Deep Learning for Sign Language Recognition: Exploring VGG16 and ResNet50 Capabilities

Jatin Sharma[1], Kanwarpartap Singh Gill[1], Mukesh Kumar[2], Ruchira Rawat[3]

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1]

Computer Science & Engineering, Graphic Era Hill University, Dehradun, Uttarakhand, India[2]

Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India[3]

Corresponding author: Kanwarpartap Singh Gill, Email: kanwarpartap.gill@chitkara.edu.in

For the deaf and hard of hearing to communicate with one another, sign language recognition (SLR) is an absolute must. The use of two popular deep learning models—VGG16 and ResNet50—for SLR-related tasks is examined in this study. We achieved remarkable success rates of 99.92% and 99.95% in sign language gesture recognition using the VGG16 and ResNet50 architectures, respectively. We found that these models worked well at accurately reading hand and gesture motions, which made it much easier for people to communicate with one another using sign language. With the use of cutting-edge deep learning methods, our study is making strides towards better SLR systems, which might lead to more accessible and inclusive communication.

Keywords: Sign language recognition, Deep learning, VGG16, ResNet50, Gesture recognition.

2024. In Mukesh Saraswat & Rajani Kumari (eds.), *Applied Intelligence and Computing*, 115–124. Computing & Intelligent Systems, SCRS, India. DOI: https://doi.org/10.56155/978-81-955020-9-7-13

1 Introduction

We provide the groundwork for our study of ASL recognition in the introduction, which stresses the significance of ASL recognition in fostering inclusive communication among the hard of hearing and deaf population. The majority of the world's 70 million deaf individuals communicate primarily using American Sign Language (ASL), which also makes use of facial expressions, body language, and hand gestures. Nevertheless, traditional barriers to communication persist, highlighting the necessity for advancements in ASL recognition technology. Recent advances in the field, enabled by powerful tools for extremely precise sign language gesture interpretation, have been made possible by deep learning techniques. Using the popular ResNet50 and VGG16 architectures, we set out to build and evaluate ASL recognition models in this study. With the use of deep learning, we want to create ASL identification systems that are both reliable and accessible, paving the way for more inclusive communication for those who use sign language.

1.1 VGG16

The 2014 ILSVR(Imagenet) champions utilised VGG16, a CNN architecture, to achieve victory. Many think it's one of the best vision model architectures ever. The most distinctive feature of VGG16 is its usage of a 2x2 filter with a stride of 2 and a maxpool layer with a 3x3 filter with a stride of 1 as opposed to a plethora of hyperparameters. Throughout its design, it adheres to this layout of convolution and max pool layers. Two fully connected layers (FCs) and a softmax output filter make up the final product. It contains sixteen weighted layers, which is what the "16" in VGG16 means. With an estimated 138 million parameters, this network is on the vast side (Figure 1).

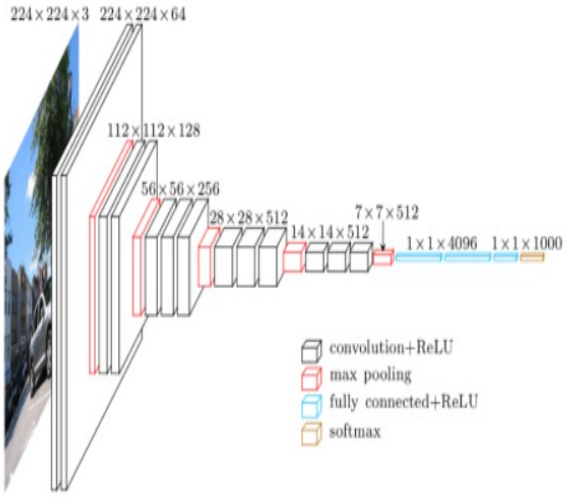

Figure 1. Architecture of VGG16

1.2 Restnet50

As a variant on the RestNet (Residual Network) design, ResNet50 represents a significant advancement in convolutional neural network (CNN) architecture, particularly when it comes to training deeper networks. ResNet50, introduced by Microsoft Research in 2015, is famous for its innovative usage of residual connections that enable the training of very deep neural networks with improved convergence and accuracy. The vanishing gradient problem is effectively addressed by ResNet50 by introducing shortcut connections that bypass numerous levels, making it easier to train networks with hundreds of layers. Object identification, picture classification, and now sign language recognition are just a few examples of how ResNet50 has paved the way for further advancements in computer vision research. ResNet50's capacity to capture complex characteristics and hierarchical representations makes it an attractive foundation for developing powerful ASL recognition models capable of accurately and efficiently interpreting sign language motions (Figure 2).

