Universal Sign Language Recognition System Using Gesture Description Generation and Large Language Model

Kanchon Kanti Podder[1], Jian Zhang[2[0000−0003−0813−2350]], and Lingyan Wang[3]

[1] Department of Electrical and Computer Engineering, Kennesaw State University, Marietta, GA 30060, USA kpodder@students.kennesaw.edu
[2] Department of Information Technology, Kennesaw State University, Marietta, GA 30060, USA jianzhang@ieee.org
[3] Department of Computer Science, Kennesaw State University, Marietta, GA 30060, USA lwang40@kennesaw.edu

Abstract. Sign language is a priceless means of communication for deaf and hard-of-hearing people to fully enable them to participate in society and interact with others. This study introduces a novel universal sign language system that uses the Gesture-script to generate a detailed description of gestures in videos, which involve continuous movement of hands, arms, heads, and body language. Subsequently, we input this description into a Large Language Model (LLM) to interpret sign language. We deployed a few-shot prompting technique for LLM, enabling it to precisely transfer the sign videos into corresponding sentences in natural language. Furthermore, the Few-shot prompting technique enables our system to interpret multiple types of sign language without pre-training or fine-tuning.

Keywords: Sign language interpretation · Large Language Model(LLM) · Masked Auto-encoder(MAE) · Few-shot prompting · Gesture

1 Introduction

Sign language communicates information through hand shapes, gestures, and facial expressions using a distinct lexicon and grammar. This communication system goes beyond hand signals and has syntax, semantics, and pragmatics like spoken languages. Sign languages vary by culture and place, with their own vocabulary and organization. This diversity reflects deaf communities' diverse linguistic heritage worldwide. Sign language helps the deaf and hard of hearing communicate, including social inclusion, education, and services, despite its complexity and variety.

Research into technological understanding and recognizing sign language is crucial to breaking down communication barriers between deaf and hearing people. The recent advancements in sign language recognition and translation highlight a broad spectrum of methodologies and challenges encountered in the field.

In literature [11] includes developing a real-time interpreter for BdSL alphabets and numerals with an impressive accuracy of 99.99% using ResNet18 and introducing a comprehensive dataset for BdSL. The authors in literature [10] explored transfer learning and CNNs for BdSL alphabet recognition, emphasizing the importance of advanced image processing in bridging communication gaps for the Deaf and Hard of Hearing (DHH) community. Similarly, the authors in literature [13] proposed a work of a Sign Language Transformer (SLT) model that aims to enhance remote healthcare accessibility, although specific results were not detailed. Another notable contribution is the development of a signer-independent Arabic Sign Language recognition system reported in literature [12], which achieved significant improvement over previous studies with an accuracy of 87.69% using a combination of CNN-LSTM-SelfMLP on segmented datasets.

On the other hand, the authors in study [9] adopted a different approach by using accelerometry and surface electromyography (sEMG) sensors to recognize Colombian Sign Language (LSC), demonstrating the effectiveness of sensor data in capturing the dynamic nature of sign language. The studies [1, 2] introduced innovative approaches through Neural Sign Language Translation (NMT) and a transformer architecture that jointly learns sign language recognition and translation, addressing the challenges of grammar and word order differences. The research by the same authors [3] proposed a SubUNets model that tackles simultaneous alignment and recognition in sign language, offering state-of-the-art hand-shape recognition accuracy. Despite these advancements, limitations persist across studies, including the need for larger, more diverse datasets, the generalizability of models to different sign languages and environments, and the practical application challenges in real-world scenarios. These limitations underscore the necessity for ongoing research and development to enhance the accuracy, robustness, and applicability of sign language recognition and translation technologies. This makes them less scalable and adaptable to new sign languages.

Currently, transformers are increasingly being utilized in many deep learning applications [17, 8, 16], including large language models, and are achieving state-of-the-art results. To overcome this gap, we propose a novel universal sign language recognition and translation system by generating hand gesture descriptions and using the knowledge-based few-shot prompting method with a Large Language Model (LLM). Instead of costly fine-tuning, it only requires a small number of prompting samples to interpret a new sign language video into natural language sentences. The contributions of the proposed research include:

- An image-description paired dataset with hand gesture images and corresponding descriptions to describe gestures.
- A self-supervised encoder augmented architecture to generate hand gesture-related descriptions from images.
- Sequentially generating video descriptions to prompt the LLM in a few-shot manner for sign language interpretation.

