International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Hybrid Vision-Based Sign Language Recognition: A Review

Prerna Charis J
Department of Artificial Intelligence and Machine Learning, Amity University, Gurugram, Haryana, India

Abstract—Sign Language Recognition (SLR) has emerged as an important research area at the intersection of computer vision and deep learning, and human and machine interaction with an objective of enabling effective communication between deaf and hearing communities. Recent advances in deep learning have improved the performance of vision-based Sign Language Recognition systems, particularly by using hybrid architectures that combine spatial features extraction and temporal sequence modelling. The goal of this review is to provide a overview of the recent developments in hybrid Vision-based Sign Language Recognition and to examine the advantages, limitation and practical deployment challenges of the current approaches. This paper provides a systematic review of the literature, the surveyed methods broadly classified into CNN-LSTM architectures, Transformer-based models and multimodal integrated frameworks which integrates visual and skeletal information. This review further investigates critical challenges affecting the deployment in real-world scenarios which includes domain shift, data scarcity, co-articulation, sign ambiguity and computational constrain. We will also discuss about emerging research direction such as self-supervised learning, cross-linguistic transfer learning, generative domain adaptation, multimodal bio signal integration, and community-centered dataset development. This survey also highlights the significant progress achieved in continuous sign language recognition while identifying the remaining technical and practical barriers that must be removed to develop robust, scalable, and user-independent SLR systems capable of operating in real-world environments.

Key Words—Sign Language Recognition (SLR), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Vision Transformer (Vit), Temporal Smoothing, Hybrid Architecture, Domain Adaptation, Self-Supervised Learning, Continuous Sign Recognition.

I. INTRODUCTION

Sign Language acts as the primary means of communication for nearly 70 million people worldwide [1]. For example, American Sign Language (ASL) is the primary language of around 500,000 individuals in United States and Canada [2],[3]. Though being used within the large English-speaking communities, ASL is differs linguistically from spoken English and posses its own grammatic structure. Sign language should not be viewed as a simplified version of the spoken language, rather it is a complete and independent natural language with its own phonology, morphology, syntax and pragmatics. The meaning is conveyed through the interaction of multiple visual and spatial components, which includes hand form, orientation, movement path and even facial expressions [4],[5],[6]. This rich multimodal structure makes SLR a challenging task for deep learning systems.

For a machine to comprehend sign language in real time, it faces a problem that neither speech recognition nor gesture recognition fully prepared it for. Speech recognition deals with one-dimensional gesture over time. Static gesture recognition deals with spatial information in a single frame. Sign language requires both simultaneously — and the meaning of a sign language can change entirely based upon the direction of the hand movement, the speed of performance, the position of the non-dominant hand, or the alignment of the person's face. [6],[7]

This is the main motivation behind hybrid architectures: no single model type is able to handle all of these dimensions accurately. Convolutional Neural Network (CNN) is efficient at reading the spatial configurations within an

individual frames but it is unsighted to motion over time. Long Short-Term Memory network (LSTM) helps tracking the temporal sequences well but it only depends up clean, well-obtained spatial features. Putting them together, we get the combination that is able to solve the problem. [8],[9],[10]

In recent times, Transformer-based architectures, which are formerly developed for language modelling, have proved that the self-attention mechanism handovers remarkably well towards video sequences. Models such as Sign Language Transformers [11] and Swin-MSTP [12] have set new standards on continuous recognition tasks, which outperforms CNN-LSTM approaches by a considerable margin on translation metrics.

This paper reviews the field with a focus upon what practically matters. For instance, the accuracy under real-time conditions, robustness across different person who does sign and their environments, and the gap between target numbers and deployed systems. We have traced the evolution from handmade features through deep learning, observed each major architectural models, and identified the open problems that the field has not yet resolved. [4], [7]

II. LITERATURE REVIEW

A. The Rule-Based Era

The first generation of an automated SLR systems was built almost entirely by hand [13]. Researchers had extracted visual features such as edge gradients via Histogram of Oriented Gradients (HOG), local texture descriptors by means of Scale-Invariant Feature Transform (SIFT), and optical flow to capture movement, and wrote precise rules mapping those features to sign classifications. As Shahin and Ismail [4] mentioned in their study, these systems had worked reasonably within strict and controlled conditions but fell quickly when any of the variables changed. For example, different background or different skin tone or a window behind the person signing, any of these could reduce performance significantly. Therefore, there was no actual leaning by the machine; the model only learnt what it had been programmed to identify.

B. Probabilistic Sequence Models

After the rule-based era came the introduction of Hidden Markov Models (HMMs). It brought in a genuine change in concepts. Rather than the firm rules, HMMs represented the natural variability in showing how any assumed sign may be produced as probability distribution over any sequence of observations. As HMMs are able to process time-series data with variable temporal lengths and discount timing variations through the use of skipped-states and same-state transition, they are widely used in continuous speech recognition [14]. Mittal et al. [15] traced the influence of HMMs on continuous SLR systems, noticing that their ability to handle variable-length sequences and differences in timing between the signers made them drastically more robust than the rule-based predecessors. Dynamic Time Warping (DTW) presented a complementary approach, allowing the systems to compare sequences of various lengths by warping the time axis [16],[17]. Both methods were factual improvements, but they depended totally on the quality of the features provided to them and those features were still hand-made [13]. The representation constriction remained.

C. The Deep Learning Transition

The arrival of deep learning eliminated the representation constrictions. Because CNNs are able to learn their own features directly from raw pixel data; they also are capable of discovering which visual patterns were actually predictive of which signs without specifying them [18],[19]. As Jiang [6] reviews, the evolution happened rapidly as soon as GPU (General Processing Unit) hardware became available. CNN-based systems outperformed hand-made approaches on most of the standard benchmarks within a few years of the deep learning trend hitting computer vision widely [20],[21].

But CNNs were built for images and not videos. They can only see one frame at a time; with no memory of whatever came previously. For isolated sign recognition, classifying a single, pre-segmented gesture, this is often sufficient. But for continuous recognition, where signs coincide each other and boundaries must be deduced from the signal itself, it is not sufficient [22]. Thus, the hybrid CNN-LSTM architecture became the field's first significant answer to this problem [22]. This uses the CNN for extracting the spatial features frame by frame, then feed these features into an LSTM for modelling the temporal evolution of the sign [24]. Walter et al. [7] provided a useful overview of how this transition have reshaped evaluation benchmarks and deployment potentials across the field.

D. The Transformer Era

The introduction of Transformer architectures to SLR represents the most noteworthy change since the progress of deep learning [24]. Camgöz et al. [11] demonstrated in their landmark 2020 work that sign language recognition and translation can be handled together in a single end-to-end Transformer model, which in result noticeably improves on LSTM-based translation baselines on the PHOENIX-2014T benchmark. The key insight is that self-attention can directly link any two frames in a sequence, regardless of how far apart they are [25]. This is a structural advantage for continuous signing, where the definition of a current gesture may depend on something that has been signed many frames prior to it. Following architectures including the Swin-MSTP [12] and the Spatial-Temporal Transformer [26] have widened this further by introducing the mechanisms to manage both local and global temporal structure at the same time.

