SignAI: Indian Sign Language Recognition Using VideoMAE-Based Vision Transformers

Dr. Rakshith N[1], Chinmayee B[2], Meghana G K[3], Suman Kumar Matho[4]
[1,2,3,4]*Dept. of ISE, P.E.S College of Engineering, Mandya, Karnataka, India*

Abstract— **Bridging communication between the deaf and hearing communities demands reliable automated sign language interpretation. Widespread unfamiliarity with sign language among the general population creates substantial accessibility challenges. This paper presents SignAI, a video-based deep learning system for recognizing Indian Sign Language (ISL). The architecture centers on Video MAE, a vision transformer pre-trained through masked autoencoding on video data, which simultaneously encodes hand shape and motion cues from gesture clips.**

Unlike prior methods that depend on CNNs or LSTM-based recurrent networks, SignAI treats each gesture as a continuous temporal sequence. The pipeline processes raw video into 16-frame clips, applies them to a fine-tuned Video MAE backbone, and maps the extracted representations to one of 101 ISL categories. End-to-end deployment pairs a React-based upload interface with a Fast API inference server that returns both text and synthesized speech.

Experiments on the ISL-CSLTR benchmark show a Top-1 validation accuracy of approximately 74%, with well-balanced precision, recall, and F1-score, confirming the suitability of transformer-based sequence models for capturing the fine-grained spatiotemporal structure of ISL gestures.

Index Terms— **Indian Sign Language (ISL), Gesture Recognition, Video MAE, Vision Transformer, Deep Learning, Spatiotemporal Analysis, Sequence-Based Classification, Computer Vision**

I. INTRODUCTION

For individuals who are deaf or hard of hearing, sign language constitutes the primary medium of everyday communication, weaving together hand configurations, facial expressions, and body orientation to form a complete visual-spatial language. Yet this language remains largely inaccessible to the broader population, creating a persistent social barrier. Automated gesture recognition systems address this gap by translating sign sequences into text or spoken words that non-signers can understand. Recent strides in computer vision and neural network design have accelerated progress in this field.

Early approaches depended on hand-engineered features and classical classifiers, which could not adequately represent the dual spatial temporal nature of gestures. The subsequent adoption of convolutional-recurrent architectures partially solved the temporal modeling challenge, but these pipelines typically struggle with long-range dependencies spanning many frames. The emergence of transformer networks marked a turning point. Their self-attention mechanism allows each token in a sequence to attend to every other token, capturing global context that convolutional or recurrent stages miss. Extending this idea to video, Video Masked Autoencoders (VideoMAE) pre-train a transformer encoder by reconstructing heavily masked video patches, yielding rich spatiotemporal representations that transfer well to downstream recognition tasks.

SignAI builds on VideoMAE to tackle ISL recognition. The ISL-CSLTR dataset, comprising 7,671 videos spanning 101 gesture categories, provides the training and evaluation corpus. Videos are uniformly sampled to 16-frame sequences, normalized, and fed to the fine-tuned model. A full-stack deployment with a React frontend and FastAPI backend completes the system, returning prediction text and optional audio to end users. The system attains roughly 74% Top-1 accuracy, demonstrating robustness across diverse signers and recording conditions while acknowledging residual confusion among visually similar gestures

The main contributions of this work are listed below:

- VideoMAEisapplied to ISL video data, enabling joint encoding of spatial hand configurations and

temporal motion dynamics within a single transformer pass.

- Areproducible preprocessing pipeline handles frame extraction, uniform sampling to N=16 frames, resizing to 224×224, and ImageNet-based normalization.
- Sequence-level inference replaces per-frame classification, allowing the model to reason about gesture trajectory and continuity rather than isolated snapshots.
- Self-attention across the full clip enables the network to identify which spatial regions and time steps carry the most discriminative gesture information.
- A production-ready pipeline connects a React video-upload frontend to a FastAPI backend, outputting recognized gestures as text with optional text-to-speech.
- Comprehensive evaluation employs accuracy, precision, recall, F1-score, and AUC-ROC to characterize model performance across all 101 gesture classes.