2 Method and Materials

The proposed universal sign language recognition system comprises three modules: Multi-modal Image Generator, Gesture-Script module, and Sign Language Interpreter. We illustrate its architecture in Fig. 1. The Multi-modal Image Generator extracts individual frames to generate raw and landmark images from a video demonstrating sign language. The Gesture-Script module generates a comprehensive gesture description for each frame, including details such as hand shapes and hand positions. Based on these gesture descriptions, the Sign Interpreter utilizes the few-shot prompting technology to condition a frozen large language model (LLM) for interpreting the sign language.

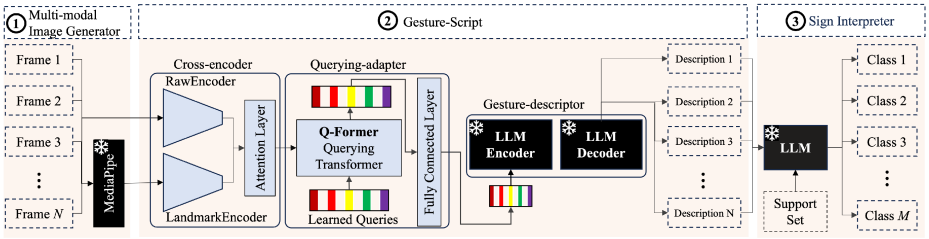


Fig. 1. System overview of our proposed sign language recognition system. Here, the black blocks indicate the frozen models, while the blue blocks denote the learnable models, and the dot edge blocks are the data or related embeddings.

2.1 Multi-modal Image Generator

The Multi-modal Image Generator serves as a pre-processing module to create the Raw Image dataset and the Landmark Image dataset, and it also associates each individual raw and landmark image with their coordinated descriptions to train our subsequent modules.

Raw Image Dataset: It is a custom dataset generated by extracting unique hand gesture frames from the WLASL [5], an open-sourced dataset of sign language-centered videos. The WLASL video dataset comprises 2,000 common American sign words performed by 100 sign-language users(signers). We extract all frames from each video, remove similar hand gesture frames, and retain only those displaying unique hand gestures. In total, we collected 10,232 raw image frames from the WASL dataset.

Landmark Image Dataset: We derived each image from the abovementioned Raw Image Dataset to create the Landmark Image Dataset, which consisted of the same amount of 10,232 images. A landmark image is synthesized with landmarks of hand joints with connecting lines superimposed on a black background

created by MediaPipe [7]. The MediaPipe detects landmarks' positions, (x, y, z), in 3D space from a 2D RGB image. Usually, it collects 66 landmarks (24 upper-body pose and 42 hand landmarks) to represent the hand and pose. We used each landmark's x and y to generate a landmark image, and an example is shown in Fig. 2.

Description associated dataset: Last, we bonded each related raw and landmark image and associated them with coordinated descriptions to create a comprehensive dataset for training the subsequent GestureScript Module. This new dataset can be denoted as $\mathcal{I} = \{(i_0, l_0, d_0), \ldots, (i_N, l_N, d_N)\}$, where i_j represents the raw image, l_j represents the corresponding landmark image, and d_j corresponds to the collected description of the gesture for the j image. Here, N is our dataset's length, comprising over 10,232 pairs of images and corresponding descriptions. A group of volunteers curated the descriptions to describe the gesture of each raw and related landmark image. Fig. 2 provides a visual representation of the dataset, showcasing the raw images, corresponding landmark images, and the accompanying descriptions of the hand gestures. It's important to note that raw and landmark images were sourced from the Raw Image Dataset and Landmark Image Dataset, as previously discussed.

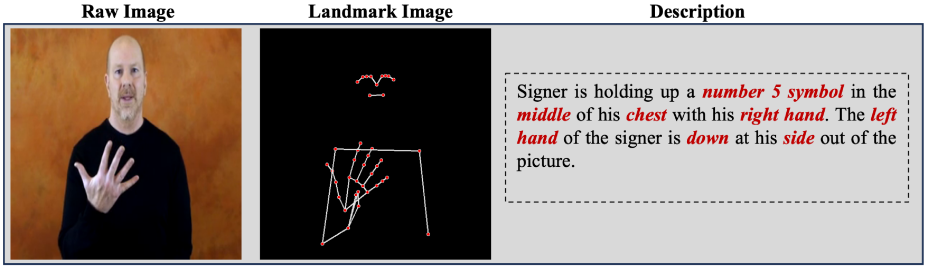


Fig. 2. Sample representation of description associated dataset. The raw and landmark images were obtained from the Raw and Landmark Image Dataset.