III. SPATIAL FEATURE EXTRACTION

It is necessary that before a system can understand what a sign means, it must learn to understand what the hand looks like in the given frame. This is considered as the spatial feature extraction problem, and it seems to be harder than it appears. The same hand form or gesture can look very different based on lighting or camera angle or skin tone or background or even based on the position of the non-dominant hand. A suitable spatial feature extractor is the one that captures what actually matters such as the geometric configuration of the hand and disregards what is not significant. [6], [27]

A. Convolutional Neural Networks

CNNs have become the default spatial backbone for virtually all modern vision-based SLR systems [28]. Their hierarchical architecture that includes learning edges at the shallow layers, shapes at the intermediate layers, and complex part configurations at the deeper layers, suits the problem naturally [29]. Pandey et al. [27] demonstrated this hierarchy clearly in their Indian Sign Language translation system, presenting how the successive convolutional layers gradually abstract from raw pixel intensities to hand-configuration representations which are largely indifferent to superficial visual variation.

The balance between accuracy and cost of computation is essential and vital for every practical system. Deep architectures like ResNet-50 or ResNet-101 performs well on standard test but are too slow for real-time deployment any embedded hardware [30]. This enhanced significant research endeavors toward other lightweight alternatives. Trpcheska, Zevnik, and Bader [31] researched MobileNetV2-based systems that runs directly on STM32 microcontroller, a kind of chip found in embedded devices, not phones, and achieved an F1-score of 0.865 at a response time of 103 milliseconds per frame. That is considerably fast enough for a natural conversation flow. MobileNetV2's key innovation which is depth wise separable convolutions accomplished a similar computational result to typical convolutions in roughly 8 to 9 times fewer operations [32].

In addition to all attention mechanisms have been merged into CNN architectures to improve their perception on complex signs without adding any significant operations [10]. Kumari and Anand [8] showed that adding a channel-wise attention layer to a ResNet backbone by effectively teaching the network to weight its own feature maps by their significance, that resulted in improved classification accuracy on WLASL-100 while retaining inference time feasible for real-time use. Kazbekova et al. [10] found parallel benefits from spatial attention in the lightweight CNN-BiLSTM system, here the attention mechanism trained to concentrate computation only upon hand regions and basically ignores the torso or any other parts of the body and the background.

On the other hand, Neural Architecture Search (NAS) gives a different philosophy which states that rather than a researcher choosing the architecture, an automated algorithm searches among the possible CNN configurations to find the one that is reliable in meeting a given constraint which can be a specific accuracy target or a latency budget or a memory limit. Walter et al. [7], in his survey went through numerous NAS-derived SLR models and found out that they consistently outperformed hand-designed architectures of comparable computational cost in a constrained deployment scenario.

B. Vision Transformers

Now we will come to the Vision Transformers (ViTs). These take a radically different approach to spatial feature extraction. In this approach, rather than processing the image through a hierarchy of local convolutions, ViTs

divides the image into grids of fixed-size patches and then process those patches with Transformer encoder, allowing each patch to apply attention to alternate patch in a single operation [26]. This gives the ViTs an advantage on the tasks where global context matters. For instance, in understanding how the configuration of one part of the hand can relate to the configuration of another.

Furthermore, Aly and Fathi [33] demonstrated a hybrid CNN-ViT architecture for ASL recognition that achieved 99.97% accuracy at 110 frames per second which delivered a result that is concurrently near-perfect and real-time, which was considered as a trade-off prior to this. The CNN component handled the low-level feature extraction competently, while the ViT component picked up the global spatial relationships which the CNN's local receptive fields would otherwise overlook. More recently, B. Alexander et al. [34] enhanced the ViT approach to video, using Video Vision Transformers (ViViTs) for word-level sign recognition and showed modelling temporal relationships through attention rather than through recurrence. This achieved competitive results without the need of training the instabilities that are associated with deep LSTMs.

C. Landmark-Based Versus Pixel-Based Approaches

Due to recent advances, a fundamental divergence in the SLR spatial systems. Perhaps you may provide the model raw RGB frames, or you may first extract a defined and systemized skeleton of key points and operate upon it? Each approach has authentic advantages; therefore, the optimal choice remains unresolved [35].

Hemanth et al. [36] demonstrated that MediaPipe based landmark extractions, which maps the hand and face into 3D skeletal coordinates, thus can achieve about 90% accuracy on sign language recognition by even using a classifier as simple as Random Forest. The reasons are transparent as a skeleton is largely invariant to lightings, skin tones, background, and even the camera quality. It represents precisely the information that differs one sign from another sign, with almost all of the visual noises are excluded. This is significantly a practical advantage for the deployment on edge devices where the computation is scarce.

The disadvantage in this is that the skeletons are representations so they discard key information. For instance, the precise texture of palm, the contact between the fingers, the slight curl of the knuckle; these all cues can differentiate between signs that share the same skeletal configuration, and that are invisible to landmark based systems. Pol et al. [37] conducted a direct comparison of landmark based systems and pixel-based systems on a same dataset and found that pixel-based models outperformed landmark based models on signs that required fine grained spatial distinctions, while landmark based models were more robust across environmental.

Abdullahi and Chamnongthai [38] proposed a middle ground stating that rather than choosing between raw pixels and negligible skeletons, they extracted prosodic (element of speech such as rhythm, intonation, pitch and stress etc.,) and angle features from the skeletal data; that is it captures not only where the hand is, but also the angles between the joints, and also the velocity of hand movement, and the rhythmic form of the signs. Their sequential learning approach on these landmark-derived features achieved around 98% accuracy on an ASL word recognition task. This suggested that rather than the contrast between landmarks and pixels, the question of which features to extract from a given representation appears have more importance.

D. Performance Gap: Isolated Versus Continuous Signing

Amongst all other findings, one of the most reliable findings in the SLR literature is that the systems which performs well on an isolated sign recognition worsens on continuous recognition, often significantly. Kumari and Anand [8] demonstrated this directly. Their CNN-LSTM attention model achieved an accuracy of 84.65% on WLASL-100 (World Level American Sign Language) continuous signing, contrasting with the typical 87 to 93% accuracy seen in the same architecture operated on isolated sign tasks. Hemanth et al. [37] also observed a similar pattern among multiple baseline models.

The gap reflects significant linguistic complexity. In continuous signing, signs do not have clear boundaries. The end of one sign influences the beginning of the next sign. This phenomenon is called co-articulation, and it is analogous to the way the spoken words blend together in a natural speech [14],[39],[40]. A model that has been trained to classify pre-segmented isolated signs has no clear

mechanism for handling this as it was tasked for them. The spatial features that it learned were clean and the features it faces at deployment are not clean. Walter et al. [7] identified this performance gap as one of the primary barriers between the current research systems and deployment that are ready for real-time.

IV. TEMPORAL MODELLING AND SEQUENCE LEARNING

To understand the spatial configuration of hand in a single frame is considered necessary but not sufficient. Sign language meaning is mostly executed in motion, that is, in the path a hand traces, the speed of a hand gesture, or the rhythm of a recurring movement. The temporal component of a SL recognition system is mainly responsible for extracting this key information from the sequences of frames. [38], [41], [37]

A. Lstm-Based Temporal Models

Long Short-Term Memory networks were amongst the first deep learning architectures that were applied to SLR temporal modelling, and they still remain widely used [42],[43]. Their gating mechanism, such as the input, forget, and output gates that regulate the flow of information through the network, allows them to retain memory over different sequences of varying length while selectively neglecting information that is no longer relevant and necessary.