II. LITERATURE SURVEY

A. Deep Learning in Sign Language Recognition

The last decade has seen machine learning displace hand-crafted approaches in gesture recognition. Rastgoo et al. [1] surveyed this trajectory comprehensively, tracing the shift from feature engineering toward end-to-end neural pipelines covering both static hand poses and dynamic gesture sequences. Among early neural architectures, Koller et al. [2] demonstrated that coupling CNN-derived spatial descriptors with hidden Markov model sequence decoders could substantially improve continuous recognition accuracy. Huang et al. [3] took a complementary route, eliminating explicit temporal boundary detection by training directly on unsegmented video streams.

B. Transformer-Based Approaches

Transformer architectures introduced a fundamentally different approach to sequence modelling that quickly found application in vision tasks. Camgoz et al. [4] pioneered a fully attention-based pipeline for simultaneous sign recognition and translation, reporting gains in capturing cross-frame dependencies that recurrent networks frequently miss. Their results motivated the broader adoption of attention mechanisms whenever long-range temporal reasoning is required.

C. CNN-Based Gesture Recognition

Convolutional networks formed the workhorse of gesture classification for several years. Kumar et al. [8] applied CNN classifiers to ISL hand shapes and achieved strong results within the finite vocabulary they studied. Ko et al. [7] and Pigou et al. [9] similarly leveraged convolutional feature extraction on both still images and short video clips; however, all these methods share a structural limitation: convolution over individual frames cannot capture gestural dynamics that unfold over time.

D. Spatiotemporal Models

To overcome the temporal blindspot of standard CNNs, researchers designed architectures that jointly process spatial appearance and temporal motion. Shi et al. [10] and Zhou et al. [11] independently proposed networks combining 2D spatial filters with temporal modules, showing that explicitly encoding motion cues improves continuous recognition benchmarks. These spatiotemporal models set an important precedent for treating video as more than a collection of independent images.

E. Detection-Based and Hybrid Models

A separate research thread has explored integrating object detectors with gesture classifiers. Tamura et al. [14] coupled YOLO-based hand localization with downstream classification to achieve high frame-level detection accuracy. Aly and Aly [6] demonstrated that unified deep learning pipelines could handle both static poses and dynamic gestures, reflecting the architectural flexibility afforded by modern networks.

F. Advanced Deep Learning Models for Complex Gestures

As datasets grew richer, research focus shifted to more challenging scenarios involving compound motions and bimanual coordination. Li et al. [12] and Li et al. [13] separately proposed frameworks for dynamic two-handed ISL recognition, illustrating how deep architectures are evolving to handle the full expressive range of sign languages rather than simplified single-hand subsets.

G. Research Motivation and Identified Gap

A clear gap persists in the literature: most methods either process frames independently, rely on sequential encoders that struggle with long-range context, or focus on hand detection without modelling full gesture sequences. Environmental variation, intra-class gesture variability, and the need for large labelled datasets compound these difficulties.

The escalating demand for seamless interactions between the hearing-impaired and the general public has driven substantial research into automated gesture translation platforms. While CNN and CNN-LSTM approaches have made progress, they often struggle to capture long-term dependencies and complex motion patterns in dynamic gestures. Detection-based methods like YOLO mainly focus on locating hands rather than understanding full gesture sequences.

Furthermore, numerous existing architectures analyze data on a per-frame basis, complicating the extraction of continuous motion traits. Changes in illumination, environmental clutter, and signing speed also degrade accuracy. To address these obstacles, we implement a VideoMAE-driven approach model that learns relationships across entire video sequences, improving the recognition of complex ISL gestures.

III. METHODOLOGY

SignAI processes ISL gestures as ordered temporal sequences rather than individual frames, capturing the motion continuity that distinguishes sign language from static pose recognition. The workflow proceeds through five stages: frame extraction, preprocessing, sequence formation, spatiotemporal feature learning with VideoMAE, and softmax classification over 101 gesture categories. A confidence gate filters low-certainty predictions before the output is rendered as text and optional speech.

The overall methodology follows a structured pipeline consisting of frame extraction, preprocessing, sequence formation, feature extraction, and classification. First, the input video is converted into frames, which are resized and standardized to ensure uniformity. A fixed number of frames (16) is selected and arranged into sequences to preserve temporal information.

