Multimodal Technologies and Interaction

MDPI

Review

A Comprehensive Review of Deep Learning Approaches for Video-Based Sign Language Recognition: Datasets, Challenges and Insights

Ulmeken Berzhanova [1,2], Aigerim Yerimbetova [1,3,*], Marek Milosz [1,4], Bakzhan Sakenov [1], Dina Oralbekova [1], Elmira Daiyrbayeva [1,5,*] and Daniyar Turgan [5]

1 Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan; berzhanovaulmekenn@gmail.com (U.B.); m.milosz@pollub.pl (M.M.); sbakzhan22@gmail.com (B.S.); dinaoral@mail.ru (D.O.)
2 Faculty of Information Technology and Artificial Intelligence, Farabi University, Almaty 050038, Kazakhstan
3 School of Engineering and Information Technology, META University, Almaty 050012, Kazakhstan
4 Department of Computer Science, Faculty of Electrical Engineering and Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland
5 Department of Software Engineering, Satbayev University, Almaty 050010, Kazakhstan; daniyar.turgan@stud.satbayev.university
* Correspondence: aigerian8888@gmail.com (A.Y.); e.daiyrbayeva@satbayev.university (E.D.)

Abstract

This study presents a comprehensive review of more than 100 research papers on sign language recognition (SLR) published between 2020 and 2026. The analysis focuses on deep learning approaches applied to video-based SLR, including spatiotemporal feature extraction, temporal modeling, attention mechanisms, motion-based representations, hybrid frameworks, transfer learning methods and other methods. Particular attention is given to how these methods model spatiotemporal dynamics and capture subtle gesture characteristics in sign language communication. The review highlights several recent developments, such as the introduction of specialized datasets, the emergence of real-time recognition systems, and the integration of multimodal fusion strategies. At the same time, persistent challenges remain, including data scarcity in low-resource sign languages, limited linguistic standardization of datasets, and insufficient model interpretability. The findings underline the importance of developing scalable and generalizable models capable of handling diverse datasets and user variability. The distinct contributions of this review are fourfold: (1) a comprehensive synthesis of over 100 studies published between 2020 and 2026, covering the full spectrum of deep learning architectures for video-based SLR; (2) a structured six-category taxonomy enabling systematic cross-architectural comparison; (3) a comprehensive focus on low-resource sign languages, which remain underrepresented in the existing literature; and (4) a critical analysis of the current benchmark landscape for low-resource sign languages, identifying key gaps and outlining strategic directions for future dataset development. These contributions are intended to guide further research toward more robust, inclusive, and universally applicable SLR systems.

Keywords: sign language; sign language recognition; video recognition; low-resource languages; artificial intelligence; deep learning; datasets

Check for updates

Academic Editor: Stephan Schlögl

Received: 28 March 2026
Revised: 9 May 2026
Accepted: 13 May 2026
Published: 22 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

SLR systems are fundamental technological tools in creating an inclusive information society by providing equal access to digital resources for people with hearing and speech impairments. As part of this challenge, the issue of video-based SLR is scientifically relevant, particularly due to the need for a special approach that takes into account the unique vocabulary, grammar, and ethnocultural features of the language, which are not fully covered by current universal models [1–5].

Sign language is a rich and complex means of communication, involving spatiotemporal hand movement, facial expression, and body position. For its effective recognition, it is necessary to integrate modern computer vision with Natural Language Processing (NLP) techniques, and take into account the limited computing resources available for applications on mobile devices and in cloud environments [6–10].

In the current technological landscape, the development of SLR systems is gaining particular importance as artificial intelligence and computer vision become central tools for building inclusive communication solutions. These systems serve as critical interfaces between deaf and hearing communities, enabling access to digital services and social participation for people with hearing and speech impairments. These systems for people with hearing and speech impairments perform the function of communication with the outside world. SLs represent complex and distinctive linguistic systems with their own structure, grammar, and expressive mechanisms.

Video recognition in sign language faces various challenges, including the complexity of the data collection and annotation process, the uncertainty in the signal, and the underdevelopment of the architecture that is balanced for context and the use of deep learning architectures capable of effectively capturing contextual information and adapting to the specific characteristics of sign languages [11–15]. Addressing these challenges requires advanced video stream analysis techniques, including object detection, motion tracking, and sequential classification, which form the technical foundation of modern SLR systems [16–20]. Figure 1 demonstrates the architecture of the collection processing system used for sign language video recognition.



Figure 1. Process flow for SLR.

This review examines studies related to SLR and explores the importance of developing "inclusive" models that support low-resource languages, as well as enabling their use in international SLs. The implementation of SLR systems opens the way to inclusive growth in the fields of education, employment, and social communication [21–24]. This review will review current approaches to video recognition in SL, with a focus on the low-resource Kazakh Sign Language (KSL), and suggest areas for further research and development. The development of SLR systems requires a complex architecture that includes video stream processing, feature extraction, and classification components. Solutions for this architecture should be tailored to the unique characteristics of SL, taking into account spatial and temporal aspects, as well as cultural features.

As shown in Figure 2, the structure of the sign language video recognition system includes several main components:

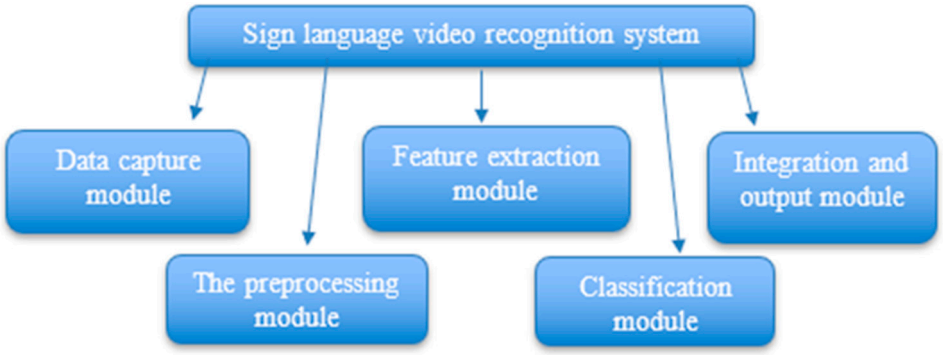


Figure 2. Architecture showing the structure of video-based SLR systems.

1. Data Capture Module. The data capture module contains devices that capture a video stream, such as a high-resolution camera. The main requirement is the ability to accurately capture human hand movements and facial expressions, as well as body shape.

2. Pre-Processing Module. The pre-processing stage aligns frames, removes noise, restores data, and extracts regions of interest (e.g., hand gestures and facial expressions) within them, which is crucial for reducing computational overhead and improving the quality of data sent to subsequent processing stages.

3. Feature Extraction Module. Deep learning techniques such as convolutional neural networks (CNNs) can be used to extract spatial and temporal features. For example, a three-dimensional convolutional neural network (3D CNN) can be used for motion analysis in combination with a Long Short-Term Memory (LSTM) or a Gated Recurrent Unit (GRU).