Abdullahi and Chamnongthai [41] verified the effectiveness of multi-stacked Bidirectional LSTMs for ASL word recognition, achieving roughly 98% accuracy by processing sequences both forward direction and backward direction simultaneously. This bidirectional architecture is particularly well suited for sign recognition because the meaning of a current hand configuration often can be influenced by both what had come before and what may come after. Mittal et al. [15] showed that modified LSTM architectures combined with a Leap Motion hand tracking data could achieve robust results on continuous recognition tasks, and also demonstrated that the architecture is flexible enough to handle non-camera input sensory system as well.

The limitations of LSTMs become very apparent in any long continuous signing sequences. The vanishing gradient problem which means where backpropagating gradient signal becomes exponentially small over many time steps making learning difficult, this means that LSTMs effectively lose memory of what occurred more than about 100 frames ago. This is acceptable for short and dense vocabulary tasks. For extended continuous signing tasks, it is a real limitation. Baihan et al. [9] addressed this through the hybrid optimisation, combining the CNN spatial features with the LSTM temporal modelling and then applying a hybrid evolutionary gradient optimiser (CNNSa-LSTM) that improves training stability on long continuous sequences. Paul et al. [45] similarly found that Adam-optimised CNN-LSTM systems converged faster and generalised better than standard Stochastic Gradient Descent (SGD)-trained alternatives, suggesting that the procedure for optimisation matters as much as the architecture for real-world deployment.

B. 3d Convolutional Networks

Now we will see about the three-dimensional Convolutional Networks (3D-CNNs), it takes a diverse approach to temporal modelling. Rather than handling a sequence frame-by-frame and gathering temporal memory in a recurrent state, it treats a short video clip as a three-dimensional data volume (height × width × time) and the learned three-dimensional convolutional filters are then directly to this volume. This enables the network to learn spatial and temporal patterns as a single operation, without separating them processing into two different stages.

Sharma and Kumar [45] used 3D-CNNs specifically to ASL recognition (ASL-3DCNN) and found out that the architecture captured significant motion distinctions between similar signs more reliably than the 2D-CNN+LSTM baselines. But this cost high for the computational load, nearly 190 milliseconds per clip. Pol et al. [38] conducted a systematic one to one comparison of LSTM and 3D-CNN methods on the same continuous recognition task. He found that 3D-CNNs outperformed LSTMs on gestures where the main information was passed in the motion, while LSTMs outperformed 3D-CNNs on longer sequences where temporal context mattered most. The practical implication is that neither of the architecture takes lead unconditionally, therefore, the best choice hinge on the sign vocabulary and the constraints in the deployment.

C. Transformer-Based Temporal Models

Now transformers have become the dominant architecture for long-range temporal modelling in SLR, replacing LSTMs [46]. The fundamental advantage of the Transformer is its self-attention mechanism, it means that it can directly relate any two-time steps in a sequence with constant-time operation, in spite of their distance [47]. Whereas an LSTM needs to spread information through each and every intermediate steps to link frame 1 to frame 100, but a Transformer can be directly joined to any frame from any other frame [46].

Camgöz et al. [11] demonstrated this advantage decisively in their Sign Language Transformers paper (2020). By treating continuous sign language recognition and translation as a single task that is the model is trained to produce both a sign language recognition output and translation by word level simultaneously. And the result was that they achieved BLEU-4 scores on PHOENIX-2014T that significantly outdone the previous breakthroughs. Cui et al. [26] extended this with spatial-temporal Transformer that gives separate attention heads to spatial and temporal dimensions, enabling the model to give its attention mechanisms for each type of information rather than merging them.

Alyami and Luqman [12] introduced Swin-MSTP, which combines Swin Transformer backbone with multi-scale temporal perception module that handles temporal information at multiple time resolutions simultaneously. The awareness behind multi-scale temporal processing is that sign language contains structures at multiple timescales such as individual phoneme-level hand configurations changes at one speed, while sign-level gestures and sentence-level rhythms changes at different speed [14]. By processing all of these simultaneously, Swin-MSTP more than twice the BLEU-4 baseline scores of previous CNN-LSTM approaches on PHOENIX-2014T. Kiran et al. [48] found out similar advantages when relating Transformer-based architectures to sign language translation tasks over multiple vocabulary sizes.

For under-resourced sign languages where labelled training data is scarce, the standard supervised Transformer approach faces obvious limitations [49]. Aloysius, Kalaiselvi Geetha, and Nedungadi [50] addressed this through unsupervised pretraining: a Conformer model (which combines convolutional and self-attention layers) is first trained on large amounts of unlabelled signing video to develop general sign language representations, then fine-tuned on the small labelled dataset available. The results showed substantially stronger cross-domain generalisation than purely supervised baselines, suggesting that self-supervised pretraining may be the field's most practical path forward on data-scarce tasks.

V. REAL-TIME PREDICTION STABILITY

A system that achieves high accuracy on a test set but produces erratic predictions in a live video stream is not a useful system. Frame-to-frame noise is a pervasive problem in deployed SLR, and addressing it is as important as maximising raw accuracy. [36]

A. The Flickering Problem

Zuo, Wei, and Mak [51] described a problem called the flickering problem in continuous SLR. Since models are evaluated on each frame or small segment independently, any variations in lighting, or hand position, or camera motion can cause the prediction to change of flicker rapidly between the adjacent classes. The output of the system's best prediction changes numerous times in a second, which makes the output uncomprehensible and destroys the confidence in the system. This problem is particularly serious during sign transitions, where the hand is in an intermediate gesture that does not cleanly match to any sign.

B. Exponential Moving Average

The most common solution is Exponential Moving Average (EMA) smoothing, applied to the model's output probability distribution over time [52]. Zuo et al. [51] formulate this as:

$$(S_t = \alpha \cdot x_t + (1 - \alpha) \cdot S_t - 1)(1)$$

C. Sliding Window Majority Voting

Hemanth et al. [36] employ a sliding window majority voting strategy, which collects predictions over a window of N consecutive frames and outputs the class that appeared most frequently within the window. Unlike EMA, majority voting is fully committed — it always outputs the winning class, rather than a weighted blend — which makes it more interpretable in practice. The cost is a fixed latency equal

to the window length. Hemanth et al. found that windows of 15–30 frames (0.5–1.0 seconds at 30 FPS) provided a good balance between stability and responsiveness for their real-time system. The approach is now standard in deployed continuous SLR pipelines.

D. Confidence Thresholding And Rejection

As'ari et al. [54] introduce confidence thresholding as a complementary approach: rather than always producing a prediction, the system only outputs a recognised sign if the model's confidence exceeds a defined threshold. Below the threshold, the system outputs a null class — indicating that no recognisable sign is being produced, which is the correct answer during sign transitions or when the hand is at rest. This reduces spurious flicker substantially during transition periods and makes the output more honest about its own uncertainty. The threshold must be tuned carefully. in case if it is too high and the system loses genuine signs and if it is too low and it becomes vague from an unsmoothed baseline.