The sequence of frames is then passed to the VideoMAE model, which learns meaningful spatiotemporal features using a transformer-based self-attention mechanism. The model captures both spatial details such as hand shape and temporal patterns such as motion across frames. Finally, the extracted features are given to a classification layer that predicts the corresponding gesture class. The predicted output is displayed as text and can also be converted into speech, making the system useful for real-world communication.

3.1 Dataset Description

Training and evaluation rely on the ISL-CSLTR benchmark ISL-CSLTR which contains 7,671 video clips and 144,448 individual frames distributed across 101 ISL phrase categories. Clips record different signers performing sentence-level gestures under varying illumination, clutter, and execution speeds. This diversity is intentional it encourages models that generalize beyond the recording conditions of any single session. The corpus is partitioned into training, validation, and test splits to allow unbiased performance estimation.

Each video is processed into fixed-length frame sequences, which are used as input to the transformer-based model. This repository can be accessed at: ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition - Mendeley Data, accessed on 2 December 2024

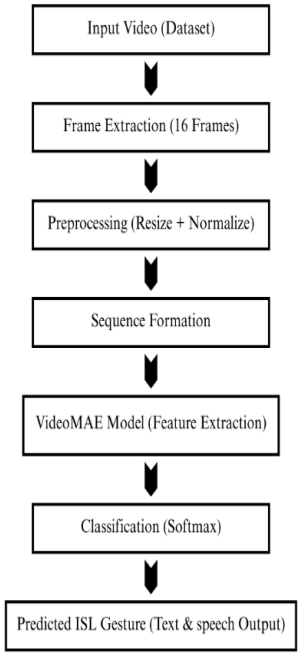

Fig. 3.1: Overall workflow of the proposed SignAI

3.2 Frame Extraction

The input video is first decomposed into individual frames. Let the input video be represented as:

$$V = \{f_1, f_2, f_3, \ldots, f_n\}$$

where f_i represents the i^{th} frame.

From each video, a fixed number of frames $N = 16$ are selected using uniform sampling to maintain consistency across all samples. This ensures that both short and long videos are represented using the same temporal length while preserving the essential motion information.

3.3 Preprocessing

Each extracted frame is processed before being used, to maintain consistency and improve model performance.

1. Resizing

All frames are resized to a fixed spatial resolution of:

$224 \times 224.$ (2)

This matches the input requirement of the VideoMAE model and reduces computational complexity.

2. Normalization

Image pixels are standardized utilizing the ImageNet dataset distribution metrics:

$$I_{norm} = \frac{I-\mu}{\sigma} \quad (3)$$

where:

- I= input image
- μ= mean
- σ= standard deviation

This step reduces variations caused by lighting and improves convergence during training.

3.4 Sequence Formation

After preprocessing, frames are grouped into fixed-length sequences:

$$S = \{f_1, f_2, f_3, \ldots, f_{16}\} \quad (4)$$

Each sequence represents a complete gesture over time. This sequence-based approach allows the model to learn motion patterns and temporal relationships, which are important for understanding dynamic gestures in sign language.

Sliding window techniques may be applied for longer videos to generate multiple sequences, improving training data diversity.

3.5 Patch Embedding

Each frame in the sequence is divided into non-overlapping patches of size 16×16. If the input frame is of size 224×224, the total number of patches per frame is:

$$\frac{224 \times 224}{16 \times 16} = 196 \quad (5)$$

Every individual patch undergoes flattening before being routed into a linear projection module. These embeddings are then combined with positional encodings so that both spatial and temporal information are preserved.

3.6 Feature Extraction using VideoMAE

The embedded patches are passed into the VideoMAE transformer model, which uses self-attention mechanisms to learn spatiotemporal features.

1. Self-Attention Mechanism

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

where:

- Q= Query
- K= Key
- V= Value
- d_k= dimension scaling factor

This mechanism helps the model focus on important regions in frames and learn relationships across different time steps.

2. Spatiotemporal Learning

- Spatial features: hand shape, orientation, position
- Temporal features: motion patterns, gesture transitions

The transformer captures both simultaneously, unlike CNN-based models that focus mainly on spatial features.