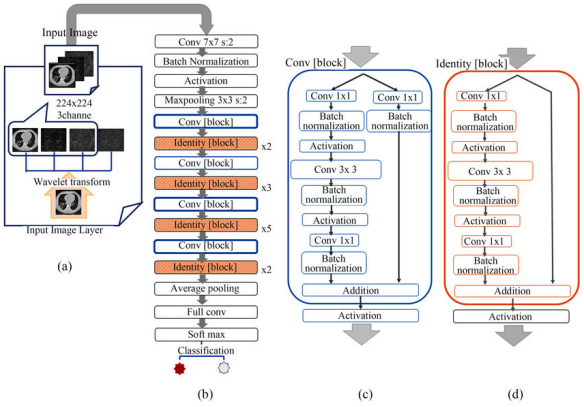

Figure 2. Architecture of RestNet50

2 Literature

Research in the area of sign language recognition (SLR) is crucial if we are to improve communication for the deaf and hard of hearing. Hand gestures, facial expressions, and body postures communicate meaning in sign language, which is a visual-gestural language. Recent advances in machine learning and computer vision have allowed SLR systems to progress considerably, opening the door to new possibilities for overcoming communication barriers.The significance of using high-quality datasets to construct accurate recognition algorithms is emphasised in ASL Citizen, which is presented by Desai et al. [1], an essential addition to studies on SLR. The goal of the community-curated dataset known as ASL Citizen is to enhance the identification of distinct sign languages. To facilitate comprehensive model training and evaluation, ASL Citizen compiles sign language examples from a variety of sources, making it a rich and diverse resource. Song and colleagues [2] provide a new approach to SLR by integrating wearable technology for sign language detection [3]. An organohydrogel-based wearable electronic skin capable of recognising sign language motions in difficult environments is demonstrated in this study. This integration makes it easy for users to use sign language for real-time communication in many different settings. Zhu et al. [4] propose a multiscale temporal network for continuous sign language recognition. Recognising the inherent temporal dynamics of sign language gestures, their approach focuses on capturing both short- and long-term temporal relationships. The model's incorporation of multiscale temporal information enhances recognition accuracy, which is particularly useful in situations involving continuous sign language recognition. Alyami and colleagues [5] tackle the problem of Arabic Sign Language (ASL) isolation recognition using a transformer-based technique.

Their findings highlight the significance of considering language-specific aspects in SLR research [6]. The proposed approach successfully identifies Arabic sign language motions by adapting the model architecture to match the unique features of ASL, including the forms and movements of the hands. In their study, Alves and colleagues [7] looked at the possibility that skeletal picture representation may enhance Libras (Brazilian Sign Language) recognition. Libras sign language is similar to others in that it relies on elaborate physical movements to convey meaning. By employing skeletal image representation techniques, their approach enhances Libras gesture recognition ability by capturing crucial temporal and spatial features. Working with updated picture models, Hu and colleagues [8] [9] [10] aim to improve continuous sign language recognition. They investigate ways to enhance existing image-based models for sign language recognition applications, cognizant of the necessity of model adaptation for practical use. Through the use of domain-specific information and model parameter adjustments, their techniques considerably improve the performance of continuous SLR tasks. An isolated sign language recognition approach that uses finger characteristics derived from posture data was proposed by Akdag and Baykan [11] [12]. They take into account a variety of inputs, such as physical motions and hand gestures, because they know sign language is multimodal. In particular for complex sign language motions, their approach enhances identification accuracy by combining complementary modalities [14] [15]. Prior work by Wadhawan and Kumar [13] [16] includes comprehensive reviews of the literature and deep learning-based SLR systems for both static and dynamic sign motions. Their research shows how deep learning architectures and methods for generating datasets have advanced, and it gives a comprehensive overview of how SLR approaches have evolved [17] [18]. Their work contributes to the advancement of SLR systems by integrating existing information and proposing new approaches. Continuous work in cross-lingual adaptation, real-time recognition, and multimodal integration has progressed SLR research beyond the listed successes. Novel sensor technologies, gesture detection in different cultural contexts, and everyday SLR application development might be future research subjects [19]. The SLR community is working on these solutions in the goal of fostering inclusive communication environments and empowering sign language users throughout the world [20].