VI. HYBRID AND MULTIMODAL SYSTEM DESIGN

Most of the capable SLR systems in the current literature have one characteristic that is they do not commit to a single model type or a single data modality. The architectures that combine spatial and temporal strengths, or that combines multiple complementary data streams gives the best result. [7], [55], [56]

TABLE I Summary of Key Benchmark Results Across Reviewed Architectures

Model	Architecture	Dataset	Accuracy	Year
MobileNet [32]	Lightweight CNN [57]	STM32 Edge [57]	94.8% for 26.1	2023 [57]
CNN-LSTM Attn [8]	CNN-LSTM [8]	WLASL-100 [8]	84.65% [8]	2022 [8]
CNNSa-LSTM [9]	CNN-LSTM HO [9]	Custom [9]	98.7% [9]	2024 [9]
Bi-LSTM Stack [41]	RNN [38]	ASL Words [38]	98.6% [38]	2022 [38]
ASL-3DCNN [45]	3D CNN [59]	ASL [59]	99.38% [59]	2022 [59]
SL Transformer [11]	Transformer [11]	PHOENIX-14T [11]	21.80 BLEU-4 [61]	2020 [11]
Swin-MSTP [12]	Swin-T [12]	PHOENIX-14T [12]	~95% and above	2023 [12]
CNN-ViT [33]	CNN+ViT [33]	ASL [34]	98.97% [61]	2025 [33]

A. Cnn-Lstm Hybrids

CNN-LSTM hybrids are among the widely used and adapted hybrid architectures in sign language recognition as they efficiently integrate the spatial and temporal learning. Here, the CNN is responsible to extract the compact and discriminative spatial features from each video frames, while the LSTM handles how these features change and evolve during the signing sequences. As'ari et al. [54] proved the practical application of this hybrid architecture in specialised healthcare applications, where their SLR model performed reliably without any major architectural modifications. This highlights the flexibility of the model across different real-time scenarios.

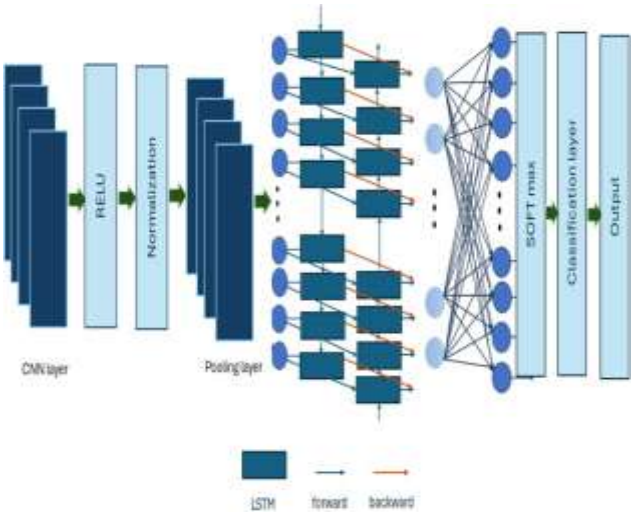

Fig 1. CNN_BiLSTM architecture by A. Abdelhamed et al. [58]

However, despite these advantages, several limitations of CNN-LSTM hybrid system have been identified. Baihan et al. [9] found out that training a standard CNN-LSTM model on long continuous signing sequence often led to instability

in gradient and slow convergence, making is unstable in learning and inefficient in computation. To handle this issue, they introduced a hybrid optimization strategy that combines evolutionary search technique with gradient descent.

In addition to all these, Kazbekova et al. [10] verified that by replacing standard LSTM layers with BiLSTM layers significantly improved contextual understanding in continuous sign language recognition. BiLSTM analyses the sequence in both forward and backward directions. Whereas normal LSTM only processes information from past frames alone. So BiLSTM allows the mode to interpret complex and ambiguous sign beginnings whose meanings becomes clear only after seeing the subsequent movements, this is a common phenomenon in natural continuous signing.

Therefore, BiLSTM-based architecture provides richer temporal context and improved recognition accuracy in complex signing tasks.

B. Transformer-Based Hybrids

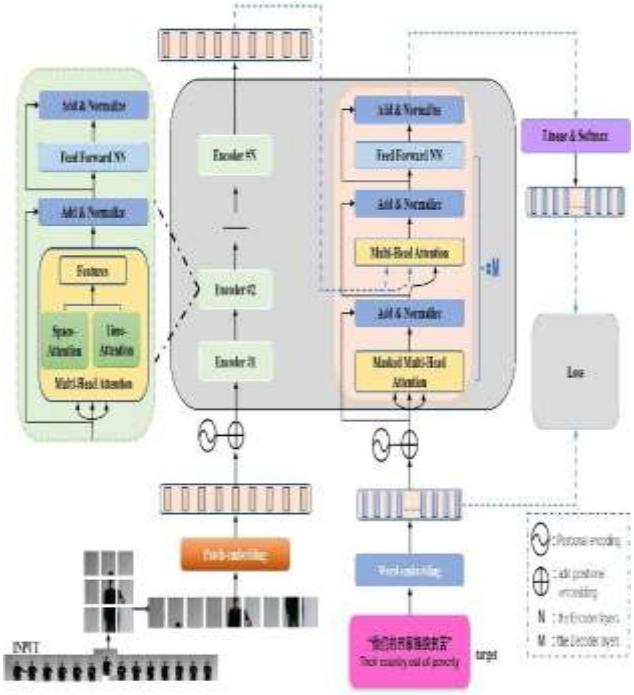

Fig 2 Transformer-based hybrid architecture by Cui et al. [26]

Shin et al. [62] proved the transferability of Transformer-based approaches to Korean Sign Language, achieving strong results without architectural changes by only retraining on language-specific datasets. This suggests that the architecture learns genuinely all general sign language representations rather than any specific ones. This adaptability highlights the potential of Transformer-based systems for developing the universal sign recognition frameworks.

C. Multimodal Fusion

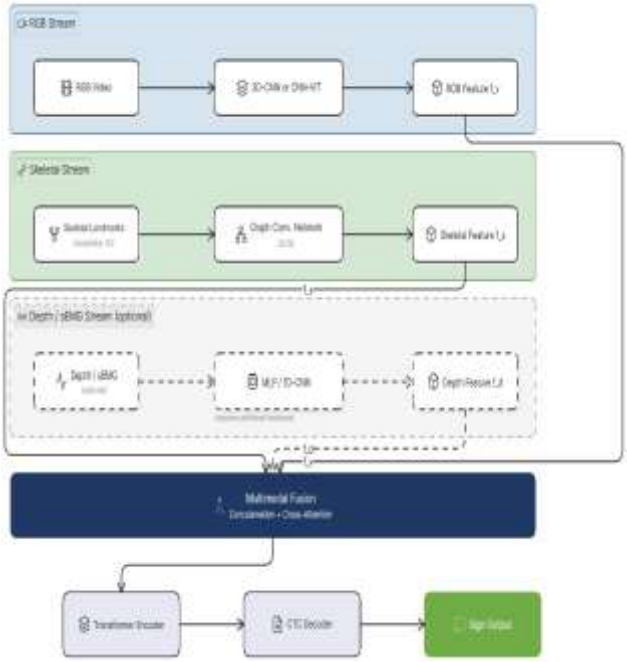

Fig 3 Multimodal architecture based on W. Vijitkunsawat et al. [63]

