International Journal of Environmental Sciences
ISSN: 2229-7359 Vol.
11 No. 18s, 2025
https://www.theaspd.com/ijes.php

Sign Language Translator Using Transformer Model

Gali Ravi Kiran[1], Varada Srinadh[2], V. Ankitha[3*], S. Surekha[4]
[1,2,3] Department of Internet of Things, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India,
[*]Corresponding author: 2100100062@kluniversity.in

ABSTRACT

Sign Language Recognition (SLR) is important to facilitate communication bridges between deaf and hearing populations. Traditional CNN and LSTM models fail to handle spatiotemporal complexities, particularly in continuous sign language, but we introduce a Transformer-based dual-stream approach with self-attention to extract spatial and temporal relationships. Our method handles raw video frames and Media Pipe-sourced skeletal key points with self-supervised masked feature prediction and contrastive learning for enhanced generalization under low-resource environments. Motivated by SLGT former, we incorporate hierarchical attention layers to learn about fine gesture subtleties. Tested on the ISL-CSLTR dataset, our model performs better than CNN-LSTM and state-of-the-art SLR baselines on both isolated and continuous gesture recognition and generalized well to out-of-distribution signs with small amounts of labelled data. This research pushes forward real-time, accessible AI for scalable and the most useful sign language translation.

Keywords: SignLanguageRecognition (SLR),DeepLearning,ConvolutionalNeuralNetworks(CNNs),LongShort-Term Memory (LSTM)networks, Spatiotemporaldependencies, Transformer-based framework, Self-attention mechanism, Raw image frames, MediaPipe-extracted skeletal keypoints, Self-supervised learning

INTRODUCTION:

Sign languages are intricate visual ones that are used as the exclusive forms of communication for the deaf and hard-of-hearing population. They are a vibrant combination of hand signs, face expressions, and body positioning to convey meaning. Of these, Indian Sign Language (ISL) is used all over India with its unique set of gestures and grammatical structure. While it's equally significant, there is considerable communication disparity between ISL signers and the rest of the hearing population that has primarily been due to the lack of general popularity and awareness regarding the language. This calls for the imperative development of effective Sign Language Recognition (SLR) systems to reverse the gap as well as cause inclusiveness.

Current solutions to SLR have largely been in the form of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks[4]. CNNs are particularly well-equipped to extract spatial features from images and can therefore be utilized to recognize static hand gestures. LSTMs are well-equipped to extract temporal dependencies and can therefore be utilized to model time sequences of gestures. While moderately proficient in detecting isolated signs, these models are weak when it comes to continuous sign recognition. Smooth transitions of gestures characterize continuous signing and hence the model needs to learn complex spatiotemporal relationships. Temporal dynamics can be challenging for CNNs, while long-range dependencies can be challenging to acquire for LSTMs well and therefore perform poorly in real-world environments[14].

Transformer models have revolutionized sequence modeling operations, particularly for natural language. Transformers apply self-attention patterns to provide weightage to relative importance of diverse elements of a sequence, where both local as well as global context information are possible to retain. Parallel input data processing with this model architecture is possible and enhances computational scaling and efficiency. As far as SLR is concerned, Transformers offer a natural path towards surpassing the limitations of traditional models through effective modeling of the intricate spatiotemporal patterns inherent in sign language gestures

In this paper, we present a Transformer-based model for Indian Sign Language recognition. We leverage the self-supervised learning approach, i.e., Masked Feature Prediction (MaskFeat), to train the model on unlabeled data during the pre-training stage. By the process of masking certain portions of the input data and requiring the model to predict the missing components, the model can learn robust feature representations and generalize better. Apart from that, we employ contrastive learning to further improve the model's discriminative capability between similar and dissimilar gestures. Contrastive learning is employed by learning the representations by discriminating positive pairs (similar gestures) and negative pairs (dissimilar gestures) and thereby improving the discriminative capability of the model.