Spatiotemporal representations are typically learned using deep neural network architectures. Convolutional layers capture spatial patterns across frames, while short-term relationships are modeled using cyclic mechanisms such as LSTM or GRU modules. In dynamic gesture recognition, three-dimensional convolutional neural networks (3D CNNs) are often integrated with cyclic layers for synchronous encoding of motion dynamics and sequential context.

4. Classification Module. This module performs gesture analysis and thoroughly interprets the sifted feature images. Transformer-based architectures model continuous sub-constraints in sequences, allowing them to effectively encode contextual relationships for subsequent classification.

5. Integration and Output Module. A special module for displaying gestures allows the user to initiate the solution of expected gestures in the text or voice version, which increases user satisfaction and adaptability.

A highly reliable SLR and accurate detection capability require a comprehensive integration of all components [4]. Despite the public recognition of inclusive communication, efforts in this area face significant obstacles, including variable levels of language complexity, insufficient resources, and limited hardware capabilities Therefore, a design that provides high reliability must provide a balance between computational power, quality assurance, integrity of values, and availability across a wide range of deployment scenarios.

The study provides a comprehensive overview of current deep learning strategies used in SLR, focusing on their methodological peculiarities and inherent limitations. The present study synthesizes the results of further research conducted after the video data, with a view to identifying current research directions and incomplete solutions to current research directions and technical challenges.

The present review also contributes a structured analytical framework for evaluating deep learning approaches across low-resource sign language contexts, with KSL serving as a representative case study. This paper outlines future research directions in SLR, emphasiz-

ing the need for generalizable, computationally efficient, and linguistically inclusive models capable of performing across diverse datasets and real-world deployment conditions.

As shown in Table 1, the present review differs from prior surveys in several substantive respects. While existing reviews provide useful overviews of CNN and RNN-based approaches, none of them provide comprehensive coverage of Transformer and GCN-based architectures for video-based SLR, nor do they address low-resource sign languages or follow a formal systematic methodology with explicit inclusion criteria and bias assessment. The proposed six-category taxonomy was derived inductively from the systematic analysis of 103 included studies rather than adopted from prior classification schemes, and reflects the specific architectural landscape of video-based SLR as it has evolved between 2020 and 2026. This taxonomy integrates temporal modeling capacity, multimodal fusion capability, and real-world applicability as organizing principles, i.e., the criteria that are absent from the categorical frameworks used in prior reviews. Together, these distinctions substantiate the originality of the present work and its contribution to the systematic understanding of deep learning approaches for video-based SLR.

Table 1. Comparison of the present review with closely related survey articles.

Dimension	Al Abdullah et al. [18]	Violet et al. [20]	Present Review
Publication year	2024	2025	2026
Papers covered	∼80	∼60	103
Period covered	Up to 2023	Up to 2024	2020–2026
Focus on video-based SLR	Partial	Partial	Exclusive
CNN/recurrent neural network (RNN) coverage	Yes	Yes	Yes, comprehensive
Transformer-based models	Limited	Limited	Yes, comprehensive
Graph convolutional network (GCN)-based models	Not covered	Limited	Yes, comprehensive
Self-supervised/transfer learning	Not covered	Not covered	Yes, covered
Structured architectural taxonomy	Not provided	Not provided	Yes, six-category
Low-resource sign languages	Not addressed	Not addressed	Yes, dedicated focus
Systematic PRISMA methodology	No	No	Yes
Bias assessment	No	No	Yes
Integrative discussion with roadmap	No	No	Yes

Several systematic reviews on deep learning-based SLR have been published in recent years [18]; however, most existing reviews either focus narrowly on isolated gesture recognition, cover only a limited architectural scope, or do not address low-resource sign languages. The present review distinguishes itself in three key respects. First, it provides a unified six-category taxonomy that systematically organizes approaches ranging from spatial feature extraction to self-supervised multi-stream learning, enabling structured cross-architectural comparison. Second, it explicitly addresses the challenge of low-resource SLR with a dedicated focus on KSL as a representative underrepresented case. Third, the formal systematic methodology includes explicit inclusion and exclusion criteria, a four-stage PRISMA-compliant filtering process, and bias assessment that ensures reproducibility and transparency that distinguish the present work from narrative surveys in the field.

2. Methodology

This review adopts a structured, systematic methodology to ensure transparency, reproducibility, and comprehensive coverage of recent advances in video-based SLR using

deep learning techniques. The overall approach is aligned with established systematic review practices and is designed to provide a consistent and unbiased analysis of the literature.

A systematic search was conducted across four major scientific databases: Scopus, Web of Science, IEEE Xplore, and Google Scholar. The search was performed in January 2026 and covered publications from January 2020 to January 2026 in order to capture the most recent developments in the field. The search queries were applied to titles, abstracts, and keyword fields using combinations of terms related to sign language recognition, gesture recognition, deep learning, and video-based analysis. Specifically, the following query structure was used: ("sign language recognition" OR "gesture recognition") AND ("deep learning" OR "neural network" OR "convolutional neural network" OR "transformer" OR "recurrent neural network") AND ("video" OR "video-based" OR "continuous"). In addition, targeted searches were conducted to ensure adequate coverage of specific methodological directions, including graph convolutional networks, attention-based architectures, continuous recognition systems, and approaches designed for low-resource sign languages.

To ensure methodological consistency, explicit inclusion and exclusion criteria were applied. Studies were included if they were published in peer-reviewed journals or conference proceedings within the specified time frame, focused on video-based SLR using at least one deep learning component, reported quantitative performance metrics such as accuracy, Word Error Rate (WER), BLEU score, or mean Average Precision (mAP), and were available in English as full-text publications. Studies were excluded if they addressed only static image-based gesture recognition without temporal modeling, relied exclusively on traditional machine learning methods without deep learning, lacked sufficient methodological detail, or represented duplicate or superseded versions of previously published work.

The study selection process followed a PRISMA-compliant multi-stage filtering procedure. Initially, 1050 records were identified across all databases. After the removal of approximately 320 duplicate entries, 730 unique records remained. In the screening stage, titles and abstracts were reviewed, resulting in the exclusion of 540 studies that were not directly related to video-based SLR or did not involve deep learning techniques. The remaining 190 articles were subjected to full-text evaluation, during which 87 studies were excluded due to insufficient methodological rigor, lack of accessibility, or misalignment with the scope of this review. Ultimately, 103 studies were retained for detailed analysis. The full selection process is illustrated in a PRISMA flow diagram (Figure 3).

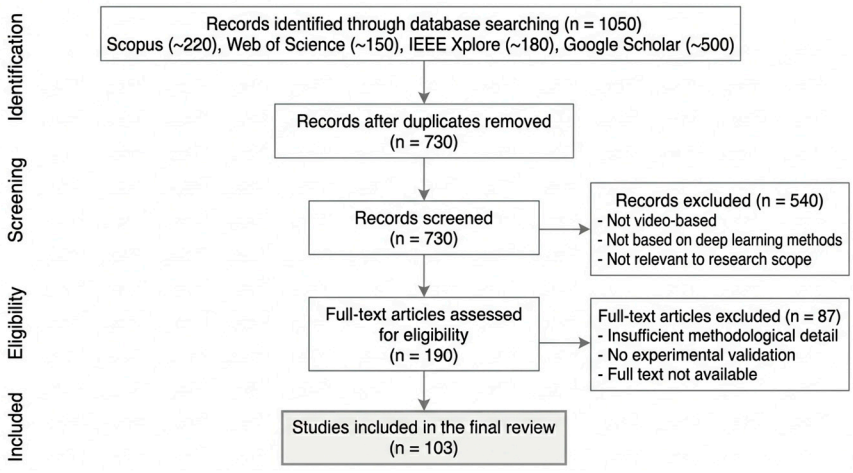


Figure 3. PRISMA flow diagram of the study selection process.