2.2 Gesture-script Module

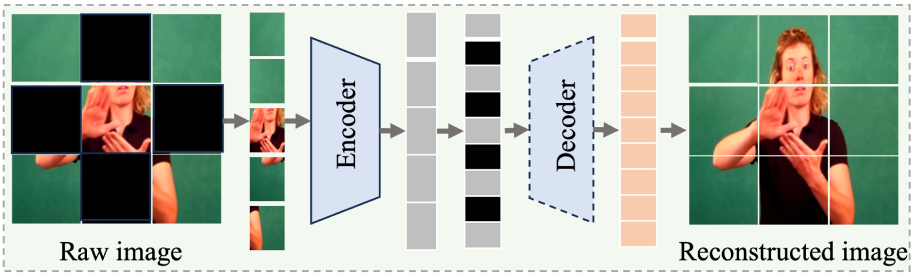
Our Gesture-script module is an image-description generation system that can detect the hand gesture and translate it into a text-based description. Its overall module architecture is illustrated by Blocks 2 in Fig. 1. It consists of three components: Cross-encoder, Querying-adapter, and the Gesture-descriptor. The Cross-encoder comprises two Masked Autoencoders (MAEs) to extract embedding from raw and landmark images, respectively; then, an attention layer is deployed to investigate the underlying relations between two images to generate a cross-image embedding. Querying-adapter is based on a Q-former [6] to adapt these cross-image embeddings to optimize the subsequent Gesture-descriptor, a frozen LLM model, to precisely describe the gesture in the input images.

Cross-encoder: As Fig. 1 shows, our Cross-encoder is based on two MAEs and an attention layer. The MAE is a self-supervised learning algorithm to acquire visual representations from unlabeled data [4]. An MAE learns to reconstruct randomly masked-out images, compelling the model to identify and represent essential image features, even when partially or fully obscured. This learning methodology enables us to train an MAE separately from the main training stream and not require any labeled data, making it a stand-alone and self-supervised process. Based on a pre-trained ViTMAE-Large model [4], we obtained the RawEncoder and the LandmarkEncoder by training two MAEs by Raw Image Dataset and Landmark Image Dataset, respectively.

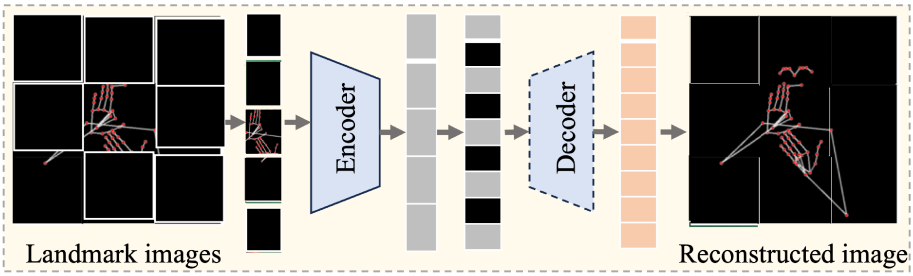
To efficiently capture the hand gesture's shape, position, and orientation, we pre-train these two MAEs using self-supervised learning. Both MAEs are based on a well-trained ViTMAE-Large model with an encoder-decoder architecture, and we use our Raw Image Dataset and Landmark Image Dataset to fine-tune RawMAE and LandmarkMAE, respectively. In this process, the RawMAE and LandmarkMAE models were trained to generate visual representations by reconstructing randomly masked-out raw images and landmark images, respectively, ensuring the models' ability to discern and depict essential features. We highlight this pre-training process in Fig. 3, which enables us to efficiently and effectively learn two MAEs capable of representing raw and landmark images by abstract embeddings. After training, we discarded their decoders to keep only the encoder to form our well-trained RawEncoder and LandmarkEncoder. Then, we deployed a cross-attention layer [15] to learn the latent relations from the outputs of these two encoders to enhance gesture detection and present them as cross-image embeddings.