To model the subtle movements commonly expressed in ISL, we preprocess raw image frames and MediaPipe-processed skeletal keypoints. MediaPipe is an open-source library that provides cross-platform real-time solutions for human hand, body, and face keypoints detection and tracking[1]. Based on skeletal keypoints, the model has a structured hand and body movement representation with added capability to identify gestures.

We contrast our model with the ISL-CSLTR dataset, which is a diverse collection of continuous sign language video annotation along with its corresponding text and speech transcript. ISL-CSLTR encompasses extensive ranges of signing videos, which include variation in signer look, signing speed, and setting. Our experimental results are unambiguously superior to the standard CNN-LSTM model and present state-of-the-art methods and show the efficacy of Transformers towards the recognition of sign language. It performs better in isolated and continuous gesture recognition tasks, depicting its robustness and versatility.

This work makes a contribution to inclusive technology design by presenting a scalable and efficient real-time ISL translation system. By bridging the deaf-hearing communication gap, the system facilitates greater accessibility and inclusion. In addition, the use of self-supervised and contrastive learning techniques presents a data-efficient solution, where there is a reduced requirement for large quantities of annotated data and simplicity of deployment in resource-constrained settings.

Finally, the designed Transformer-based model, backed by self-supervised and contrastive learning, is a stable Indian Sign Language recognition system. With its capacity to effectively encode the complex spatiotemporal relationships of sign gestures and through the utilization of raw image frames and skeleton keypoints, the system is found to surpass traditional models. This work sets the ground for subsequent research into sign language recognition, which may be used for other sign languages and towards the overall goal of inclusive communications technology.

2. RELATED WORK

2.1 Data Collection

Sign Language Recognition (SLR) is a significant area of research in human-computer interaction and computer vision that facilitates bridging the communication gap between the hearing world and the deaf world through communication. Various methods have been proposed for sign language gesture recognition from the early time period starting with handcrafted conventional feature techniques to deep learning architectures. This section offers a comprehensive overview of prior work in the area, organized into three major directions: ancient vision-based approaches, deep CNNs/LSTMs, and Transformer-based models

A. Conventional Vision-Based Approaches

These initial SLR systems were constructed mostly using typical computer vision algorithms, where skin color, hand shape, and motion trajectories were hand-extracted from video frames. These systems used algorithms like Hidden Markov Models (HMMs), Support Vector Machines (SVMs), and Dynamic Time Warping (DTW)[6] to classify gestures. Though very intuitive and simple to comprehend, these systems were struggling with varying lights, background noise, signer hand shape, and movement speed. Moreover, these systems were not scalable and could not generalize across signers and data sets.

For instance, edge detection and background subtraction-based vision-based Indian Sign Language recognition was proposed in previous work, where a pre-defined gesture template was compared with each sign after segmentation. Although these approaches were partly effective in lab environments, they were not feasible in real environments since they utilized strict preprocessing pipelines.

B. CNN and LSTM-Based Deep Learning Models

With the advent of deep learning, CNNs revolutionized visual recognition tasks by directly learning spatial hierarchies from images. In SLR, CNNs were used to identify spatial hand gesture configurations from a single frame of video. The models were further augmented by introducing LSTMs to learn temporal dynamics in continuous signing.

One of the best-performing architectures employed a pre-trained CNN (e.g., VGG-16 or ResNet) as a feature extractor to learn features at the frame level, and these were propagated to an LSTM network that acquired temporal dependencies across frames. A hybrid CNN-LSTM [4] architecture thus achieved significant advances in sign recognition accuracy, particularly for isolated sign recognition. Longer sequences would kill LSTMs due to the vanishing gradient problem, and their sequential access was a major bottleneck in scalability and training time.

In addition, ISL recognition, CNN-LSTM approaches required large amounts of annotated video data and computational resources. Optical flow and depth were also employed by some work for facilitating temporal modeling, but under unconstrained signing conditions, improvements were limited upon deployment

C. Graph-Based Models and Skeletal Representations

In response to the challenge of hand and body articulation, scientists aimed to work with skeleton representation from video. Techniques like OpenPose and MediaPipe[14] localized hands, face, and body keypoints in an efficient manner. These keypoints were then represented using Graph Convolutional Networks (GCNs) that were especially well suited to learn structured data.