The most robust systems in the recent literature integrate multiple complementary data streams rather than relying on single modality [63]. Zhou et al. [56] combined RGB video with skeletal representation using a 3D Deformable CNN architecture, enabling the model to adapt its spatial sampling to the variations of the hand movements across frames. Their Multi-Stream Graph Convolutional Network (MS-GCN) component produces the relationships between skeletal key points as a graph structure, where the edges represent the anatomical connections between the joints. The resulting system showed improved robustness especially under challenging real-world conditions such as

partial occlusions, messy background and varying lightening, where conventional camera only systems often experience substantial degradation

Rakesh Kumar et al. [55] extended the multimodal method further by incorporating enhanced Transformer-based learning over integrated visual and positional feature streams. In their system, they used a Transformer encoder to integrate spatial (CNN-extracted) and temporal (LSTM-processed) features with the skeletal joint position embeddings, and achieved state-of-the-art results on their continuous recognition tasks. These development in the architecture reflects a broader trend that is rather than choosing among CNN, LSTM, Transformer, and skeletal approaches independently, the best system results from incorporate all of them, by using each of their strength where they contribute most [64].

VII. DOMAIN ADAPTATION AND GENERALISATION

Domain shift is the most significant practical problem in SLR. A model trained using a dataset captured in a controlled environment with proper lighting and background condition often gives a degraded performance when deployed in real-world scenarios. Variations such as difference in camera quality, lightings and signer appearance and their styles can reduce recognition accuracy by 20 to 40%. Walter et al. [7] verified and stated this degradation systematically across multiple model families and concluded that domain robustness, not raw accuracy, is the primary challenge for the real-world deployment.

A. Generative Approaches

Zhang, Chen, and Chen [65] addressed the problem of domain shift using a Generative Adversarial Approach (GAN) based framework called SignGAN. Their model is trained to transform signing videos captured in controlled environment into realistic representation of real-world condition while preserving the identity of the sign. The generator component learns to synthesise variation of the given sign captured under different conditions such as different lighting, different background, different camera angle and effectively developing domain-adapted training data without requiring additional real-world data. This approach demonstrated meaningful improvements in accuracy on cross-domain evaluation sets, though it requires careful architectural design and optimization to avoid mode collapse and to ensure that critical sign information is preserved through the domain transformation.

B. Self-Supervised Pretraining

Aloysius, Kalaiselvi Geetha, and Nedungadi [50] take a different approach: rather than transforming the training data, they change what the model learns before it ever sees labelled examples. Their Conformer model is first pre-trained on large volumes of unlabelled signing video through a masked prediction task, that is, the model learns to predict the content of masked video segments from surrounding context, in much the same way that BERT learns language representations from masked tokens in text. By the time the model is fine-tuned on the small labelled dataset, it has already learned rich general-purpose representations of how signing works. The result is substantially stronger cross-domain generalisation than a model trained from scratch on the same labelled data.

Algethami et al. [66] demonstrate that similar transfer learning principles apply across sign languages. Their work on Continuous Arabic Sign Language Recognition shows that models pre-trained on one sign language and fine-tuned on another achieve significantly better performance than models trained from scratch on the target language alone — even though Arabic and English sign languages share almost nothing lexically. The architectural representations, it appears, generalise at a level that is deeper than vocabulary.

C. Cross-Linguistic Generalisation

As we have seen earlier, Shin et al. [62] further provided evidence for transferability of a transformer-based architecture using their Korean Sign Language recognition study. They used a transformer model which was originally designed and evaluated on ASL recognition tasks, to achieve competitive result by retaining only the final classification layer on a language specified data. The spatial and temporal feature extractions were transferred successfully without structural modifications. This finding demonstrates cross-linguistic generalisation in transformer-based architecture. It suggests that the pretraining of a transformer architectures need to be

performed only once on large scale dataset, and the resulting learned representations can then be adapted to other sign languages with much smaller dataset requirements.

VIII. OPEN CHALLENGES

In this section we will see concerning several technical and social challenges that are unsolved despite the genuine progress documented throughout this review, as a substantial gap remains between what a system performs in controlled environment and real-world deployment.

A. Data Scarcity

The first challenge is data scarcity [67]. Deep learning models requires large number of datasets, however, SLR dataset remains limited. For example, WLASL, one of the largest available ASL datasets, contains approximately 21,000 videos across 2,000 word-level signs, represents only a small fraction of the vocabulary diversity found in natural communication [68]. Walter et al. [7] note that the limitation of data is not merely technical problem but also social. The effective collection and annotation of sign language video require the active participation of Deaf community members, and most existing datasets were collected by researchers with limited community participation. Consequently, how Deaf people actually sign based on regional variation, age-related differences, and individual signing style are underrepresented. Baihan et al. [9] stated that models trained on such limited datasets generalize poorly to signers not represented in training, which is a serious challenge for a technology that is meant to work for everyone.

B. Co-Articulation

The second challenge is Co-articulation. It is the blending of consecutive signs at their boundaries in natural communication and this remains one of the most difficult problems that is unsolved in SLR [79]. Zuo, Wei, and Mak [51] identify it as the primary factor contributing the performance gap between isolated and continuous sign recognition systems. The challenge is that there is no universal rule for how signs blend since the exact form of co-articulation vary depending on the signing speed, the signer's style, neighbouring signs and the sentence context [70]. As a result, a model that is trained to handle co-articulation for one dataset will often fail to generalise across other datasets [71]. Although, Aloysius et al. [50] showed that self-supervised pretraining helps in temporal understandings, but it did not fully resolved problem.

C. Sign Ambiguity And Prosody

Many sign language gestures differ only through subtle variations in movement, speed, orientation [68]. Abdullahi and Chamnongthai [38] documented several pairs of ASL signs which shares nearly identical hand shapes but vary in the speed of the movement, the path traced, or the rhythm of the hand. Distinguishing these signs requires models that are capable of learning prosodic information which are the signing equivalent of stress and intonation in speech. However, these informations are currently underrepresented in both the feature representation and optimization objectives [71]

D. Deployment Constraints

The third challenge is the deployment constrains. The gap between benchmark accuracy and real-world utility is an engineering problem as in real world, occlusions and background noise can significantly impair data quality [72]. As'ari et al. [54] noted that the systems which perform well on pre-processed benchmark video often fails when it is deployed on smartphone cameras, which have different noise characteristics, frame rates and environmental variation. Trpcheska et al. [31] stated that even MobileNet-based systems required extensive quantisation and pruning before achieving acceptable real-time performance. Walter et al. [7] provided deployment challenges that includes hardware variability, network latency, battery limitations and privacy concerns.