3 Input Dataset

Our ASL recognition study makes use of training, testing, and evaluation sets as inputs. The testing set and training set are both retrieved from "/kaggle/input/american-sign-language-recognition/test_set" and "/kaggle/input/american-sign-language-recognition/training_set" respectively. In order to get the data ready, we use a utility function named load_images to resize the images to 64x64 pixels and arrange them next to their labels. Stratify ensures that data is spread evenly among all labels throughout the train-test split method. By splitting the dataset into training and testing subsets, we may better understand its composition. There are 12,071 pictures in the testing set and 48,281 in the training set, totaling 40 unique symbols. For the purpose of validating models and evaluating performance, a collection of 8,000 assessment photographs is also prepared. This dataset, as shown in Figure 3, is used to train and evaluate our ASL recognition algorithms.

```
Total number of symbols:  40
Number of training images:  48281
Number of testing images:  12071
Number of evaluation images:  8000
```

Figure 3. Input Dataset Utilized

4 Proposed Methodology

The proposed method lays forth a systematic approach to developing accurate and efficient ASL recognition models. The first and foremost step is to load the dataset's training, testing, and evaluation sets from the specified folders. This dataset is used to train and evaluate ASL recognition methods. Following data loading, preparation methods get the data ready for model training. By encoding the labels into binary vectors one-hot, the models are able to effectively comprehend category information. This preprocessing step ensures that the data is properly formatted and prepared to be fed into the neural network models. Following this, the models are initialised with an emphasis on two well-known architectures: ResNet50 and VGG16. These models provide distinct architectural and performance advantages; they form the basis of the ASL recognition system. What makes VGG16 unique and allows it to capture photographs with fine details so well are its deep convolutional layers. On the other hand, ResNet50 manages the vanishing gradient problem with residual connections, enabling the training of deeper neural networks. After initialization, the models' architectures are investigated extensively to understand their parameters and underlying structure. You need this knowledge to optimise your model's performance and understand the assessment results. In order to evaluate how well VGG16 and ResNet50 perform, confusion matrices are generated for each model. Featuring an exhaustive rundown of the accuracy of sign language gesture categorization, these matrices highlight potential misclassifications or areas that require improvement. Future optimisation and refining efforts may be guided by knowing the pros and cons of each model using confusion matrix research. This method provides a structured framework for developing and testing ASL recognition models, with the end goal of making sign language communication more accessible. Using extensive testing and analysis, our goal is to construct robust and trustworthy ASL recognition systems using deep learning.

5 Results

5.1 Model Accuracy Comparison

We examined the architectures of VGG16 and ResNet50 deep learning models for ASL recognition. On both the test and evaluation datasets, the VGG16-based model achieved a perfect score of 99.992%. With a score of 100 on the evaluation dataset and 99.95% on the test pictures, the ResNet50 model proved to be very accurate. These outcomes show that, with little changes, both models can effectively identify ASL motions. In general, the high accuracy rates of VGG16 and ResNet50 models indicate that they have the potential to facilitate easy communication for sign language users, leading to improvements in accessibility and inclusivity. As seen in Figures 4 and 5, more research into these models has the potential to produce ASL identification systems that are both more capable and more applicable in the real world.