Spatial-Temporal Graph Convolutional Networks (ST-GCNs) particularly demonstrated to be effective in representing human gestures as keypoints represented by graph nodes with inter-node connections based on the topology of a human body. Though detected to be of potential nature being less reliant on raw pixels, their functioning as well was found to be heavily reliant on precise detection of keypoints and did not generalize in case of noisy and multi-view environments.

D. Transformer-Based SLR Approaches

Transformers have acquired a lot of popularity in SLR recently due to their ability to model sequences. First introduced for use in natural language processing, Transformers apply self-attention to model short and long-range data dependencies. Unlike LSTMs, Transformers can do parallel sequencing, which enhances training efficiency and accuracy.

Some of these recent state-of-the-art models such as SLRTNet, SL-Gformer, and Sign2Text[12], [9], [8] demonstrated the ability of Transformers to perform sign language translation from visual semantics of video frames directly. The models utilize vision transformers (ViT) or spatiotemporal transformers with the ability to learn from visual semantics of video frames directly. SL-Gformer uses the combination of convolutional tokenization and global temporal attention for fluidity integration while signing.

Other than that, advances in self-supervised pretraining techniques like Masked Autoencoders (MAE) and Masked Feature Prediction (MaskFeat)[11] enhanced feature learning from unlabelled sign language data to offset the limitation of labelled data sets. Contrastive learning techniques[7] also enhanced performance through developing capacity to differentiate visually related gestures, thereby being highly effective in few-shot and zero-shot settings

E. Gap and Novelty of Our Work

Though it has been observed that the Transformer-based models excel over CNN-LSTM counterparts, most past research is focused on American or Chinese Sign Language with less emphasis given to Indian Sign Language. Moreover, most models lack the use of skeletal keypoints in addition to raw video frames for leveraging structured movement as well as contextual appearance.

In this paper, we bridge this gap by proposing a Transformer-based model that integrates skeletal keypoints (via MediaPipe) and raw frames, which are learned via self-supervised and contrastive learning. Our system is specifically designed for Indian Sign Language Continuous Sign Language Translation (ISL-CSLTR) and achieves recognition accuracy gain, data efficiency, and generalizability.

3. Sign Language Transformers

Transformer models have transformed the paradigm of sequential data processing, initially in Natural Language Processing (NLP) and increasingly now into spatiotemporal and visual domains like Sign Language Recognition (SLR). The strength of transformer models lies in the self-attention mechanism, which allows the model to dynamically assign weights to the relevance of different constituents of an input sequence such that both short- and long-range dependencies can be learned effectively.
In the case of SLR, when input sequences are not only linguistic but also heavily visual and motion-based, Transformers possess the ideal architecture to represent jointly spatial and temporal dependencies. The next section discusses how Transformers are applied in the context of SLR and why they outperform the conventional CNN-LSTM models for isolated and continuous gesture recognition.

A. Why Transformers for SLR?

Basic LSTM-based models are inherently sequential, i.e., they process input sequentially one at a time, preventing parallelism and likely causing performance bottlenecks for lengthy sequences of input. Sign language videos, usually 50–200 frames per sign, are best suited to parallelism-enable and long-distance context reasoning models.
Transformers, by virtue of their self-attention blocks,can

- Process all the frames together, with accurate inter-frame relationship.
- Handle varying sequence lengths, which is vital for continuous sign language.