For each selected study, relevant information was systematically extracted, including publication metadata, the sign language addressed, the type of deep learning architecture

employed, datasets used, evaluation metrics, reported performance, and identified limitations. Based on this information, the studies were categorized according to the taxonomy proposed in Section 3, enabling a structured comparison of existing approaches.

To reduce selection bias, the search and screening procedures were conducted independently by two authors, with any disagreements resolved through discussion. Nevertheless, certain limitations remain. In particular, a degree of publication bias is expected, as many studies report relatively high performance metrics, which may reflect a tendency to publish positive results. Additionally, no formal risk-of-bias assessment tool was applied, as existing instruments are primarily designed for clinical research contexts. Instead, methodological quality was evaluated based on criteria such as dataset scale, experimental design, clarity of evaluation protocols, and reproducibility of results.

Overall, the adopted methodology ensures a transparent and reproducible framework for analyzing the current state of deep learning approaches in video-based SLR.

3. Taxonomy of Deep Learning Approaches for SLR

This review provides a systematic analysis of video-based SLR technologies, focusing on systems designed to support people with hearing and speech impairments. This review also focuses on modern deep learning architectures, highlighting key industry challenges such as high tagging variability, lack of annotated datasets, and significant computational complexity.

To ensure a coherent and structured analysis of the literature, existing approaches are divided into major technological paradigms. These include methods for acquiring spatial and spatiotemporal objects, approaches for modeling time sequences, Transformer-based and focusing architectures, graph-based spatial and temporal modeling methods, hybrid full SLR and sign language translation systems, and transfer learning strategies and other auxiliary methods.

This comprehensive collection of publications ensures coverage of recent industry impacts and contributions, and was selected based on rigorous criteria of relevance and quality. This systematic classification allows for a critical evaluation of current achievements, identification of methodological trends, and identification of challenges and promising areas open to future research.

Bubble position reflects temporal modeling capacity (vertical axis) and data efficiency (horizontal axis) (See Figure 4). Dashed arrows indicate the main evolutionary progression across architectural generations, from spatial feature extraction approaches toward hybrid end-to-end systems.

3.1. Spatial and Spatiotemporal Feature Extraction Approaches

One of the most widely used approaches for video-based SLR is the use of deep learning architectures to obtain spatial and spatiotemporal features. This methodology is effective in mitigating the problem of insufficient data and improving the representation of features similar to those that can be extracted from higher levels of visual abstraction [25].

Dynamic gesture recognition in American Sign Language (ASL) has been widely studied, particularly with 3D CNN-based approaches that have achieved good results. These models can classify up to 100 words in real time with high accuracy, although efficiency is often limited by the volume and diversity of available datasets [26]. Malaysian Sign Language (MSL) has been shown to achieve over 90% accuracy using the Convolutional Block Attention Module-ResNet (CBAM-ResNet) architecture. However, the results appear to be highly dependent on dataset quality, with over-classification and generalization decreasing in more complex and invisible scenarios [27].

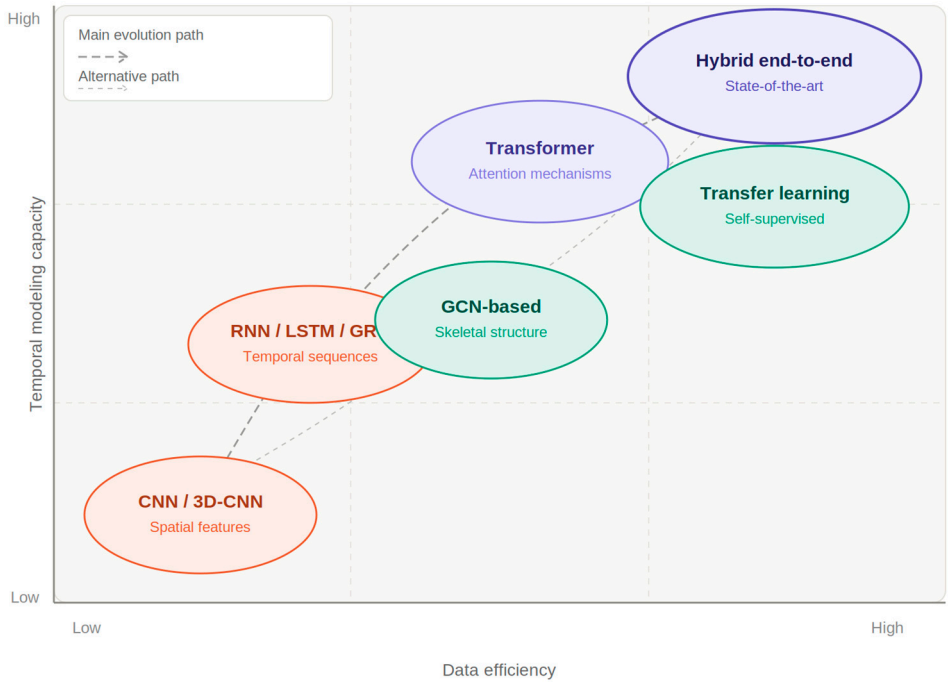


Figure 4. Evolutionary trajectory of deep learning architectures for video-based SLR.

Comparative studies of CNN and 3D CNN models have shown that 3D CNNs offer the best performance, achieving 91% accuracy in Trinidadian and Tobago Sign Language (TTSL) and 83% accuracy on the ASL dataset. Despite these good results, the evaluation of resilience to dataset and environmental changes has not been adequately considered, limiting the practical application of such systems in real-world interactive environments [28].

Building on previous research, this study [29] presents a deep learning-based SLR system aimed at reducing communication barriers in the deaf community. This system is based on a CNN architecture adapted to recognize the five basic signs of ASL. Furthermore, it supports a variety of input formats, including video sequences and real-time camera data, making the system more flexible in use. Despite these advantages, the proposed approach has primarily been evaluated under controlled conditions. Its performance on larger and more diverse datasets, as well as its robustness in real-world settings, has not yet been fully explored. This limitation highlights the need for further practical validation and scalability analysis.

A study of Afaan Oromo Sign Language (AOSL) achieved 92.98% accuracy using a pipeline with pre-processing, Gabor features, and a ResNet-50 classifier [30]. However, most studies still focus on isolated gestures, and generalization across datasets remains limited. This highlights the need for improved temporal modeling and adaptive representations for practical use.