Querying-adapter: The extracted cross-image embeddings can provide comprehensive information about the given gesture images. However, there is an apparent gap between those embeddings and the frozen Gesture-descriptor module to generate precise text context to depict the gesture. The original LLM-based Gesture-descriptor is trained to describe the overall scenario from the images and will not focus on the gesture details. For example, by giving the images in Fig. 2, the Gesture-descriptor will generate a description as "A person makes a pose in front of the camera.", which is a literaturally correct statement for the images but is meanless for us to interpret the sign-language. We need to enable the Gesture-descriptor to generate a description with details of the hand and gesture, as shown in the right column of Fig. 2. To bridge this gap, we proposed the Querying-adapter to adapt these cross-image embeddings to optimize the subsequent Gesture-descriptor in describing the gesture precisely.

As shown in Fig. 1, the core of the proposed Querying-adapter is a Q-Former, a lightweight transformer that acts as an information bridge between a frozen image encoder and a frozen language model (LLM)[6]. The Q-Former is a trainable module that extracts a fixed number of embeddings from the image encoder regardless of the image resolution. It utilizes learnable query vectors to extract relevant visual features from the image encoder and then transfers the most



(a) Training of RawEncoder



(b) Training of LandmarkEncoder

Fig. 3. The self-supervised training process for our proposed RawEncoder and LandmarkEncoder: They are both trained with their decoder parts that will be discarded later.

useful visual feature to the LLM for text generation. Then, a Fully Connected Layer is deployed to linearly project the query embeddings from the Q-Former to match the dimension of the subsequent LLM-based Gesture-descriptor. We used our description associated dataset $\mathcal{I}$ to fine-tune a Q-Former that was initialized with BERT[14] pre-trained weights and trained our Fully Connected Layer. They were fine-tuned end-to-end with the frozen RawEncoder, frozen LandmarkEncoder, and the subsequent frozen LLM-based Gesture-descriptor. Then, the well-trained Querying-adapter could provide optimized Query embeddings to enable subsequent modules to describe the hand gesture precisely.

Gesture-descriptor: The Gesture-descriptor is a frozen LLM encoder-decoder. It receives the Query embedding from our Querying-adapter and output descriptions to depict the gesture of sign language. These query embeddings serve as soft visual prompts to condition the LLM on the visual representation. These prompts will take advantage of the generative capability of the LLM to enable the Gesture-descriptor to depict the precise details of the gesture, which can be used to interpret the corresponding sign language. A well-trained Gesture-descriptor is capable of generating a meaningful description of the hand gesture in raw and landmark images. For example, it will depict the gesture for a person

using sign language as *"The signer's hands are at chest level, with the right hand below the left. Both palms face the body. ...."*

2.3 Sign Interpreter:

The Sign Interpreter is the final step of our system. It will transfer frame-by-frame gesture descriptions of a sign language video into natural language sentences, facilitated by a few-shot prompt framework, as depicted in Block 3 of Fig. 1. Few-shot prompting is a technique that uses a small number of examples to guide a large language model (LLM) towards generating the desired output. Let $\mathcal{D} = \{v_0, \ldots, v_m\}$ represents our dataset consisting of m videos to represent M unique sign language classes (usually $m \geqq M$), where each video v_i is associated with a sequence of sign gestures. Additionally, we have a set of n natural language sentences $\mathcal{S} = \{s_0, \ldots, s_n\}$, where each sentence s_j corresponds to a specific sequence of signs in $\mathcal{D}$. Our goal is to develop a prompt P that can effectively map the sign language sequences to the corresponding sentences. To achieve this, we create a set of few-shot examples or support set, denoted as $\mathcal{E} = \{(\overline{d_0}, s_0), \ldots, (\overline{d_h}, s_h)\}$, consists of each sign language classes that presented in $\mathcal{D}$. Here, $\overline{d_i}$ is a frame-by-frame description generated from our Gesture-Script for a video, and $s_i \in \mathcal{S}$ is the corresponding natural language sentence. $\mathcal{E}$ is a tiny dataset with a length $h \ll m$, which guides the LLM to generalize and generate coherent natural language representation from a given gesture description $\overline{d_i}$. By giving a support set $\mathcal{S}$, this approach is generalized to any sign language recognition as $\mathcal{S}$ contains $\overline{d_i}$ describes a hand gesture, and s_i represents the corresponding natural language representation in that particular sign language. This enables our proposed model to interpret any sign language video into natural language sentences with a small support set $\mathcal{E}$ instead of computing-costly fine-tuning.