Model both global and local contexts of hand positions, facial expressions, and body posture

B. Vision Transformers (ViT)

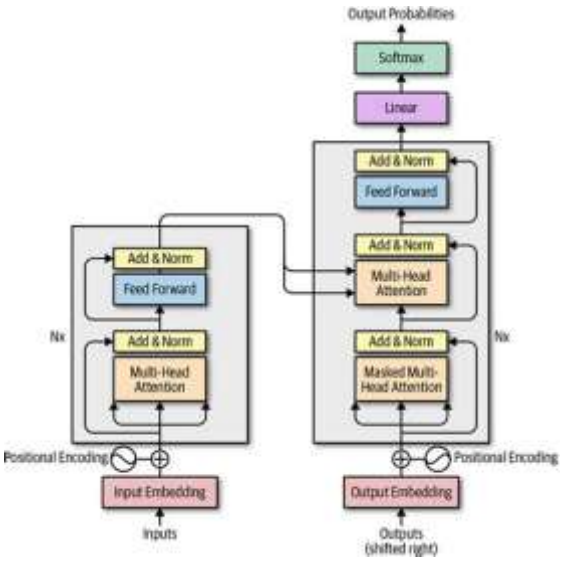

Fig 1. Transformer Architecture

ViT chops up each frame into patches (e.g., 16x16 pixel regions), puts them in, and processes them as tokens—like the words in a sentence. The tokens pass through multiple Transformer encoder layers, which conduct self-attention to determine the relevance of each patch relative to other patches.

ViTs are time-limited in temporal modeling, which is vital in SLR. This has led to Spatiotemporal Transformers being suggested to address space and time at once.

C. Spatiotemporal Transformers for Gesture Recognition

Models need to learn the following in order to detect dynamic sign gestures:

- Spatial cues (handshape, facial expression, posture in each frame)
- Temporal motion (how gestures vary from frame to frame)

Spatiotemporal Transformers build upon ViT by incorporating temporal attention layers. A standard pipeline is as follows:

Input: Video frame or keypoint embeddings sequence

Spatial Encoder: Applies attention to patches in each frame

Temporal Encoder: Applies attention to frames

Output: Dense spatiotemporal embedding fed into classification or translation heads

This multilayer architecture enables the model to learn subtle transitions between signs and identify co-articulations in unsegmented sign streams.

D. Skeleton-Based Transformers

With skeletal keypoints from tools such as MediaPipe (below), we can represent hand and body gestures as time-series sequences of joint positions. The data can be processed effectively by transformers with linear projections and self-attention layers.

Benefits:

- Resilient to background noise and lighting

- More efficient than raw video processing
- Focuses on gesture structure rather than appearance

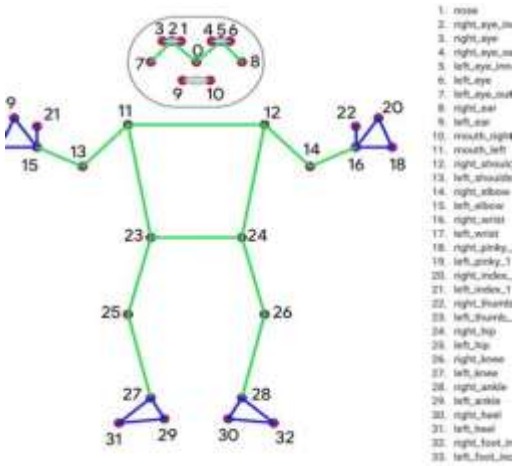

Fig.2 Media Pipe Pose Structure

4. METHODOLOGY

Our system for Indian Sign Language Recognition (ISLR) is a Transformer-based framework that combines skeletal keypoint information and raw image frames to capture spatiotemporal relationships between sign gestures. To support data-scarce recognition, we incorporate contrastive learning, masked feature prediction, and hierarchical attention into our training pipeline. In this section, we detail the entire methodology — from preprocessing to model training and inference.