Despite the strong benchmark performance reported across many CNN- and 3D CNN-based SLR systems, these architectures exhibit several structural limitations that affect their applicability in real-world environments. Most convolutional approaches primarily rely on appearance-level spatial representations, making them sensitive to illumination variation, background clutter, camera viewpoint changes, and signer-specific visual characteristics. Furthermore, although 3D CNNs partially capture short-term temporal dynamics, their ability to model long-range contextual dependencies remains limited due to the localized nature of convolutional operations. As a result, many models demonstrate high performance under controlled laboratory conditions but experience noticeable degradation in signer-independent and cross-dataset evaluation settings. These limitations explain the subsequent shift in SLR research toward recurrent, attention-based, and Transformer-driven

architectures designed to improve temporal reasoning, contextual understanding, and generalization across diverse signing conditions.

Across the reviewed studies, CNN and 3D-CNN architectures demonstrate strong performance on controlled, single-dataset benchmarks, with accuracy figures frequently exceeding 90%. However, a consistent pattern emerges: performance degrades substantially when models are evaluated on unseen signers or cross-dataset scenarios. The primary reason is that convolutional features remain sensitive to appearance-level variations—background clutter, illumination changes, and signer-specific morphology—which are not adequately addressed by spatial convolution alone. Furthermore, most 3D-CNN models reviewed here were trained and evaluated on datasets of limited scale and homogeneity, which inflates reported accuracy and masks generalization weaknesses. These observations suggest that spatial feature extraction approaches require complementary temporal modeling and domain adaptation strategies to achieve robust real-world performance, particularly for low-resource sign languages where dataset diversity is inherently limited.

3.2. Temporal Sequence Modeling

Architectures with RNNs and LSTMs are commonly used in SLR systems to model time-dependent behavior. This preserves the ordinal context of movement along with video signals and provides a comprehensive explanation of the dynamics of the gesture process. For example, in experiments conducted with the IISL2020 database containing 11 characters of Indian Sign Language (ISL), a 97% accuracy was achieved by combining LSTM and GRU layers [31]. However, since these results are obtained with relatively simple and repetitive phrases, the effectiveness of this model has not been fully demonstrated in complex or continuous speech sequences. Similarly, a multimodal system developed for Mexican Sign Language (MSL) used LSTM and GRU layers to analyze temporal dynamics. This system combines hand, body, and facial functions with an OpenCV AI Kit with Depth (OAK-D) camera to measure depth (see Figure 5) [32]. However, such architectures still require further research into their potential for adaptation to other language groups or natural communication conditions.

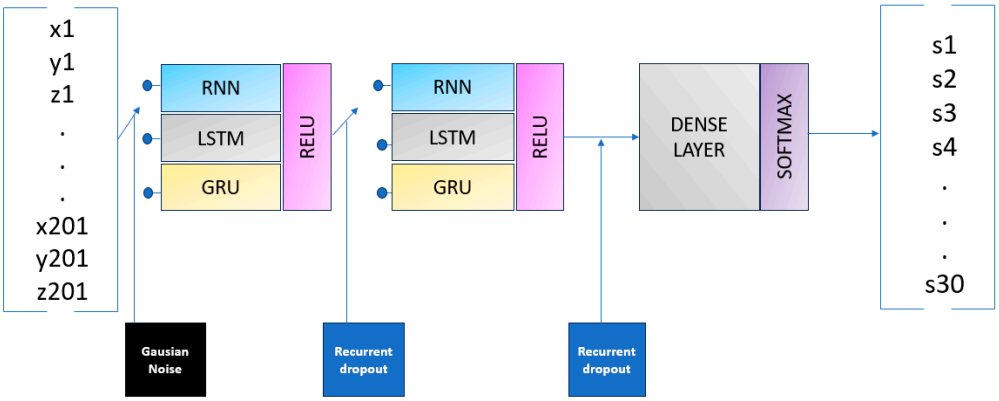


Figure 5. Comparative representation of RNN, LSTM, and GRU models.

The PoseNet-LSTM architecture achieved 98% accuracy in real-time text translation [33], but by integrating OpenCV, MediaPipe, and LSTM tools to recognize short phrases in Indian Sign Language (ISL), we achieved 99.17% accuracy in training and 97.78% accuracy in testing [34]. However, the fact that these models can perform poorly when motion overlaps and is constrained by short circuits is one of the issues that needs to be addressed. In summary, these works demonstrate the effectiveness of recurrent architectures for temporal modeling, while also exposing limitations in robustness, scalability, and generalization to complex real-world gestures.

In the context of Arabic Sign Language (ArSL) [35], keypoint representations are used, suggesting that pre-processing is crucial for model accuracy and achieving an accuracy of over 88.5%. In a recent study [36], a MediaPipe LSTM-based ASL recognition system was proposed, which accurately recognized three words, indicating reliable performance for very constrained tasks, but insufficient evidence for continuous or large-vocabulary SLR.

Recurrent architectures such as LSTM and GRU have demonstrated strong sequential modeling capabilities, achieving high accuracy on constrained tasks with short, well-defined phrase sequences. However, a fundamental limitation shared across these models is their tendency to compress entire gesture sequences into fixed-size hidden states, which inevitably leads to information loss for longer or more complex signing sequences. This compression bottleneck becomes particularly problematic in continuous SLR, where individual signs overlap and temporal boundaries are ambiguous. Additionally, the high accuracy figures reported in several studies, including 99.17% on ISL, are typically obtained under controlled laboratory conditions with limited vocabulary and a small number of signers, which does not reflect the complexity of real-world signing environments.

From a broader architectural perspective, recurrent sequence modeling represented a major advancement over purely spatial CNN-based approaches because it enabled explicit temporal reasoning across sign sequences. However, the reviewed studies consistently indicate that recurrent architectures encounter scalability limitations as sequence complexity increases. In tasks involving continuous SLR that are characterized by long signing streams, overlapping gestures, rapid motion transitions, and signer variability, recurrent hidden-state compression frequently results in information loss and unstable temporal alignment. Furthermore, the sequential processing nature of these models imposes limitations on parallelization efficiency and increases inference latency, consequently reducing their suitability for real-time deployment scenarios. These structural constraints were a significant motivation for the transition to attention-driven and Transformer-based architectures. These architectures provide direct modeling of long-range temporal dependencies without reliance on compressed sequential memory representations.

3.3. Transformer-Based and Attention-Driven SLR

Modern research is actively developing new approaches that allow systems not only to recognize individual gestures, but also to translate entire sentences in real time. The authors of [37] proposed a novel Transformer architecture jointly trained to handle both Continuous Sign Language Recognition (CSLR) and sign language translation (SLT). The system employed the Connectionist Temporal Classification (CTC) loss function, which connects recognition and translation tasks without requiring explicit temporal alignment. This design significantly improved performance on the challenging RWTH-PHOENIX-Weather-2014T dataset. The suggested networks outperformed both video-to-speech translation models and gloss-to-speech translation models, in some cases boosting translation quality by more than half, achieving Bilingual Evaluation Understudy with 4-gram (BLEU-4) scores that increased from 9.58 to 21.80.