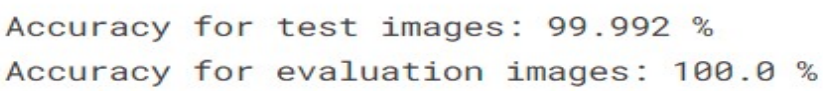

Figure 4. Accuracy Of VGG16

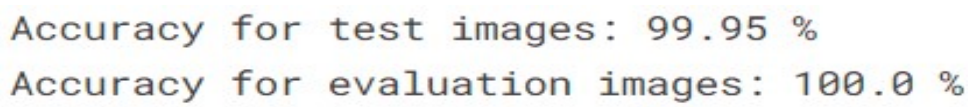

Figure 5. Accuracy Of Resnet50

5.2 VGG16 Accuracy and Loss Plot

The performance of the American Sign Language (ASL) recognition model that was trained and verified using VGG16 across a number of epochs is demonstrated by the accuracy and loss graphs. The accuracy plot provides a visual representation of the performance of the model on both the training dataset and the testing dataset as it is being trained. The fact that the computer becomes better at recognising American Sign Language gestures is demonstrated by the fact that its accuracy improves across both datasets as the number of epochs grows. Models that are able to generalise successfully to unknown input are essential to the development of accurate American Sign Language (ASL) recognition systems. The fact that the training and testing accuracy curves converge provides evidence that this is the case. The loss plot is used to illustrate the loss of the model if there is a rise in the disparity between the predicted and actual labels. The learning process of the model is strengthened by a lowering loss trend since it indicates that the model is reducing the number of prediction mistakes. When it comes to situations that occur in the real world, the dependability of the model is improved since it does not overfit to the training data. This may be observed by seeing that the training and testing loss curves are closely matched. The accuracy and loss plots of the American Sign Language (ASL) recognition model that was trained using VGG16 are displayed in Figures 6 and 7. These plots demonstrate how the model learns to identify ASL movements and can corroborate its performance.

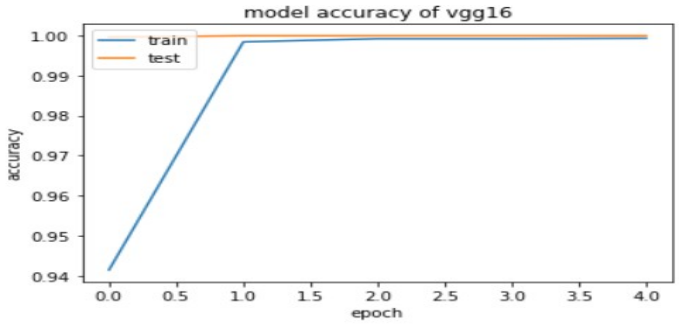

Figure 6. Model accuracy of VGG16

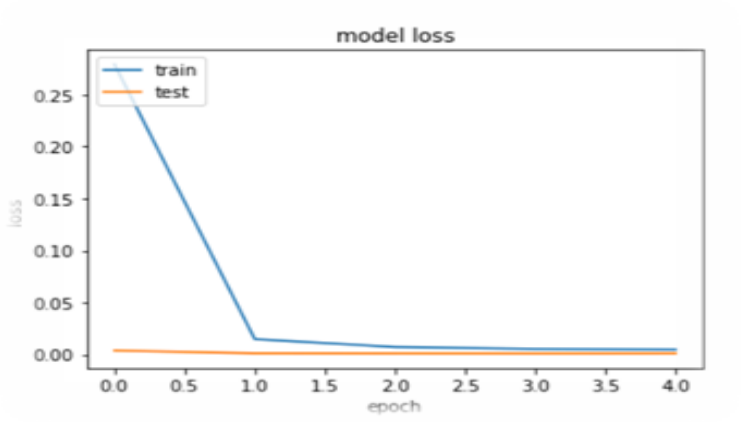

Figure 7. Model loss

5.3 Resnet50 accuracy and loss plot

Training and validation performance throughout several epochs are displayed in the accuracy and loss graphs of the ASL recognition model that is built on ResNet50. The accuracy plot displays the model's performance on the training and testing datasets for ASL gesture categorization across successive epochs. As training progresses, both the training and testing accuracy curves climb, showing that the model can learn and recognise ASL signs. The fact that these curves are converging shows that the model works well with unknown data and holds up under real-world conditions. If we compare the expected and actual labels, we can see the model's loss on the loss plot. The learning process of the model is validated when the loss trend is lowering, since it shows that the model is minimising prediction errors. If the training and testing loss curves are quite close to one other, it means that the model isn't getting too inflated by the training data, which means it will perform better on real-world ASL identification tasks. Figures 8 and 9 show the accuracy and loss graphs, which demonstrate the training dynamics of the ASL recognition model based on ResNet50. The model accurately understands ASL motions.