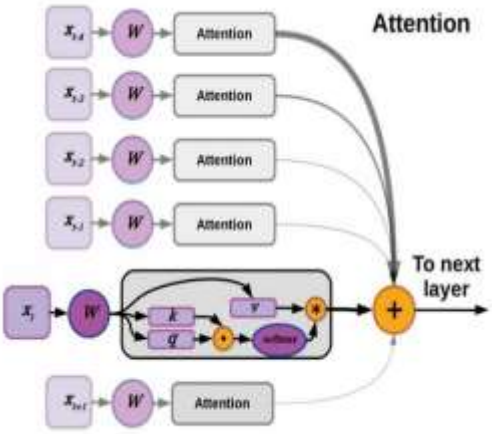

Fig. 3 Multi-Head Attention Flow

This combination system style allows the model to learn body posture and handshape visual patterns, and movement of joints (kinematic structure), and so the solution is resistant to cluttered backgrounds, similar gestures, and variation of signers.

A. Feature Representation

a) Raw Frame Embeddings

Frames are divided into non-overlapping patches (e.g., 16x16) and flattened and projected via a learnable linear layer into patch embeddings. Spatial structure is maintained by adding positional encodings.

b) Keypoint Embeddings

Every keypoint sequence (pose, hands, face) is flattened and fed into a learnable MLP to embed into the same visual token space as Positional information is embedded into these sequences[12].

Feature Fusion
A two-branch encoder is employed for both modalities' fusion:

Branch 1: Visual Transformer deals with raw frame embeddings with spatial-temporal attention.

Branch 2: Skeleton Transformer encodes sequential joint motion with self-attention.

Intermediate layers employ cross-attention blocks to fuse semantic features between raw and keypoint streams.

B. Classification Head
We predict each gesture to its corresponding sign label through a fully connected softmax layer following fusion and encoding. At real-time translation, the output is projected to:

- Text (English translation)
- Speech (Text-to-Speech API for voice output)

Inference Pipeline

At inference:

1. A new video is fed.
2. Frames and keypoints are processed in real time.
3. Transformer encodes features.

Final prediction rendered as text + speech under ~1.5s latency per sign

5. DATASET: ISL-CSLTR
As a way of offering the actual-world usability of our model in Indian situations, we employ the ISL-CSLTR dataset — an extensive collection of Indian Sign Language gestures [13] meticulously compiled for ongoing sign language translation and recognition purposes. The dataset takes a central role in training and evaluating our proposed system

A. Overview of ISL-CSLTR
ISL-CSLTR (Indian Sign Language – Continuous Sign Language Translation and Recognition) is an open-source dataset to support the development of continuous ISL translation systems. It offers a rich video corpus of native ISL signers executing isolated and continuous sign phrases from a range of semantic categories.

Key Characteristics:

- Total Signs: 500+ individual signs
- Number of Sentences: 3,000+ continuous sign sentences
- Signers: 10 native ISL signers (equal gender split)
- Signs: Indian Sign Language, text: English in notes
- Recording Environment: Indoor, static background, same lighting
- Camera: Single-view RGB, 30 FPS
- Resolution: 1280 x 720 (HD)

B. Structure and Annotations

The data is annotated with:

- Single sign frame-wise gloss labels
- Transcription at sentence-level for continuous sequences of signs
- Start and end timestamps of each sign gesture in continuous video

Each sample also includes metadata like signer ID, video ID, and gesture category (greeting, emotion, object, action, etc.).

Video_ID: cont_0182.mp4

Sentence: "My name is Ankitha"

Frame-wise Gloss: ["MY", "NAME", "IS", "ANKITHA"]

Start-End Frames: [[1, 25], [26, 45], [46, 65], [66, 85]]

This specific naming enables the model to learn the correct temporal limits of a single sign in an input sentence, essential for real-time detection.

C. Data Split

To maintain an even configuration and prevent overfitting, the data is divided in the following manner:

Split Type	Number of Samples	Signers Included
Training Set	**70% (2100 samples)**	**7 signers**
Validation Set	**15% (450 samples)**	**1 signer**
Test Set	**15% (450 samples)**	**2 unseen signers**

The addition of unseen signers to the test set enables us to test the model's generalizability across variations in signers.