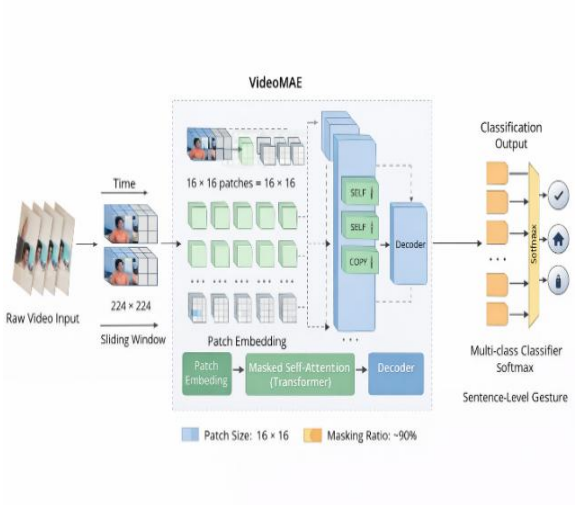

Fig. 3.6.1: Architecture of the Video MAE Model for Spatiotemporal Feature Extraction

3.8. Classification

The output features from the transformer are passed to a classification head consisting of a fully connected layer followed by a SoftMax activation function.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (7)$$

where:

4 C= number of classes (101)

5 z_i= output logits

The predicted class is obtained using:

$$\hat{y} = \arg\max P(y_i) \quad (8)$$

3.9. Confidence Filtering

To improve prediction reliability, a confidence score is calculated from the softmax output. Predictions below a certain threshold can be filtered out to reduce incorrect classifications.

$$\text{Confidence} = \max P(y_i) \quad (9)$$

This ensures that only high-confidence predictions are considered valid outputs.

3.10. Output Generation

The final predicted gesture is displayed as text output to the user. Additionally, the system supports optional speech output, converting the recognized text into audio to enhance usability and accessibility.

3.11. System Integration

The complete system is implemented using:

- Frontend: React-based interface for video upload
- Backend: FastAPI for handling API requests and model inference

The frontend captures user input and sends video data to the backend, where preprocessing and inference are performed. The predicted gesture is then returned and displayed to the user.

IV. RESULTS AND DISCUSSION

SignAI was assessed on the full ISL-CSLTR corpus: 7,671 video clips, 144,448 frames, and 101 gesture categories. Evaluation accuracy, precision, recall, F1-score, and AUC-ROC to give a multi-faceted view of classification quality and generalization.

The fine-tuned VideoMAE backbone achieved a Top-1 validation accuracy of 74%, with precision at 77% and both recall and F1-score at 74%. These closely matched figures indicate that the model avoids the precision-recall trade-offs that can inflate one metric at the expense of another, and that spatiotemporal representations extracted from 16-frame sequences are well-suited to sentence-level ISL categorization.

Training dynamics reveal that accuracy climbs sharply in the first five epochs before levelling off near 74% on the validation set, while the training loss descends monotonically. A modest gap between training and validation loss is present from epoch 7 onward, consistent with mild overfitting on sequence data a well-known behaviour in video transformers trained on relatively small corpora. The gap remains contained, however, indicating acceptable generalization to unseen signers and conditions.

The confusion matrix presents the classification performance across all gesture classes. The strong diagonal pattern shows that most gestures are correctly classified, while small off-diagonal values indicate occasional misclassifications. These errors mainly occur for gestures that are visually similar or have slight variations in motion, which is expected in real-world datasets. The Precision–Recall curve illustrates the relationship between precision and recall across different thresholds, with an average precision of approximately 0.79. The ROC curve shows a macro-average AUC value of around 0.98, indicating strong class separability and reliable classification performance.