The Transformer encoder and decoder utilize a scaled dot product attention mechanism, which is formulated as follows:

$$\text{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \text{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \quad (1)$$

The formulas presented here describe the scaled dot product attention mechanism, which is a core component of both encoding and decoding in the Transformer architecture, where Q and K denote the query, key, and value representations, respectively, and d_k represents the key dimensionality scaling factor.

To further enhance representation capacity, multi-head attention is applied, defined as

$$\mathrm{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_\mathrm{h})\mathrm{W}^\mathrm{O} \tag{2}$$

where each head is computed as

$$\mathrm{head}_i = \mathrm{Attention}(\mathrm{QW}_i^Q, \mathrm{KW}_i^K, \mathrm{VW}_i^V)$$

In this formulation, h denotes the number of parallel attention heads, while W_i^Q, W_i^K, and W_i^V represent the learnable projection matrices for the query, key, and value representations, respectively.

By computing h independent attention functions in parallel, the multi-head mechanism enables the model to jointly attend to information from different representation subspaces at different temporal positions. In the context of video-based SLR, this property is particularly valuable: individual attention heads can specialize in capturing distinct aspects of gesture articulation, such as handshape trajectories, upper-body posture transitions, and facial expression dynamics. Since sign language meaning is conveyed simultaneously across multiple articulators, this parallel multi-channel attention provides a more complete and semantically grounded representation of the sign sequence than single-head attention alone.

$$\mathrm{DecoderOutput} = \mathrm{softmax}\left(\frac{\mathrm{Q_{dec}K_{enc}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V_{enc}} \tag{3}$$

where Q_{dec} denotes the decoder query representations, while K_{enc} and V_{enc} correspond to the key and value representations obtained from the encoded video features. The scaling factor $\sqrt{d_k}$ stabilizes gradient magnitudes during training by preventing the dot products from growing excessively large in high-dimensional spaces.

This cross-attention mechanism enables the decoder to dynamically align predicted glosses or words with the most informative temporal regions of the input sign sequence. Such alignment has been demonstrated to enhance contextual translation consistency and strengthen the modeling of long-range dependencies between gesture phases and linguistic units in continuous SLR tasks. In practice, cross-attention significantly improves sequence-level recognition and translation quality. However, the global attention operation also increases computational cost and memory requirements, which remain important limitations for real-time and low-resource deployment scenarios.

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, X) \tag{4}$$

where T denotes the number of output tokens in the target sequence, y_t represents the target token at position t, and X corresponds to the encoded video feature sequence.

Minimizing $\mathcal{L}$ during training encourages the model to assign progressively higher probability mass to the correct output tokens at each decoding step, effectively penalizing confident incorrect predictions more heavily than uncertain ones. This loss function trains both the recognition and translation components jointly in an end-to-end manner, creating a unified optimization objective that connects visual feature extraction, temporal modeling, and sequence generation. Lower values of $\mathcal{L}$ correspond to closer alignment between the predicted probability distributions and the reference translation sequences, which empirically correlates with improved BLEU scores and reduced Word Error Rate (WER) on standard SLR benchmarks such as RWTH-PHOENIX-Weather-2014T [38]. For more accurate and reliable recognition of sign language [39], studies take into account not only hand gestures, but also body movements and facial expressions.

In [40], a multi-branch architecture was proposed to address the computational complexity and limitations of Transformer-based models for Korean SL. The model integrates CNN and Transformer features, lightweight Multi-Head Self-Attention (MHSA), and a Grain module with 3×3 convolutions, achieving 89.0% accuracy on a 77-label dataset and 98.3% on a laboratory dataset while reducing computational cost.

Within this multi-branch Transformer-based architecture, a scaled dot product attention mechanism is applied, defined as

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where Q, K and V denote the query, key, and value representations, respectively, and d_k represents the scaling factor associated with the key dimensionality. This approach estimates the pairwise similarity of all temporal features, which are then converted into attention weights that can be normalized by the softmax function, allowing the model to prioritize the most important parts of the feature sequence. In practice, this allows Transformers to capture long-term dependencies between gesture components, including hand gestures, body posture changes, and contextual coarticulation within cues that span sequences. Unlike other iterative architectures, the attention-based model directly links to distant frames, making CSLR very effective.

$$P(y|x) = soft\max(W\mathbf{h}_{cls} + b) \tag{6}$$

where $\mathbf{h}_{cls}$ denotes the global sequence representation produced by the Transformer encoder, while W and b correspond to the learnable classification parameters.

The softmax function converts the linear classification scores into normalized probability values across all label classes, allowing the model to assign a confidence score to each candidate label class, enabling probabilistic decisions during inference.

From a representation learning perspective, the Transformer encoder learns context-sensitive embeddings that capture the temporal gesture evolution across sign sequences, and the classification layer maps these embeddings to discrete sign label probabilities.

Although Transformer-based architectures have demonstrated high performance in SLR tasks, several challenges remain: most Transformer-based SLR models rely on large labeled datasets, limiting their applicability to resource-scarce SLs.

In SLR, Transformer methods rely on visual cues and skeletal coordinate sequences rather than accurately modeling the geometric relationships in the joints of the hand, which resemble dynamic variations in joint angles. Transformers are traditionally weighted like all temporal cues, which reduces the sensitivity to subtle local motion that is essential for visual cue separation. Therefore, the inclusion of geometric features and adaptive temporal attention in the Transformer circuit model provides a promising direction for robustness and generalization in SLR systems.

Although Transformer-based architectures enable integrated end-to-end training of SLR and translation, several limitations remain. First, Transformer models require large annotated datasets, which are difficult to obtain for many SLs. Second, the models rely primarily on visual token embeddings extracted from raw video features, which may still contain redundant background information. A third obstacle is that the trans-architecture directly ignores the geometric structure of the arms (e.g., the corners of the frame). Furthermore, since the mechanism of transient attention is the same for all tokens, there is a risk of incomplete activation of the subtle dynamics of finger movements. For this reason, combining geometric properties with a flexible focusing mechanism to enhance the reliability and adaptability of the system remains an unresolved problem.

A cross-attention (CA) module was proposed in [41] to process multimodal data such as RGB and optical flow in a single system for the purpose of analyzing complex gestures. Although this proposed approach reduces computational costs and improves recognition quality, it does not solve the problem of maintaining a balance between model complexity and speed. While the Contrastive Visual–Textual SLR (CVT-SLR) system described in [42] outperforms both single-modal and multimodal methods, its reliance on pre-trained models may limit its flexibility in new databases or new domains.

To facilitate CSLR, the Inter-Gloss Attention (IIGA) module was used in [43]. Although this method performs well in terms of WER, it requires careful pre-processing steps such as background removal to work effectively, which means that the system remains sensitive to the interference from the real environment.

Also, in [44], the authors proposed a Relative Spatial Transformer (RST) architecture that encodes spatial relative relationships instead of absolute coordinates. While this solution improves the model's ability to distinguish spatial relationships, its scalability to larger vocabularies or specific use cases remains an open question. In summary, although Transformer-based SLR models offer high accuracy and the ability to analyze complex sequences, their high computational requirements, dependence on database quality, and inability to fully handle subtle geometric features such as finger gestures remain major challenges for researchers.