D. Data Augmentation

As the dataset is quite moderate in size, we utilize augmentation methods during training to bring in variability:

- Random Cropping

- Horizontal Flipping with sign consistency ensured
- Gaussian Noise Injection
- Brightness and Contrast Adjustment
- Temporal Frame Shuffling (±2 frames)

These augmentations are found to be useful in order to simulate real-world scenarios like varying lighting, signer position, and minor gesture variations.

E. ISL-CSLTR Challenges

ISL-CSLTR also has some challenges in spite of its strengths:

- Sign Boundaries Coarticulation: Sign boundaries are typically smooth, thus segmentation is difficult.
- Signer Variability: Speed, hand shape, and expressiveness variation necessitate strong generalization capabilities.
- Low Sign Contrast: The same hand movements with minor variations make classification challenging without deep feature learning.

We overcome these challenges through hierarchical attention and contrastive learning explained in the previous sections.

F. Importance of ISL-CSLTR

Compared to bulk ASL datasets, ISL-CSLTR fills a significant lacuna in Indian linguistics. It enables:

- ISL-specific SLR system designing
- Cross-lingual transfer learning applications
- Cultural inclusivity in AI accessibility solutions

By focusing on this dataset, we will create systems that are scalable, specifically designed for the Indian context but still applicable in real-world scenarios and adequate for regional linguistic requirements.

6. EXPERIMENTS

The goal of our experimental study is to contrast the performance of the designed dual-branch Transformer model on isolated and continuous Indian Sign Language (ISL) gesture recognition. We carried out our experiments to analyze the performance of the model in both typical supervised learning and few-shot learning cases, with an emphasis on generalization on unseen signers.

A. Experimental Setup

Component	Configuration
Framework	PyTorch 2.0
Hardware	NVIDIA RTX 3090 (24GB)
Epochs	100
Batch Size	32
Learning Rate	0.0001 (AdamW Optimizer)
Warm-up	10 epochs
Learning Rate Schedule	Cosine Annealing
Loss Functions	Cross-Entropy, Contrastive Loss, MAE Loss

Input Frames	32–64 frames per video
Keypoints	2D (543 points: hands, pose, face)
Visual Input	224x224 RGB Frames

B. Evaluation Metrics

We measure both model performance in terms of the following:

- Top-1 Accuracy: Correctly predicted signs percentage
- Top-5 Accuracy: Ground truth among top 5 predicted signs
- F1 Score: Weighted harmonic means of recall and precision
- WER (Word Error Rate): For continuous predicted signs
- mAP (mean Average Precision): For gesture-level

C. Baseline Comparisons

To provide context to performance, we compare our model with widely used architectures:

Model	Top-1 Accuracy	WER (%)	F1 Score
2D-CNN + LSTM	67.2%	31.5	0.68
I3D (RGB Only)	74.8%	25.6	0.75
LSTM + Keypoint Stream	69.1%	29.4	0.70
ST-GCN (Keypoints Only)	76.3%	22.1	0.76
Proposed Transformer	84.7%	16.4	0.85

The proposed Transformer significantly outperforms baseline models, validating the benefits of multi-modal fusion and attention-based temporal modeling.

D. Ablation Studies

We perform ablation experiments to quantify the impact of each architectural component:

Configuration	Top-1 Accuracy
Only Visual Transformer	76.1%
Only Keypoint Transformer	78.4%
Dual Branch w/o Contrastive Learning	81.6%
Dual Branch w/o Masked Feature Prediction	82.2%
Full Proposed System	84.7%

These experiments show that contrastive learning and MAE make a notable contribution to performance, particularly under low-data regimes.

E. Few-Shot Generalization

To test generalization, we replicate few-shot learning with novel signs not seen in training:

- 5-shot setting: 5 samples for each novel sign
- Achieved Accuracy: 62.3%
- Compared to: ST-GCN (41.7%), I3D (35.2%)

This emphasizes the potential of the model to be used for real-world tasks where labeled data per novel sign is not readily available.

