visited on 07/15/2024).

[5] W. Holmes and I. Tuomi, "State of the art and practice in ai in education," *European Journal of Education*, vol. 57, pp. 542–570, Oct. 2022. DOI: 10.1111/ejed.12533. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/ejed.12533.

[6] S. Garg and S. Sharma, "Impact of artificial intelligence in special need education to promote inclusive pedagogy," *International Journal of Information and Education Technology*, vol. 10, pp. 523–527, 2020. DOI: 10.18178/ijiet.2020.10.7.1418.

[7] N. D. Center, *Importance of effective communication between deaf and hearing individuals*, 2019. [Online]. Available: https://nationaldeafcenter.org/wp-content/uploads/2022/11/Importance-of-Effective-Communication-Between-Deaf-and-Hearing-Individuals.pdf.

[8] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018. DOI: 10.1109/cvpr.2018.00812.

[9] N. Vasani, P. Autee, S. Kalyani, and R. Karani, "Generation of indian sign language by sentence processing and generative adversarial networks," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2020. DOI: 10.1109/iciss49785.2020.9315979.

[10] J.-H. Kim, E. J. Hwang, S. Cho, D. H. Lee, and J. Park, *Sign language production with avatar layering: A critical use case over rare words*, N. Calzolari, F. Béchet, P. Blache, *et al.*, Eds., ACLWeb, Jun. 2022. [Online]. Available: https://aclanthology.org/2022.lrec-1.163 (visited on 07/22/2024).

[11] L. Hu, L. Gao, Z. Liu, and W. Feng, *Continuous sign language recognition with correlation network*, arXiv.org, Mar. 2023. DOI: 10.48550/arXiv.2303.03202. [Online]. Available: https://arxiv.org/abs/2303.03202.

[12] J. Shin, A. S. Musa Miah, M. A. M. Hasan, *et al.*, "Korean sign language recognition using transformer-based deep neural network," *Applied Sciences*, vol. 13, p. 3029, Jan. 2023. DOI: 10.3390/app13053029. [Online]. Available: https://www.mdpi.com/2076-3417/13/5/3029 (visited on 10/21/2023).

[13] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, vol. 21, p. 2437, Apr. 2021. DOI: 10.3390/s21072437.

[14] A. Duarte, S. Palaskar, L. Ventura, *et al.*, "How2sign: A large-scale multimodal dataset for continuous american sign language," *DIGITAL.CSIC (Spanish National Research Council (CSIC))*, Jun. 2021. DOI: 10.1109/cvpr46437.2021.00276.

[15] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, arXiv.org, Dec. 2017. DOI: 10.48550/arXiv.1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762.

[16] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv.org, Dec. 2014. [Online]. Available: https://arxiv.org/abs/1412.6980.

[17] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv.org, Apr. 2015. [Online]. Available: https://arxiv.org/abs/1409.1556.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001. DOI: 10.3115/1073083.1073135. [Online]. Available: https://dl.acm.org/citation.cfm?id=1073135.

[19] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries*, ACLWeb, Jul. 2004. [Online]. Available: https://aclanthology.org/W04-1013.