Transformer-based architectures signify a substantial conceptual shift in the domain of SLR research, wherein they supersede conventional compressed sequential memory mechanisms, instead favoring direct global attention modeling across the entire sign sequences. This enables improved contextual reasoning, stronger long-range temporal dependency modeling, and more effective integration of multimodal information compared to recurrent architectures. The reviewed studies demonstrate that Transformers consistently achieve superior benchmark performance in continuous SLR and sign language translation tasks, particularly on large-scale datasets with complex temporal structure.

Nevertheless, there are several significant limitations that have yet to be satisfactorily addressed. Firstly, Transformer models generally necessitate substantial annotated datasets and considerable computational resources, which limits their applicability to low-resource sign languages and edge-device deployment scenarios. Secondly, it has been demonstrated that attention mechanisms in themselves do not inherently encode geometric relationships between finger joints and body articulators. This has the effect of limiting sensitivity to subtle fine-grained motion patterns that are linguistically important in sign communication. Thirdly, the elevated parameter complexity of numerous Transformer-based systems has been demonstrated to increase the cost of training and to reduce the efficiency of inference in real-time applications. The aforementioned limitations have led to a surge of interest in graph-based representations, lightweight attention mechanisms, and hybrid multimodal architectures. These architectures are designed to enhance geometric awareness, computational efficiency, and robustness under real-world signing conditions.

3.4. Graph-Based Spatial–Temporal Modeling for SLR

GCN can effectively model the spatial and temporal patterns in a sign language sequence by analyzing the coordinates in the skeletal joints. In the pose-based word-level sign language recognition (WSLR) framework, spatial and temporal features are separated in [45], spatial dependence is determined by GCN, and temporal relationships are modeled using the Bidirectional Transformer Encoder (BERT). While this approach improves prediction accuracy by approximately 5% on the WLASL dataset compared to conventional pose-based methods, its generalization to diverse datasets and complex real-world conditions remains underexplored.

Complex video backgrounds and the challenges posed by multiple lighting scenarios were addressed with structures based on multiple frames. The multi-stream Sign Language Image Deformation Network (SL-IDN) is used for study by researchers [46], with 27 main nodes across four feature streams, and offering strong generalization across WLASL, AUTSL, and CSL datasets, yet performance on large-scale, unconstrained videos has not been fully validated. With a low computational resource requirement during comparison, multi-branch attention models that can incorporate spatial and temporal graph features for dynamic hand movements with deep neural representations have achieved accuracies of 94.12%, 92.00%, and 97.01% on the MSRA, DHG, and SHREC'17 datasets, respectively [47]. However, these high accuracies often rely on preprocessed or idealized skeleton data, which may limit robustness in noisy or occluded environments.

The lightweight pipelines incorporated into the residual GCN containing constrained configurations effectively model the spatial and temporal dependencies, reduce computational cost, and maintain high performance, which yielded an accuracy rate of 27.62% in WLASL-300, 26.97% in WLASL-1000, and 100% in LSA-64 [48]. Despite these results, performance degradation is evident on large-scale and complex datasets, indicating that scalability is a major limitation.

To better capture the local topology of the arm and wrist, HAND-aware graph convolutional networks (HA-GCNs) outperformed previously presented models in AUTSL by using subgraphs of all joints and bones, providing a drop-graph function that included the dataset [49]. Graph convolution focusing networks (GCARs) improved their performance by integrating multi-stage graph scrolling and focusing mechanisms [50]. Similarly, models including a spatial–temporal graph convolutional network (ST-GCN), BiLSTM, and 3D pose parameterization showed competitive results on the RWTH-PHOENIX-2014 and Chinese datasets (Split I) [51].

However, these approaches remain computationally intensive, often requiring careful hyperparameter tuning and limiting their real-time use. Lightweight models based on handcrafted frameworks and residual GCNs showed superior results on the WLASL and MINDS-Libras datasets, while reducing computational cost [52]. This implies a trade-off between model complexity, accuracy, and computational efficiency, which must be carefully balanced depending on the application context. GCN-based recognition is advancing automated translation systems by shifting the focus from raw pixel analysis to structured pose and motion analysis. However, performance heavily relies on idealized skeletons, and combining multiple models increases complexity and the risk of over-classification. These limitations require the combination of graphical representations and additional modules to enhance robustness, capture spatiotemporal dependencies, and generalize to invisible signers.

GCN-based approaches offer a fundamentally different representational paradigm compared to pixel-based methods: by explicitly modeling the skeletal graph structure of the human body, they achieve greater robustness to background noise and illumination variation. The reviewed studies confirm that GCNs consistently outperform pose-based baselines on datasets such as WLASL and AUTSL when skeleton data quality is high. However, this strength becomes a vulnerability in real-world conditions: GCN performance degrades significantly when pose estimation fails due to the occlusion, fast motion, or low image resolution, because erroneous skeletal inputs propagate directly into graph convolution. Furthermore, most reviewed GCN models rely on idealized, preprocessed skeleton data that may not be available in deployment settings. The computational cost of multi-stream GCN architectures also limits their applicability in real-time and mobile scenarios. These trade-offs suggest that GCNs are most effective as a complementary component within hybrid architectures, rather than as standalone solutions.

From a system-level perspective, GCN-based approaches address one of the central weaknesses of appearance-driven SLR models by explicitly modeling the structural relationships between body joints, hand articulators, and motion trajectories. This representation significantly improves robustness to background variability, illumination changes, and certain forms of visual noise that commonly affect RGB-based methods. As a result, graph-based representations have become increasingly important for signer-independent recognition and multimodal fusion pipelines.

Nevertheless, the reviewed studies also reveal several persistent challenges. The effectiveness of GCN architectures remains highly dependent on the quality and stability of pose estimation, meaning that errors caused by occlusion, motion blur, low-resolution video, or inaccurate skeletal extraction directly propagate into graph representations and reduce recognition reliability. In addition, many multi-stream GCN systems introduce considerable computational overhead due to the simultaneous processing of spatial, temporal, and multimodal graph structures. These constraints limit scalability and complicate deployment in real-time and mobile-assistive systems. Consequently, recent research trends increasingly explore lightweight graph architectures, adaptive graph attention mechanisms, and hybrid Transformer–GCN frameworks that attempt to balance structural awareness with computational efficiency and temporal modeling capacity.

3.5. Hybrid End-to-End SLR and Translation Models

Sign language interpretation models tend to be computationally expensive and have an insufficient understanding of the global context. The Transformer–CNN hybrid approach combines the local functions of a CNN with the global representation of a Vision Transformer (ViT) to effectively capture subtle gestures and achieve 99.97% accuracy, 110 Frames Per Second (FPS), and just 5.0 Gigabytes of Giga Floating-Point Operations Per Second (GFLOP) on Alphabet's ASL dataset [53]. However, these results are often obtained on controlled datasets, and the model's robustness under real-world variations, such as occlusion or lighting changes, remains unclear.

Kondo et al. [54] have recently presented their work on the efficiency of capturing angular features present in Japanese SLR cameras and have compared the results of CNN and ViT-based architectures, as shown in Figure 6.