IX. FUTURE SCOPE

A. Self-Supervised And Semi-Supervised Learning

The most impactful near-term direction is probably self-supervised learning. The success of BERT and GPT in natural language processing — where models pre-trained on massive amounts of unlabelled text then fine-tune to specific tasks with a fraction of the data — has a clear analogue in sign language. Aloysius et al. [50] have demonstrated the proof of concept. The next steps are scaling the pretraining corpus (unlabelled signing video is relatively abundant) and developing pretraining objectives that are specifically suited to the structure of sign language, rather than borrowing directly from speech or text objectives.

B. Multimodal And Biosignal Integration

The performance ceiling of RGB-based systems is constrained by what is captured by camera alone. Zhou et al. [56] and M. Rakesh Kumar et al. [55] demonstrated that integrating RGB data with skeletal representation improves recognition robustness substantially. The next future step is integrating biosignals such as surface Electromyography (sEMG) which captures muscle activity from the forearm and provides information about the configuration of hand that is independent of variation in the visual condition [73]. Early work on EMG-augmented SLR has shown hopeful results [74], however, the hardware requirements currently limit their adaption.

C. Community-Centred Data Collection

As Walter et al. [7] argued that the data problem in SLR is not purely technical only but extends to social aspect. To solve this, we require the active participation of Deaf communities in the dataset design, data collection, and evaluation. These datasets designed by and is for deaf signers and will more clearly represent natural signing variation, hence, more useful for real-world deployment. In addition to all that is mentioned, this will be ethically defensible datasets collected primarily for research purpose [1]. Algethami et al. [66] demonstrated the importance of this for Arabic Sign Language, where community involvement in data collection led to more representative dataset compared to previous work.

D. Continuous, User-Independent Evaluation

The field needs evaluation benchmarks that should reflect the challenges encountered during real-world deployment. Zuo et al. [51] proposed evaluation protocols for continuous sign recognition that measures not just accuracy but also latency, stability, and the quality of segmentation. B.Alexander et al. [34] reasoned for word-level recognition benchmarks that include signers from diverse geographic regions, age groups, and signing backgrounds. Such evaluation would provide much easier way to identify systems that are genuinely suitable for deployment rather than systems that perform well on controlled tasks and not in practical deployment.

X. CONCLUSION

Sign language recognition has attained advancements due to the development of deep learning methods that are capable of combining spatial and temporal informations. The combination of CNNs for spatial feature extraction and LSTMs and Transformers for temporal modelling has resolved much of the core problem under controlled conditions. Transformer-based architectural models have set new standards on continuous recognition and translation tasks, and lightweight CNN backbones have made real-time processing feasible on edge hardware [27].

However, despite these achievements, several challenges remain unresolved. One of which is domain shift which continues to reduce performance when models are deployed in an environment that is different from their training conditions [69]. Also, there is co-articulation in continuous signing which remains a major obstacle to have an accurate recognition. In addition to these, the limitation in the dataset creation because of less involvement of communities raises concerns regarding quality of data and practical applicability.

The most promising future progress is the advances in self-supervised learning [49], multimodal integration and transfer learning. Cross-linguistic transfer learning, as demonstrated by Shin et al. [62], Algethami et al. [66], and Aloysius et al. [50], suggests that deep learning models may be capable of learning generalized sign language representations that can adapt multiple sign language even with limited retraining requirements.

Overall, the fundamental system required for effective sign language recognition are now well established. The next stage in progress will depend on the deployment of more representative dataset, robust framework and stronger collection of datasets with community involvement that can ensure future system are accurate and practical.

REFERENCES

1. D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. R. Morris, “Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective,” arXiv:1908.08597, 2019.

2. J. P. Morford and M. L. Carlson, “Sign perception and recognition in non-native signers of ASL,” Lang. Learn. Dev., vol. 7, no. 2, pp. 149–168, 2011, doi: 10.1080/15475441.2011.543393.
3. J. L. Mathews, A. L. Parkhill, D. A. Schlehofer, M. J. Starr, and S. Barnett, “Role-reversal exercise with Deaf Strong Hospital to teach communication competency and cultural awareness,” Am. J. Pharm. Educ., vol. 75, no. 3, p. 53, 2011, doi: 10.5688/ajpe75353.
4. N. Shahin and L. Ismail, “From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: Survey, taxonomy and performance evaluation,” Artif. Intell. Rev., vol. 57, p. 271, 2024, doi: 10.1007/s10462-024-10895-z.
5. “Correction,” Nature, vol. 448, p. 998, 2007, doi: 10.1038/448998b.
6. J. Jiang, “Sign Language Recognition Methods: Applications and Advances of Deep Learning Technology,” Highlights Sci. Eng. Technol., vol. 124, pp. 385–390, 2025, doi: 10.54097/ryyycp94.
7. A. K. Walter, G. Srivastava, and L. Kumari, “Sign language recognition in the deep learning era: A comprehensive study of model performance, robustness and deployment considerations,” Int. J. Sci. Res. Arch., vol. 15, no. 3, pp. 398–407, 2025, doi: 10.30574/ijsra.2025.15.3.1699.
8. D. Kumari and R. S. Anand, “Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism,” Electronics, vol. 13, no. 7, p. 1229, 2024, doi: 10.3390/electronics13071229.
9. A. Baihan, A. I. Alutaibi, M. Alshehri, et al., “Sign language recognition using modified deep learning network and hybrid optimization: A hybrid optimizer (HO)-based optimized CNNSa-LSTM approach,” Sci. Rep., vol. 14, p. 26111, 2024, doi: 10.1038/s41598-024-76174-7.
10. G. Kazbekova, Z. Ismagulova, G. Ibrayeva, A. Sundetova, Y. Abdrazakh, and B. Baimurzayev, “Real-time lightweight sign language recognition on hybrid deep CNN-BiLSTM neural network with attention mechanism,” Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 4, 2025, doi: 10.14569/IJACSA.2025.0160452.
11. N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden, “Sign Language Transformers: Joint end-to-end sign language recognition and translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 10020–10030, doi: 10.1109/CVPR42600.2020.01004.
12. S. Alyami and H. Luqman, “Swin-MSTP: Swin transformer with multi-scale temporal perception for continuous sign language recognition,” Neurocomputing, vol. 617, Art. no. 129015, 2025, doi: 10.1016/j.neucom.2024.129015.
13. L. Jamieson, F. Moreno-García, and E. Elyan, “A review of deep learning methods for digitisation of complex documents and engineering diagrams,” Artif. Intell. Rev., vol. 57, p. 136, 2024, doi: 10.1007/s10462-024-10779-2.
14. S. C. W. Ong and S. Ranganath, “Automatic sign language analysis: A survey and the future beyond lexical meaning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 873–891, Jun. 2005, doi: 10.1109/TPAMI.2005.112.
15. A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, “A modified LSTM model for continuous sign language recognition using Leap Motion,” IEEE Sensors J., vol. 19, no. 16, pp. 7056–7063, Aug. 2019, doi: 10.1109/JSEN.2019.2909837.
16. G. Bailador, C. Sanchez-Avila, J. Guerra-Casanova, and A. de Santos Sierra, “Analysis of pattern recognition techniques for in-air signature biometrics,” Pattern Recognit., vol. 44, no. 10–11, pp. 2468–2478, 2011, doi: 10.1016/j.patcog.2011.04.010.
17. W. N. Khotimah, T. Anggita, and N. Suciati, “Indonesian sign language recognition using Kinect and dynamic time warping,” Indones. J. Electr. Eng. Comput. Sci., vol. 15, no. 1, pp. 495–503, 2019, doi: 10.11591/ijeecs.v15.i1.pp495-503.
18. M. A. Hamza, A. Subahi, N. A. Alghanmi, et al., “Deep fusion based transfer learning with bald eagle search algorithm for sign language recognition to assist individuals with hearing and speech impairments,” Sci. Rep., vol. 15, p. 32752, 2025, doi: 10.1038/s41598-025-10660-4.
19. I. A. Adeyanju, O. O. Bello, and M. A. Adegboye, “Machine learning methods for sign language recognition: A critical review and analysis,” Intell. Syst. Appl., vol. 12, Art. no. 200056, 2021, doi: 10.1016/j.iswa.2021.200056.
20. O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, “Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in