3 Experimental Study

3.1 Experimental Setup

In this Work-in-progress paper, we only conducted experiments to assess the pre-training process of our proposed RawEncoder and LandmarkEncoder. Their performances were evaluated by how precisely the fine-tuned RawMAE and LandmarkMAE can reconstruct the raw and landmark images, respectively. As we introduce in the section 2.2, the RawMAE consists of the RawEncoder and a decoder, and the LandmarkEncoder consists of the LandmarkEncoder and a decoder. Hence, the reconstruction results can reflect the RawEncoder's and LandmarkEncoder's performance in feature extraction. In these experiments, we took 9,227 images from the Raw Image Dataset to train RawMAE and the rest of 1,005 for validation. We designed our experimental setup such that the LandmarkMAE was trained and validated on the same corresponding landmark images as the RAWMAE, that is, 9,227 training images and 1,005 validating

images from the Landmark Image Dataset. By adopting this structured experimental design, we ensured that the LandmarkMAE model followed a similar training and validation regimen as RawMAE.

The experiment was conducted using the PyTorch deep learning framework and Python 3.7. The model underwent training using a computational setup consisting of a 128GB random access memory (RAM) and a 24GB NVIDIA GeForce graphics processing unit (GPU). The combined parameter count for the raw image MAE and skeleton image MAE was 329,541,888. In order to train both RawMAE and LandmarkMAE, a masking ratio of 0.75% was employed. The models underwent training for a total of 1000 epochs, with a batch size of 24.

3.2 Preliminary Results

In our preliminary experiment, We fine-tuned our RawMAE and LandmarkMAE with our data set. The rationale behind fine-tuning the MAEs on raw and landmark images was to obtain more coherent embeddings of hand, face, and skeleton compared to the original pre-trained version of the MAE. We evaluated our fine-tuned RawMAE against the reconstructed result from the original pre-trained ViTMAE-Large. As shown in Fig. 4(a), the reconstructed raw image using the pre-trained ViTMAE-arge is more blended and blurred, and it misses the sharpness and edges of the gesture. The fine-tuned RawMAE trained on our custom dataset preserved the sharpness and edges of the fingers, eyes, mouth, and hand, although it may be limited in reconstructing clothing and background. The reconstruction of the fine-tuned LandmarkMAE also outperformed the pre-trained ViTMAE-Large, as shown in Fig. 4(b).

These preliminary results show that our fine-tuned RawMAE and LandmarkMAE performed better than the original ViTMAE-Large model and prove that our proposed RawEncoder and LandmarkEncoder are both efficient and effective in extracting features to present the hand gesture. In our future work, we will investigate other modules, such as the Gesture-script Module and the Sign Interpreter, and assess the end-to-end performance of our proposed sign language recognition system.

4 Conclusion

This research proposes a universal sign language interpretation system that does not require pre-training or fine-tuning on a specific sign language video dataset. The system only requires a few-shot prompt containing a few frame-by-frame descriptions of gestures generated by the proposed Gesture-script to an LLM for interpreting any sign language sign video. The preliminary results in the work-in-progress show the efficiency and effectiveness of our proposed feature-extracting module and make a promising future for the entire system. We believe that this system will significantly help the deaf and hard-of-hearing community by providing a sign language interpreter to non-sign language users.