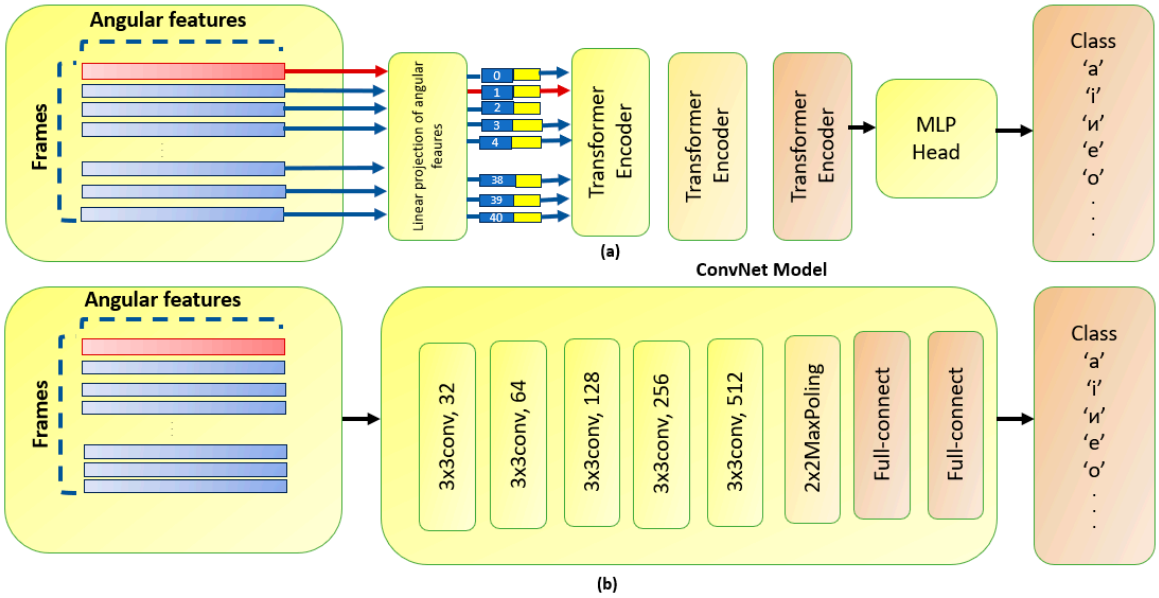


Figure 6. The architectures of the ViT and CNN models are: (**a**) the ViT model consisting of three encoder stages and (**b**) the CNN model consisting of 8 layers deep.

Let $\mathbf{p}_i$, $\mathbf{p}_j$, and $\mathbf{p}_k \in R^3$ denote the 3D coordinates of finger joints extracted using the MediaPipe pose estimation framework. The angular feature θ represents the angle formed at the joint $\mathbf{p}_j$ by the vectors connecting adjacent joints:

$$\theta = \arccos\left(\frac{(\mathbf{p}_i - \mathbf{p}_j) \times (\mathbf{p}_k - \mathbf{p}_j)}{\| \mathbf{p}_i - \mathbf{p}_j \| \| \mathbf{p}_i - \mathbf{p}_j \|}\right) \tag{7}$$

where $\mathbf{p}_i$, $\mathbf{p}_j$ and $\mathbf{p}_k$ denote the 3D coordinates of adjacent finger joints, and θ represents the angular relationship between the corresponding skeletal vectors. The vector of values used as a divisor is used to smooth the values and to ensure that they are stable over the current measurement and the distance between the shots. Here, the function is a normalized dot product of the function, which converts the angle parameter measured in radians into the format of the view. These angle parameters provide a compact and reliable description of finger movements, and are less sensitive to changes in translation and illumination than when using raw coordinate symbols. The resulting angle feature vectors are stored in a temporary location and are introduced as input tokens through the ViT encoder feature in the SLR pipeline.

Despite the achievements described above, several limitations remain unresolved across current SLR approaches. RGB-based methods remain susceptible to background noise and illumination changes, while skeleton-based methods partially mitigate these issues but rely on raw Cartesian joint coordinates that do not capture geometric relationships between finger joints. The angular feature representations offer a more compact and robust alternative, as they are invariant to translation and scale changes, making them better suited for real-world deployment. Using a skeleton, these methods partially eliminate the limitations of using joint coordinates in visualization, which are often characterized by not taking into account the geometric relationships of the finger joints, and use raw (unprocessed) Cartesian coordinates.

The angular difference representations given by the coordinates of the joints in the skeleton represent a compact form of the movement of the joints of the hand, and the model, in turn, produces semantically meaningful descriptions. These characteristics are not inherent in the transmission, zoom adaptation, and changes in lighting conditions, factors that make them suitable for use in real-world SLR systems.

Several research questions remain open in this area. First, the integration of angular skeletal representations with Transformer-based architectures for contextual modeling remains limited. Second, the deep temporal modeling of angular motion dynamics in sign sequences has not been fully addressed. Third, systematic comparative studies evaluating angular feature implementations across multiple deep learning architectures under controlled experimental conditions are lacking. These gaps represent concrete directions for future investigation. Second, the problem of deep temporal modeling of sign language sequences in angular motion dynamics, which is focused on frame-level classification performance, has not yet been fully addressed in the current study. Third, comparative studies that can assess how similar angular feature implementations are used in many deep learning architectures under controlled experimental conditions remain lacking.

Research on hybrid architectures involving convolutional spatial feature extraction and global contextual modeling in the Transformer when implemented on geometric skeleton representations is limited. And the generalization efficiency of SLR systems that rely on angular features in real-world writing situations, such as signers and datasets, is still an open question.

Subsequent research should systematically investigate Transformer-based and hybrid deep learning architectures that incorporate angular skeletal features, with a focus on tem-

poral sequence modeling, robustness to real-world variation, and computational efficiency for practical SLR deployment.

Lip-reading-based Audio–Visual Speech Recognition (AVSR) models using an Active Shape Model (ASM) and CNN reduced WER by 6.59% and reached 95% lip-reading accuracy [55]. A hybrid CNN–BiLSTM–GAN pipeline using MediaPipe enabled recognition, translation, and gesture generation, achieving a classification accuracy above 95% and a BLEU of 38.06 [56]. TwoStream-SLR/SLT models with dual visual encoders achieved state-of-the-art SLR/SLT performance [57]. Nevertheless, many hybrid models still use static modality fusion or simple temporal aggregation, ignoring adaptive weighting of manual and non-manual features, which limits their robustness in natural sign language communication.

A hybrid deep learning system combining CNN-based spatial features, manual keypoint priors, LSTM temporal modeling, and Transformer-based long-range dependencies was proposed for Kannada SL, which achieved 97.6% training, 96.75% validation, and 81% testing accuracy on a new medical domain dataset and outperformed traditional CNN-LSTM and manual keypoint (HKP)-LSTM baselines [58]. A multi-scale context-aware network (MSCA-Net) with multi-scale motion and temporal modules achieved the highest CSLR results [59]. Specialized models improved ASL question recognition (98.91% accuracy) [60], boosted efficiency with spiking neural networks [61], or enabled fast Sistem Isyarat Bahasa Indonesia (SIBI)-to-Indonesian translation (96–99% accuracy) [62].