sign language videos," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 9, pp. 2306–2320, Sept. 2020, doi: 10.1109/TPAMI.2019.2911077.
21. O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep Sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," Int. J. Comput. Vis., vol. 126, pp. 1311–1325, 2018, doi: 10.1007/s11263-018-1121-3.
22. R. Rastgoo, K. Kiani, and S. Escalera, "Word separation in continuous sign language using isolated signs and post-processing," Expert Syst. Appl., vol. 249, pt. B, Art. no. 123695, 2024, doi: 10.1016/j.eswa.2024.123695.
23. S. Yang and Q. Zhu, "Continuous Chinese sign language recognition with CNN-LSTM," in Proc. 9th Int. Conf. Digit. Image Process. (ICDIP), SPIE, vol. 10420, 2017, Art. no. 104200F, doi: 10.1117/12.2281671.
24. P. Xie, M. Zhao, and X. Hu, "PiSLTRc: Position-informed sign language transformer with content-aware convolution," IEEE Trans. Multimedia, vol. 24, pp. 3908–3919, 2022, doi: 10.1109/TMM.2021.3109665.
25. K. Alomar, H. I. Aysel, and X. Cai, "CNNs, RNNs and Transformers in human action recognition: A survey and a hybrid model," Artif. Intell. Rev., vol. 58, p. 387, 2025, doi: 10.1007/s10462-025-11388-3.
26. Z. Cui, W. Zhang, Z. Li, and Z. Wang, "Spatial–temporal transformer for end-to-end sign language recognition," Complex Intell. Syst., vol. 9, pp. 4645–4656, 2023, doi: 10.1007/s40747-023-00977-w.
27. S. Pandey, S. Tahseen, R. Pathak, H. Parveen, and M. Maurya, "Real-time vision-based Indian sign language translation using deep learning techniques," Int. J. Innov. Res. Comput. Sci. Technol., vol. 13, no. 3, 2025, doi: 10.55524/ijircst.2025.13.3.6.
28. W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, "BEST: BERT pre-training for sign language recognition with coupling tokenization," Proc. AAAI Conf. Artif. Intell., vol. 37, no. 3, pp. 3597–3605, 2023, doi: 10.1609/aaai.v37i3.25470.
29. C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1915–1929, Aug. 2013, doi: 10.1109/TPAMI.2012.231.
30. M. Fernández-Sanjurjo, M. Mucientes, and V. M. Brea, "Real-time multiple object visual tracking for embedded GPU systems," IEEE Internet Things J., vol. 8, no. 11, pp. 9177–9188, Jun. 2021, doi: 10.1109/JIOT.2021.3056239.
31. A. Trpcheska, F. Zevnik, and S. Bader, "Towards real-time vision-based sign language recognition on edge devices," in Proc. IEEE Sensors Appl. Symp. (SAS), Naples, Italy, 2024, pp. 1–6, doi: 10.1109/SAS60918.2024.10636604.
32. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
33. M. Aly and I. S. Fathi, "Recognizing American Sign Language gestures efficiently and accurately using a hybrid transformer model," Sci. Rep., vol. 15, Art. no. 20253, 2025, doi: 10.1038/s41598-025-06344-8.
34. A. Brettmann, J. Grävinghoff, M. Rüschoff, and M. Westhues, "Breaking the barriers: Video vision transformers for word-level sign language recognition," arXiv:2504.07792, 2025.
35. R. K. Pathan, M. Biswas, S. Yasmin, et al., "Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network," Sci. Rep., vol. 13, Art. no. 16975, 2023, doi: 10.1038/s41598-023-43852-x.
36. S. V. Hemanth, V. D. Ram, R. A. Raj, P. Nithin, V. Niharika, and T. B. Kumar, "Real-time sign language recognition using advanced computer vision," Int. Res. J. Adv. Eng. Hub, vol. 3, no. 5, 2025, doi: 10.47392/IRJAEH.2025.0368.
37. A. Anturkar, A. Khot, A. Andure, A. Ghosh, A. Magadum, A. Bahadur, and M. Pol, "Real-time sign language to text translation using deep learning: A comparative study of LSTM and 3D CNN," Int. J. Comput. Appl., vol. 187, no. 55, 2025, doi: 10.5120/ijca2025925946.
38. S. B. Abdullahi and K. Chamnongthai, "American Sign Language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach," IEEE Access, vol. 10, pp. 15911–15923, 2022, doi: 10.1109/ACCESS.2022.3148132.
39. A. Duarte et al., "How2Sign: A large-scale multimodal dataset for continuous American Sign Language," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, 2021, pp. 2734–2743, doi: 10.1109/CVPR46437.2021.00276.