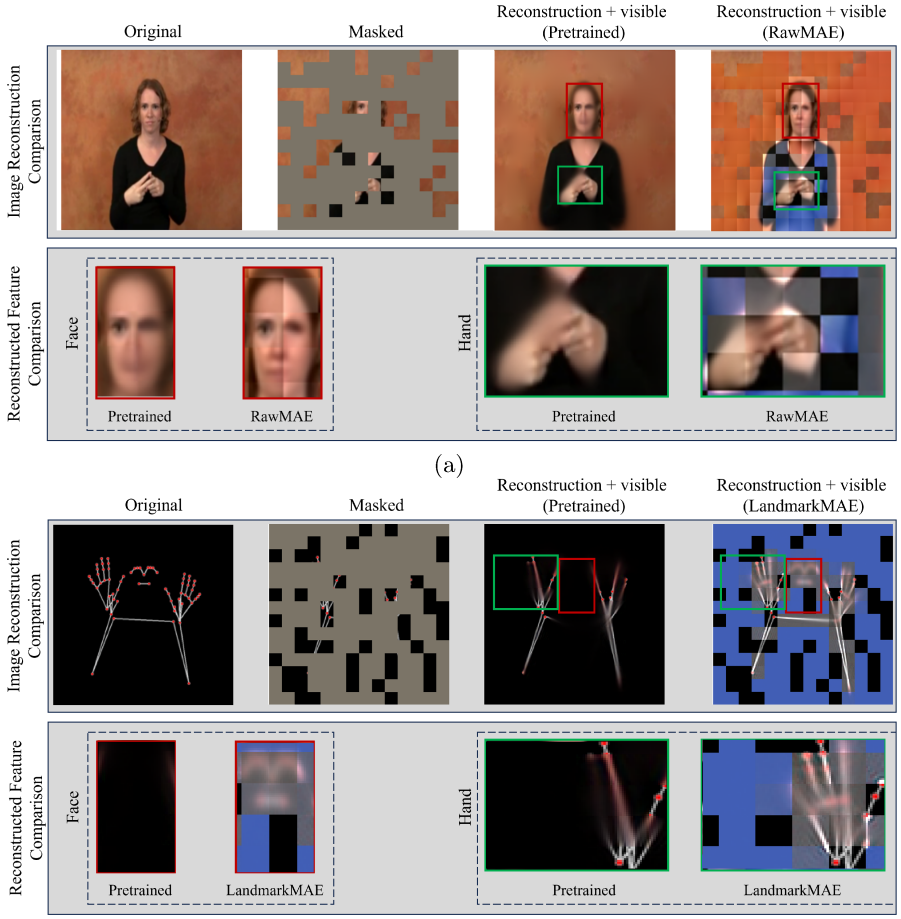


Fig. 4. Reconstruction results comparison: our proposed RawMAE and Landmark-MAE outperform the original Pretrained ViTMAE-Large in retaining the features of gestures, including the edges of the fingers, eyes, mouth, and hand. a).Pretrained ViTMAE-Large vs. RawMAE and b). Pretrained ViTMAE-Large vs. LandmarkMAE.

Acknowledgment

This work is supported in part by the NSF under Grants CCSS-2245607 and CCSS-2245608.

References

1. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
2. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
3. Cihan Camgoz, N., Hadfield, S., Koller, O., Bowden, R.: Subunets: End-to-end hand shape and continuous sign language recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
5. Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison (2020)
6. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
7. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
8. Ma, J., Yang, C., Mao, S., Zhang, J., Periaswamy, S.C., Patton, J.: Human trajectory completion with transformers. In: ICC 2022-IEEE International Conference on Communications. pp. 3346–3351. IEEE (2022)
9. Pereira-Montiel, E., Pérez-Giraldo, E., Mazo, J., Orrego-Metaute, D., Delgado-Trejos, E., Cuesta-Frau, D., Murillo-Escobar, J.: Automatic sign language recognition based on accelerometry and surface electromyography signals: A study for colombian sign language. Biomedical Signal Processing and Control **71**, 103201 (2022)
10. Podder, K.K., Chowdhury, M., Mahbub, Z.B., Kadir, M.: Bangla sign language alphabet recognition using transfer learning based convolutional neural network. Bangladesh J. Sci. Res pp. 31–33 (2020)
11. Podder, K.K., Chowdhury, M.E., Tahir, A.M., Mahbub, Z.B., Khandakar, A., Hossain, M.S., Kadir, M.A.: Bangla sign language (bdsl) alphabets and numerals classification using a deep learning model. Sensors **22**(2), 574 (2022)
12. Podder, K.K., Ezeddin, M., Chowdhury, M.E., Sumon, M.S.I., Tahir, A.M., Ayari, M.A., Dutta, P., Khandakar, A., Mahbub, Z.B., Kadir, M.A.: Signer-independent arabic sign language recognition system using deep learning model. Sensors **23**(16), 7156 (2023)
13. Podder, K.K., Tabassum, S., Khan, L.E., Salam, K.M.A., Maruf, R.I., Ahmed, A.: Design of a sign language transformer to enable the participation of persons with disabilities in remote healthcare systems for ensuring universal healthcare coverage. In: 2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR). pp. 1–6. IEEE (2021)
14. Tenney, I., Das, D., Pavlick, E.: Bert rediscovers the classical nlp pipeline. arXiv preprint arXiv:1905.05950 (2019)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
16. Wang, X., Zhang, J., Mao, S., Periaswamy, S.C., Patton, J.: Locating multiple rfid tags with swin transformer-based rf hologram tensor filtering. In: 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall). pp. 1–2. IEEE (2022)
17. Wu, Y., Zhang, J., Wu, S., Mao, S., Wang, Y.: Cmrm: A cross-modal reasoning model to enable zero-shot imitation learning for robotic rfid inventory in unstructured environments. In: IEEE Global Communications Conference 2023 (2023)