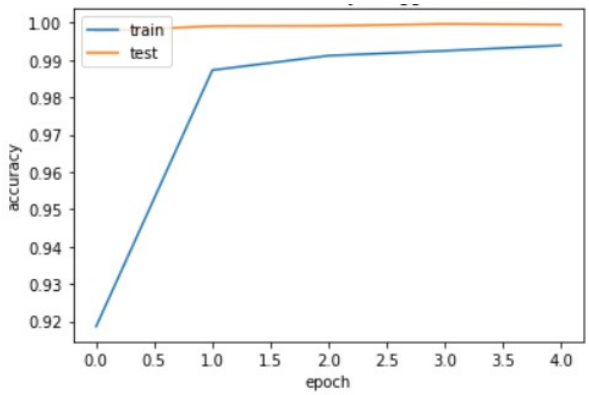

Figure 8. Model Accuracy of RestNet.

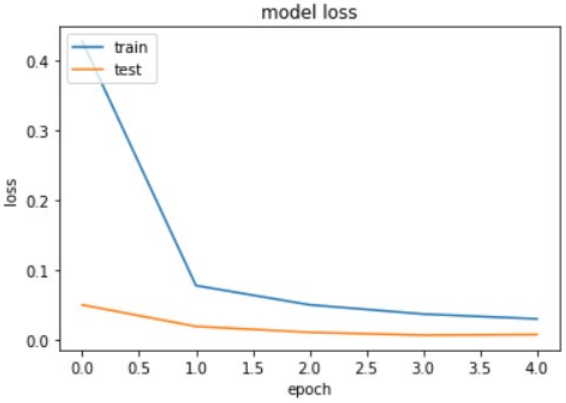

Figure 9. Model Loss

5.4 Single Image Prediction

After the picture has been resized into a NumPy array, it is then transformed to conform to the required input size of the model. Following that, the image is sent through the trained model so that the forecast may be obtained. The class probabilities are predicted by the model for each and every American Sign Language gesture; the class that is predicted is the one that has the highest likelihood. After that, the prediction is transferred to the alternative sign language gesture label that corresponds to it by means of a series of conditional statements. On the condition that the predicted probability for a certain class is the highest among all classes, the matching label is assigned to that class. At long last, the console is provided with a printout of the label that was anticipated. This approach demonstrates how the trained American Sign Language (ASL) recognition model may be used to predict the ASL gesture that is displayed in a single picture (see Figure 10).

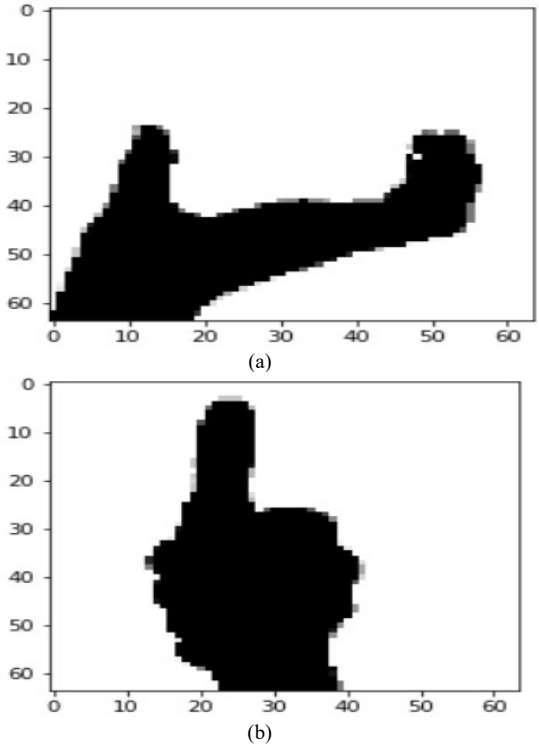

Figure 10. Sign for (a) Space (b) Best of luck

6 Conclusion

Our findings lead us to the conclusion that deep learning models, and more specifically VGG16 and ResNet50, are beneficial for the recognition of American Sign Language (ASL). In the course of comprehensive testing and analysis, both models displayed high levels of accuracy when it came to interpreting American Sign Language motions. Specifically, on the test pictures, VGG16 earned a 99.992% accuracy rate, while ResNet50 obtained a 99.95% accuracy rate. In light of these findings, it is

clear that deep learning architectures possess the capability to accurately recognise American Sign Language (ASL) gestures. This opens up intriguing possibilities for the development of communication systems that are both inclusive and accessible. In addition, the code for the single image prediction proved that the trained models could be utilised in a practical setting by enabling the recognition of American Sign Language gestures in real time from individual photographs. When everything is taken into consideration, our findings illustrate the significant progress that has been made in the field of deep learning-based American Sign Language (ASL) identification. This has implications for enhancing accessibility and inclusivity in communication for those who use sign language.