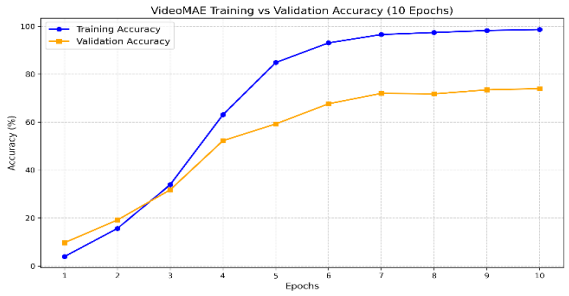

Fig. 4.1: Training and validation accuracy over epochs

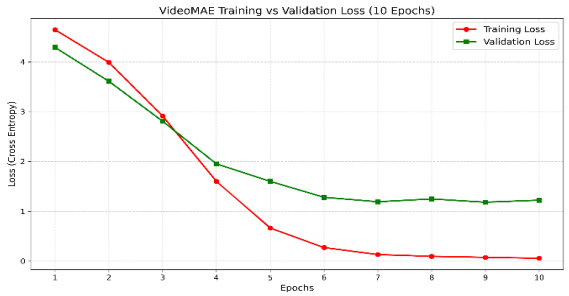

Fig. 4.2: Training and validation loss over epochs

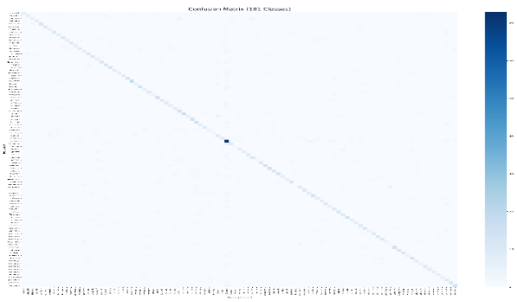

Fig. 4.3: Confusion matrix for ISL gesture classification

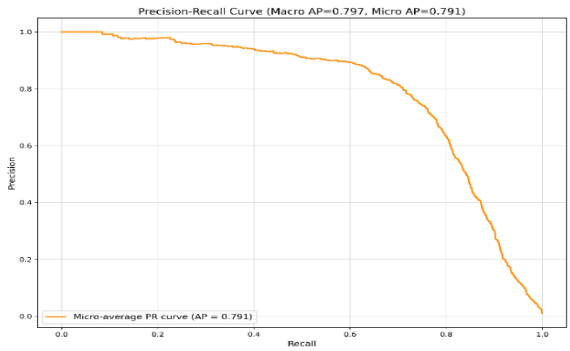

Fig. 4.4: Precision–Recall curve showing model performance

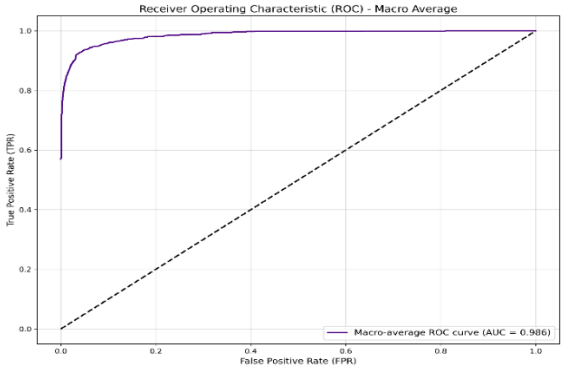

Fig. 4.5: ROC curve for multi-class classification

Table 4.1: Performance Comparison of Different Approaches

Model	Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC - ROC
CNN-Based	Spatial	65	67	64	65	0.85
CNN + LSTM	Hybrid	69	70	68	69	0.88
YOLO-Based	Detection	85	87.5	85	85	0.90
Video MAE	Transformer	74	77	74	74	0.98

The comparison findings indicate that architectures like YOLO achieve higher accuracy because they focus on identifying individual features directly. However, these models do not effectively capture the temporal relationships present in gesture sequences. In contrast, the VideoMAE model learns both spatial and temporal features together using a transformer-based architecture. This helps in better understanding of gesture sequences and provides improved robustness for sentence-level recognition tasks, making it more suitable for real-world applications.

A. Why VideoMAE over Other Models

To validate the capabilities of the proposed model, a comparison was carried out with CNN, CNN-LSTM, and YOLO-based models. Each of these models has shown good results in specific scenarios, but they have certain limitations when applied to sequence-based gesture recognition.

CNN-based models mainly focus on extracting spatial features from individual frames. While they work well for static images, they are not effective in capturing motion information across multiple frames.

CNN-LSTM models combine spatial feature extraction with sequential learning. Although they can model temporal dependencies to some extent, they still face challenges in capturing long-range relationships and complex gesture patterns.

YOLO-based models are designed for object detection and perform well in identifying hands and gestures at the frame level. However, they focus more on localization and are less effective in understanding continuous gesture sequences.

In contrast, the VideoMAE model uses a sequence-based transformer architecture that processes entire video sequences. Through self-attention, it learns relationships across frames more effectively, helping it understands complex gesture patterns. Even though it requires more computation, it provides better generalization and robustness for real-world sign language recognition tasks.