40. M. Gonzalez, "Computer vision methods for unconstrained gesture recognition in the context of sign language annotation," Ph.D. dissertation, Université Paul Sabatier–Toulouse III, Toulouse, France, 2012.
41. S. B. Abdullahi and K. Chamnongthai, "American sign language words recognition of skeletal videos using processed video driven multi-stacked deep LSTM," Sensors, vol. 22, no. 4, Art. no. 1406, 2022.
42. E. Tsironi, P. Barros, C. Weber, and S. Wermter, "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition," Neurocomputing, vol. 268, pp. 76–86, 2017, doi: 10.1016/j.neucom.2016.12.088.
43. S. A, R. R, G. J, H. Kuresan, A. A. Bala, and G. E. Visuvanathan, "Hand sign recognition using deep learning models," in Proc. Int. Conf. Recent Adv. Electr., Electron., Ubiquitous Commun., Comput. Intell. (RAEEUCCI), Chennai, India, 2025, pp. 1–9, doi: 10.1109/RAEEUCCI63961.2025.11048208.
44. K. Paul et al., "An Adam based CNN and LSTM approach for sign language recognition in real time for deaf people," Bull. Electr. Eng. Informat., vol. 13, no. 1, pp. 499–509, 2024, doi: 10.11591/eei.v13i1.6059.
45. S. Sharma and K. Berwal, "ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks," Multimedia Tools Appl., vol. 80, pp. 1–13, 2021, doi: 10.1007/s11042-021-10768-5.
46. N. Ranasinghe, D. Do-Ha, S. Maksour, T. Malepathirana, S. Seneviratne, L. Ooi, and S. Halgamuge, "FRAME-C: A knowledge-augmented deep learning pipeline for classifying multi-electrode array electrophysiological signals," arXiv:2505.18183, 2025.
47. H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," Proc. AAAI Conf. Artif. Intell., vol. 35, no. 12, pp. 11106–11115, 2021, doi: 10.1609/aaai.v35i12.17325.
48. G. R. Kiran, V. Srinadh, V. Ankitha, and S. Surekha, "Sign language translator using transformer model," Int. J. Environ. Sci., pp. 3285–3298, 2025, doi: 10.64252/9j23fm84.
49. H. Hu, W. Zhao, W. Zhou, and H. Li, "SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 9, pp. 11221–11239, Sept. 2023, doi: 10.1109/TPAMI.2023.3269220.
50. N. Aloysius, G. M, and P. Nedungadi, "Continuous sign language recognition with adapted conformer via unsupervised pretraining," arXiv:2405.12018, 2024.
51. R. Zuo, F. Wei, and B. Mak, "Towards online continuous sign language recognition and translation," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Miami, FL, USA, 2024, pp. 11050–11067, doi: 10.18653/v1/2024.emnlp-main.619.
52. M. As'ari, N. Sufri, and G. Qi, "Emergency sign language recognition from variant of convolutional neural network (CNN) and long short term memory (LSTM) models," Int. J. Adv. Intell. Informat., vol. 10, no. 1, pp. 64–78, 2024, doi: 10.26555/ijain.v10i1.1170.
53. M. R. Kumar, P. Kumar, and R. Vaseekaran, "Multi-modal fusion and transformer-enhanced learning for real-time continuous sign language recognition," in Proc. 3rd Int. Conf. Self Sustainable Artif. Intell. Syst. (ICSSAS), Erode, India, 2025, pp. 310–316, doi: 10.1109/ICSSAS66150.2025.11081311.
54. Q. Zhou, H. Li, W. Meng, H. Dai, T. Zhou, and G. Zheng, "Fusion of multimodal spatio-temporal features and 3D deformable convolution based on sign language recognition in sensor networks," Sensors, vol. 25, no. 14, Art. no. 4378, 2025, doi: 10.3390/s25144378.
55. J. C. N. Bittencourt and W. C. J. Rocha, "Optimisation techniques for compact CNN on embedded systems for gesture recognition," U.Porto J. Eng., vol. 9, no. 5, pp. 65–76, 2023, doi: 10.24840/2183-6493_009-005_002156.
56. M. A. Abdelhamed, M. Samy, B. E. Elnaghi, and A. Magdy, "A hybrid CNN-BiLSTM deep learning framework for signal detection of a massive MIMO-NOMA system," Phys. Commun., vol. 66, Art. no. 102450, 2024.
57. A. Kasapbaşı, A. E. A. Elbushra, O. Al-Hardanee, and A. Yilmaz, "DeepASLR: A CNN-based human-computer interface for American Sign Language recognition for hearing-impaired individuals," Comput. Methods Programs Biomed. Update, vol. 1, Art. no. 100048, 2021, doi: 10.1016/j.cmpbup.2021.100048.
58. K. Yin and J. Read, "Better sign language translation with STMC-Transformer," in Proc. 28th Int. Conf.

Comput. Linguistics (COLING), Barcelona, Spain, 2020, pp. 5975–5989.
59. E. L. R. Ewe, C. P. Lee, K. M. Lim, L. C. Kwek, and A. Alqahtani, "LAVRF: Sign language recognition via lightweight attentive VGG16 with random forest," PLoS One, vol. 19, no. 4, 2024, doi: 10.1371/journal.pone.0298699.
60. J. Shin, A. S. M. Miah, M. A. M. Hasan, K. Hirooka, K. Suzuki, H.-S. Lee, and S.-W. Jang, "Korean sign language recognition using transformer-based deep neural network," Appl. Sci., vol. 13, no. 5, Art. no. 3029, 2023, doi: 10.3390/app13053029.
61. W. Vijitkunsawat and T. Racharak, "GSR-Fusion: A deep multimodal fusion architecture for robust sign language recognition using RGB, skeleton, and graph-based modalities," IEEE Access, vol. 13, pp. 108235–108254, 2025, doi: 10.1109/ACCESS.2025.3581683.
62. S. S. J., E. A. Suvi, R. J. Jajin, A. Jagan, and S. P. S., "Sign language recognition using deep learning: A systematic review of models and approaches," Int. J. Innov. Res. Comput. Sci. Technol., vol. 14, no. 1, pp. 34–42, 2026, doi: 10.55524/ijircst.2026.14.1.5.
63. H. Zhang, X. Chen, and S. Chen, "Cross-domain Wi-Fi sign language recognition with GANs," in Proc. 10th Int. Conf. Commun. Broadband Netw. (ICCBN), New York, NY, USA, 2022, pp. 60–65, doi: 10.1145/3538806.3538810.
64. N. Algethami, R. Farhud, M. Alghamdi, H. Almutairi, M. Sorani, and N. Aleisa, "Continuous Arabic sign language recognition models," Sensors, vol. 25, no. 9, Art. no. 2916, 2025, doi: 10.3390/s25092916.
65. S. F. Ahmed, M. S. B. Alam, M. Hassan, et al., "Deep learning modelling techniques: Current progress, applications, advantages, and challenges," Artif. Intell. Rev., vol. 56, pp. 13521–13617, 2023, doi: 10.1007/s10462-023-10466-8.
66. D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Snowmass, CO, USA, 2020, pp. 1448–1458, doi: 10.1109/WACV45572.2020.9093512.
67. Y. Li, X. Geng, Z. Ma, Q. Miao, and C.-M. Pun, "Boundary-aware sentence-gloss alignment with semantic similarity measurement for continuous sign language recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 35, no. 11, pp. 11390–11403, Nov. 2025, doi: 10.1109/TCSVT.2025.3572520.
68. M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," IEEE Access, vol. 9, pp. 126917–126951, 2021, doi: 10.1109/ACCESS.2021.3110912.
69. M. Inan, Y. Zhong, S. Hassan, L. Quandt, and M. Alikhani, "Modeling intensification for sign language generation: A computational approach," in Findings Assoc. Comput. Linguistics: ACL 2022, Dublin, Ireland, 2022, pp. 2897–2911.
70. H. ZainEldin, S. A. Gamel, F. M. Talaat, et al., "Silent no more: A comprehensive review of artificial intelligence, deep learning, and machine learning in facilitating deaf and mute communication," Artif. Intell. Rev., vol. 57, p. 188, 2024, doi: 10.1007/s10462-024-10816-0.
71. J. Wu, L. Sun, and R. Jafari, "A wearable system for recognizing American Sign Language in real-time using IMU and surface EMG sensors," IEEE J. Biomed. Health Informat., vol. 20, no. 5, pp. 1281–1290, Sept. 2016, doi: 10.1109/JBHI.2016.2598302.
72. V. E. Kosmidou, L. J. Hadjileontiadis, and S. M. Panas, "Evaluation of surface EMG features for the recognition of American Sign Language gestures," in Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS), New York, NY, USA, 2006, pp. 6197–6200, doi: 10.1109/IEMBS.2006.259428