References

[1] Desai, A., Berger, L., Minakov, F., Milano, N., Singh, C., Pumphrey, K., Ladner, R., Daumé III, H., Lu, A.X., Caselli, N. and Bragg, D., 2024. ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition. Advances in Neural Information Processing Systems, 36.

[2] Song, B., Dai, X., Fan, X. and Gu, H., 2024. Wearable multifunctional organohydrogel-based electronic skin for sign language recognition under complex environments. Journal of Materials Science & Technology, 181, pp.91-103.

[3] Song, B., Dai, X., Fan, X. and Gu, H., 2024. Wearable multifunctional organohydrogel-based electronic skin for sign language recognition under complex environments. Journal of Materials Science & Technology, 181, pp.91-103.

[4] Zhu, Q., Li, J., Yuan, F. and Gan, Q., 2024. Multiscale temporal network for continuous sign language recognition. Journal of Electronic Imaging, 33(2), pp.023059-023059.

[5] Alyami, S., Luqman, H. and Hammoudeh, M., 2024. Isolated Arabic Sign Language recognition using a transformer-based model and landmark keypoints. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), pp.1-19.

[6] Alyami, S., Luqman, H. and Hammoudeh, M., 2024. Isolated Arabic Sign Language recognition using a transformer-based model and landmark keypoints. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), pp.1-19.

[7] Alves, C.E.G., Boldt, F.D.A. and Paixão, T.M., 2024. Enhancing Brazilian Sign Language Recognition through Skeleton Image Representation. arXiv preprint arXiv:2404.19148.

[8] Hu, L., Shi, T., Gao, L., Liu, Z. and Feng, W., 2024. Improving Continuous Sign Language Recognition with Adapted Image Models. arXiv preprint arXiv:2404.08226.

[9] Hu, L., Shi, T., Gao, L., Liu, Z. and Feng, W., 2024. Improving Continuous Sign Language Recognition with Adapted Image Models. arXiv preprint arXiv:2404.08226.

[10] Hu, L., Shi, T., Gao, L., Liu, Z. and Feng, W., 2024. Improving Continuous Sign Language Recognition with Adapted Image Models. arXiv preprint arXiv:2404.08226.

[11] Akdag, A. and Baykan, O.K., 2024. Multi-Stream Isolated Sign Language Recognition Based on Finger Features Derived from Pose Data. Electronics, 13(8), p.1591.

[12] Akdag, A. and Baykan, O.K., 2024. Multi-Stream Isolated Sign Language Recognition Based on Finger Features Derived from Pose Data. Electronics, 13(8), p.1591.

[13] Wadhawan, A. and Kumar, P., 2021. Sign language recognition systems: A decade systematic literature review. Archives of Computational Methods in Engineering, 28, pp.785-813.

[14] Camgoz, N.C., Koller, O., Hadfield, S. and Bowden, R., 2020. 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 (pp. 10023-10033).

[15] Gill, K.S., Anand, V., Gupta, R. and Pahwa, V., 2023, July. Insect Classification using Deep Convolutional Neural Networks and Transfer Learning On MobileNetV3 Model. In 2023 World Conference on Communication & Computing (WCONF) (pp. 1-5). IEEE.

[16] Wadhawan, A. and Kumar, P., 2020. Deep learning-based sign language recognition system for static signs. Neural computing and applications, 32(12), pp.7957-7968.

[17] Li, D., Rodriguez, C., Yu, X. and Li, H., 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1459-1469).

[18] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2023, March. Flower Classification Utilisizing Tensor Processing Unit Mechanism. In 2023 2nd International Conference for Innovation in Technology (INOCON) (pp. 1-5). IEEE.

[19] Min, Y., Hao, A., Chai, X. and Chen, X., 2021. Visual alignment constraint for continuous sign language recognition. In proceedings of the IEEE/CVF international conference on computer vision (pp. 11542-11551).

[20] Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K. and Fu, Y., 2021. Skeleton aware multi-modal sign language recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3413-3423).