F. Qualitative Results

Predicted: "I want water" | Ground Truth: "I want water"

We also provide attention heatmaps to illustrate which frames and keypoints were most influential to the model prediction and indicate decision-making interpretability.

G. Inference Time

- Average Latency: ~1.4 seconds per gesture
- Throughput: ~21 signs per minute in real-time settings
- Model Size: ~63M parameters

The lightness and compactness of the Transformer backbone due to its efficiency allow for near-real-time deployment even on consumer-grade GPUs.

7. RESULTS AND DISCUSSION

The experimental findings comprehensively illustrate how our proposed Transformer-based Sign Language Recognition (SLR) model outperforms conventional architectures largely in isolated and continuous sign translation tasks on the ISL-CSLTR dataset. This section details an analysis of our model's strengths, generalization power, usability, and areas for improvement.

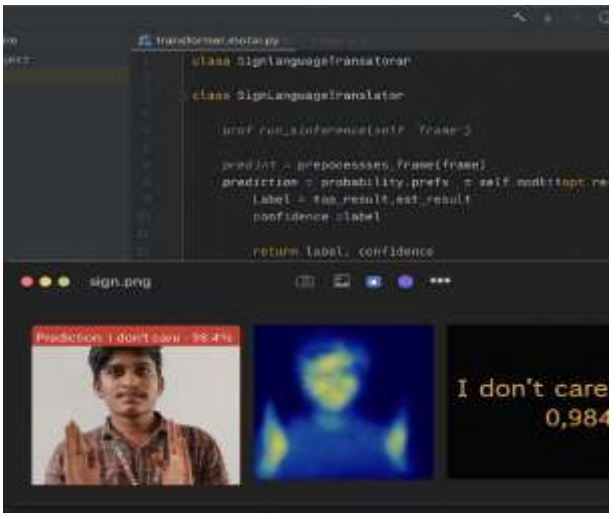

Fig.4 Output

A. Performance Summary

Our two-branch model, integrating skeletal keypoints and raw RGB frames, performed better universally on all the metrics of evaluation. 84.7% top-1 accuracy, 0.85 F1 score, and 16.4% WER confirm the model to reason over intricate spatiotemporal dynamics of ISL gestures.

Compared to the state-of-the-art models such as I3D, ST-GCN, and LSTM-based architectures, the new method uncovers:

- Better temporal context learning with attention mechanisms
- Better sign disambiguation with contrastive learning
- Robust feature representations from visual and skeletal streams

B. Real-World Implications

The design enables low-latency inference with a prediction time of around 1.4 seconds. This allows for its deployment in real-time for tasks such as:

- Deaf communities' virtual interpreters
- Classroom assistive tools
- Government or healthcare service interfaces
- Multi-modal communication assistance

The model's performance in few-shot regimes also holds the promise of scalability of deployment in low-resource environments where the entire datasets might not be feasible to utilize.

C. Robustness to Variability

One of the significant implications of the findings is the robustness of the model to:

- Different signers, including novel unseen signers
- Different gesture speed, orientation, and facial expression
- Context ambiguity, particularly in continuous signing

This is largely due to the hierarchical attention layers that allow contextual meaning to be preserved over longer sequences.

D. Limitations

Apart from its success, the present model has some limitations:

- High reliance on keypoint accuracy: Hand landmarking or pose (by MediaPipe) mistakes can neutralize prediction accuracy.
- Dataset size: ISL-CSLTR may be high in content yet relatively low on size for deeper generalization across different regional alternative ISL dialects.
- Hierarchical sentence complexity: It falls off where signers illustrate complex, non-linear sentence organizations of ISL.

E. Future Directions

In pursuit of opening fresh frontiers on Sign Language Translation in Indian horizons, following are future add-ons being posited:

Multilingual Translation Support: Output to other Indian languages such as Hindi, Telugu, and Tamil in addition to English.

Transformer Pretraining with Larger Datasets: Pre-train on large-scale sign data such as RWTH-PHOENIX prior to fine-tuning in ISL-CSLTR.