The ability of VideoMAE to learn global spatiotemporal relationships makes it more suitable for real-world sign language recognition compared to traditional methods.

V. CONCLUSION

This paper introduced SignAI, a full-stack system for automated Indian Sign Language recognition built around the VideoMAE vision transformer. By treating each gesture as a fixed-length video sequence rather than a series of independent frames, the architecture captures the continuous hand motion and temporal context essential for reliable ISL interpretation. The accompanying React-FastAPI pipeline makes the system accessible through a standard browser interface, returning recognized gestures as text and optionally as synthesized speech.

Evaluated on the ISL-CSLTR dataset across 101 gesture categories, SignAI achieved 74% Top-1 validation accuracy with a macro-average AUC-ROC of 0.98 metrics that reflect both accurate per-class prediction and excellent probability calibration. While detection-centric baselines such as YOLO attain higher frame-level accuracy, they lack the temporal reasoning needed for continuous gesture streams, making them unsuitable for real-world ISL deployment.

Remaining challenges include confusion among visually similar signs and sensitivity to extreme lighting or background variation. Future directions include data augmentation strategies targeting rare gesture classes, lightweight model distillation for edge deployment, and expansion to additional ISL vocabulary to support broader communicative contexts. SignAI represents a meaningful step toward intelligent assistive technology that fosters equal communication access for deaf and hard-of-hearing communities.

ACKNOWLEDGMENTS

The authors are grateful to their faculty supervisor and project guide for steady guidance and constructive feedback throughout this research. Institutional support from the Department of ISE, P.E.S College of Engineering, Mandya, provided the computational and laboratory infrastructure that made this work possible. The authors also thank the creators of the ISL CSLTR dataset for making the resource publicly available, and acknowledge the contributions of all individuals who participated in dataset preparation and system testing.

REFERENCES

[1] Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *IEEE Access*, vol. 9, pp. 145812–145840, 2021.

[2] S. Koller, O. Zargaran, H. Ney, and R. Bowden, "DeepSign: Hybrid CNN-HMM for continuous sign language recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2593–2607, 2020.

[3] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 32, no. 1, pp. 2257–2264, 2020.

[4] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10023–10033.

[5] M. Almasre and H. Al-Nuaim, "A vision-based hand gesture recognition system using deep learning," *IEEE Access*, vol. 8, pp. 107233–107245, 2020.

[6] S. Aly and A. Aly, "Deep learning-based sign language recognition system for static and dynamic gestures," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 8, pp. 4214–4225, 2022.

[7] M. Ko, J. Choi, and J. Kim, "Deep learning-based hand gesture recognition using CNN," in *Proc. IEEE Int. Conf. Consumer Electronics (ICCE)*, 2021, pp. 1–4.

[8] Kumar, R. Sharma, and P. Singh, "Indian sign language recognition using convolutional neural networks," *Procedia Computer Science*, vol. 167, pp. 2003–2012, 2020.

[9] L. Pigou, S. Dieleman, P. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. ECCV Workshops*, pp. 572–578, 2020.

[10] W. Shi, H. Wang, and X. Li, "Deep learning-based continuous sign language recognition using spatial-temporal networks," *IEEE Access*, vol. 9, pp. 103977–103989, 2021.

[11] Zhou, W. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 7, pp. 13009–13016, 2020.

[12] Y. Li, M. Rodriguez, and X. Zhang, "Vision-based sign language recognition using deep neural networks," in *Proc. IEEE Int. Conf.*

Image Processing (ICIP), 2022, pp. 3142–3146.

[13] K. Li, Z. Zhang, and Q. Liu, "A deep neural network framework for dynamic two-handed Indian sign language recognition," *Sensors*, vol. 25, Art. no. 3652, 2025.

[14] S. Tamura, T. Yanagi, and K. Kawamoto, "Hand gesture recognition using YOLO-based object detection and deep learning," in *Proc. IEEE Int. Conf. Artificial Intelligence and Computer Vision (ICAICV)*, 2023, pp. 112–118.

[15] P. Gupta and N. Sharma, "Sign language recognition using deep learning and computer vision techniques," in *Proc. IEEE Int. Conf. Machine Learning and Applications (ICMLA)*, 2024, pp. 589–594.