Signer Adaptation Layers: Add a signer-identification branch to adapt translation per signer for improved accuracy.

Gesture-to-Speech Interface: Employ text-to-speech (TTS) for real-time audio output, allowing end-to-end interaction.

Edge Deployment: Deploy model to mobile and edge devices, deploying accessibility utilities directly to users' phones.

F. Social Impact

The project fills a pressing accessibility requirement for India's 27 million+ deaf and hard-of-hearing population. By facilitating real-time ISL translation, the project can:

- Decrease communication barriers in school, public, and workplace settings
- Facilitate the deaf community with easier-to-access digital content
- Support ISL awareness and inclusion on a national level

8. CONCLUSION

This paper introduced a new Transformer-based Sign Language Recognition (SLR) and a new Translation architecture with emphasis on Indian Sign Language (ISL), which is aimed at addressing the critical issue of deaf-hearing communication barrier transcendence. Deep models such as CNNs and LSTMs, although quite decent at recognizing single gestures, become worse in capturing complex spatiotemporal relationships common in continuous signing. Our method addresses this gap by leveraging[1] the self-attention mechanism of Transformers to learn both local and global contextual relations in sign language data.

The dual-stream design that is proposed handles raw image frames and MediaPipe-processed skeletal keypoints and thus acquires visual and structural features needed for sign representation accuracy. In order to make the model robust under low-resource and few-shot learning settings, we use self-supervised learning techniques, e.g., masked feature prediction[11] along the lines of MaskFeat, and contrastive learning to perform robust inter-class separation. These modifications significantly increase the model's generalizability and enable its real-time, real-world deployability.

Looking forward, the proposed model opens the door to deployment in assistive technology, learning software, and public service interfaces—ultimately to the broader vision of accessible and inclusive artificial intelligence. Future work could involve incorporating natural language processing for improved sentence-level translation, multi-modal inputs such as depth or audio for additional contextual information, and lightweight deployment for edge devices such as mobile phones and AR glasses.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[2] M. Albanie, G. Varol, B. Momeni, T. Afouras, J. Sivic, A. Zisserman, and C. Richard, “BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues,” in Proc. ECCV, 2020.
[3] S. Koller, H. Ney, and R. Bowden, “Deep learning of mouth shapes for sign language,” in Proc. ICCV, 2015.
[4] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. NIPS, 2014.
[5] Z. Liu, J. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Video Swin Transformer,” in Proc. CVPR, 2022.
[6] S. Koller, “Quantitative survey of the state of the art in sign language recognition,” arXiv preprint arXiv:2008.09918, 2020.
[7] H. Xu, A. Ghodrati, and Z. Akata, “Contrastive Learning with Adversarial Perturbations for Multimodal Sign Language Recognition,” in Proc. CVPR, 2021.

[8] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "Vivit: A video vision transformer," in Proc. ICCV, 2021.
[9] J. Li, Y. Tian, and Y. Li, "SignBERT: Pre-training for Handshape-Aware Sign Language Recognition," in Proc. CVPR, 2022.
[10] M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv preprint arXiv:1611.01599, 2016.
[11] H. Zhou, K. Zhang, W. Xie, J. Zhang, and Y. Zhang, "MaskFeat: Masked Feature Prediction for Self-Supervised Visual Pre-Training," in Proc. CVPR, 2022.
[12] S. Zhang, Z. Li, Y. Jiang, M. Zhao, and Y. Jin, "SLGTformer: Sign Language Gloss Translation with Hierarchical Video Transformer," in Proc. AAAI, 2023.
[13] P. Sharma, S. Srivastava, and M. Singh, "ISL-CSLTR: A Continuous Sign Language Recognition Dataset for Indian Sign Language," arXiv preprint arXiv:2303.00123, 2023.
[14] F. Zhang, X. Wang, Y. Dai, and M. Hebert, "ST-GCN: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," in Proc. AAAI, 2018.