Region-specific systems included Uighur SLR using improved You Only Look Once version 7 (YOLOv7) [63], Ethiopian Sign Language (EthSL) with 8.82% WER [64], and Azerbaijani Sign Language (AzSL), achieving 94% accuracy [65]. Lightweight detection models like YOLOv8 [66], and Single-Shot MultiBox Detector (SSD) with MobileNet V2 [67], and Dual-branch Shuffle Attention Mechanism–You Only Look Once version X (DSA-YOLOX) [68] improved real-time CSLR with reduced cost. Advanced fusion methods such as Multi-Scaled feature Fusion–Euclidean Transformer (MSF-ET) [69] and Dual-Stream Full-Hand networks [70] set new benchmarks on CSL and Ankara University Turkish Sign Language (AUTSL) datasets.

Additional studies explored a Long-term Recurrent Convolutional Network (LRCN) for ASL [71] and introduced a multimodal ArSL dataset with MobileNet-LSTM, achieving 99.7% dependent-mode accuracy [72], as shown in Figure 7.

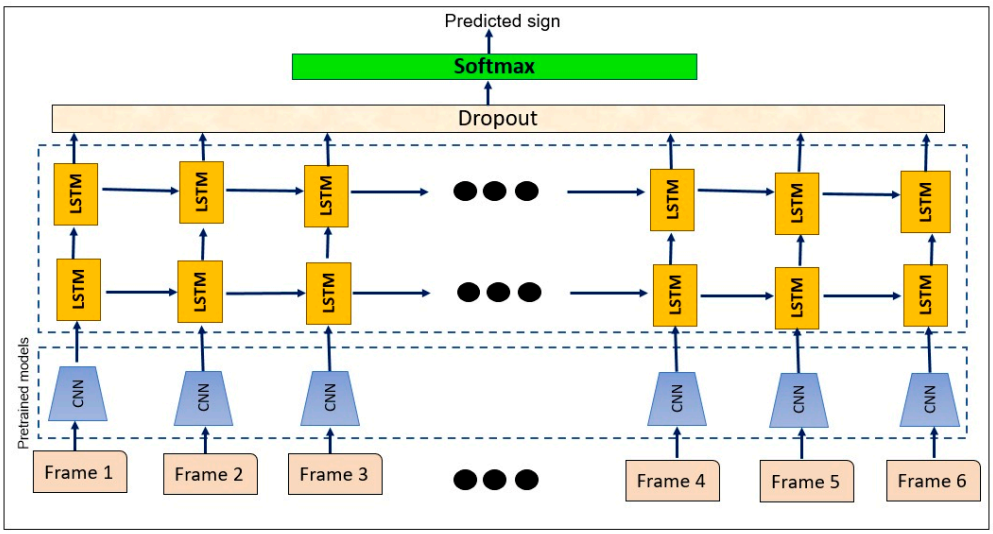


Figure 7. System structure with two CNN variants.

To formally represent the multimodal feature integration, the fused feature vector is defined as

$$\mathbf{F}_{fusion} = [\mathbf{F}_{manual} || \mathbf{F}_{nonmanual}] \quad (8)$$

where $\mathbf{F}_{manual}$ and $\mathbf{F}_{nonmanual}$ denote the feature representations extracted from manual and non-manual signing components, respectively. Vector concatenation denoted by the operator $||$ produces a unified multimodal feature vector, integrating manual and non-manual information. If the dimensionality of manual features is d_m and non-manual features is d_n, the fused vector resides in $R^{d_m+d_n}$. This finding opens the way for a recognition model to learn about cross-modality dependencies between hand gestures and facial features, which is important for sign language in this semantic sense, which is transmitted over multiple channels. However, it is not possible to predict that these two modalities can contribute similarly to the causes of all-encompassing sign language gestures.

$$P(y|x) = soft\max(W\mathbf{F}_{fusion} + b) \quad (9)$$

where $\mathbf{F}_{fusion}$ denotes the fused multimodal representation, while W and b correspond to the learnable classification parameters. The fused feature vector provides a joint representation of manual and non-manual linguistic information. Following the linear classifier, the softmax function transforms its output into a normalized probability for all feature classes, which generates a visual representation of the features that represent facial features or head movements that is sensitive to the semantic differences in the model. During training, this output layer is typically optimized using a cross-entropy loss, allowing the model to learn discriminative multimodal feature representations for SLR.

Hybrid multimodal architectures currently represent the dominant research direction in advanced SLR because they combine the complementary strengths of multiple representational paradigms, including CNN-based spatial extraction, Transformer-based contextual modeling, graph-based structural reasoning, and multimodal fusion of manual and non-manual linguistic cues. The reviewed studies consistently demonstrate that such systems achieve the strongest benchmark performance, particularly in continuous SLR and sign language translation tasks involving complex temporal semantics.

However, the increasing architectural complexity of hybrid systems gives rise to a number of significant practical challenges. Firstly, the integration of multiple feature streams substantially increases computational cost, memory consumption, and training requirements, which limits deployment feasibility on mobile and embedded assistive platforms. Secondly, many multimodal fusion strategies rely on static weighting mechanisms that assume equal importance of all modalities. However, the contribution of facial expressions, hand motion, pose dynamics, and contextual information varies significantly across sign languages and communication scenarios. Thirdly, it is evident that a significant proportion of reported benchmark results are obtained under controlled experimental conditions using relatively homogeneous datasets. This may result in an overestimation of real-world generalization capability. These observations suggest that future progress in SLR will depend not only on increasing architectural complexity but also on developing adaptive fusion mechanisms, lightweight deployment-oriented models, and more ecologically valid evaluation protocols capable of reflecting realistic signing environments.

3.6. Transfer of Training and Other Methods

Transfer learning, feature enrichment, and auxiliary strategies are used to improve the accuracy of SLR. Enriching WSLR models with subtitle-aligned features and external memory for class centroids was proposed in [73]. While this improves recognition, the reliance on external annotations may limit application to domains without aligned subtitles,

which highlights the reliance on leading datasets. The ChaLearn LAP Large-Scale Signer-Independent Isolated SLR Challenge (CVPR 2021) [74] evaluated multimodal approaches including RGB, depth, pose estimation, transfer learning, ensembles, and spatiotemporal models on the AUTSL dataset (226 gestures, 36K videos). Although the accuracy exceeded 96%, similar gestures, authenticity, and interpretation issues remained, highlighting the need for balanced datasets and contextual attention.

The research in [75] introduced Sign Hidden-Unit BERT (SHuBERT) (Figure 8), a self-supervised encoder trained on 1000 h of ASL videos, extending SHuBERT to multi-stream visual input (hands, face, body). Using MediaPipe keypoints, DINOv2 features, and K-means clustering, SHuBERT achieved state-of-the-art performance in ASL translation (How2Sign, OpenASL, Few-shot Learning Evaluation for Universal Representation of Sign Languages (FLEURS-ASL) and isolated recognition (ASL Citizen, SEM-LEX)).

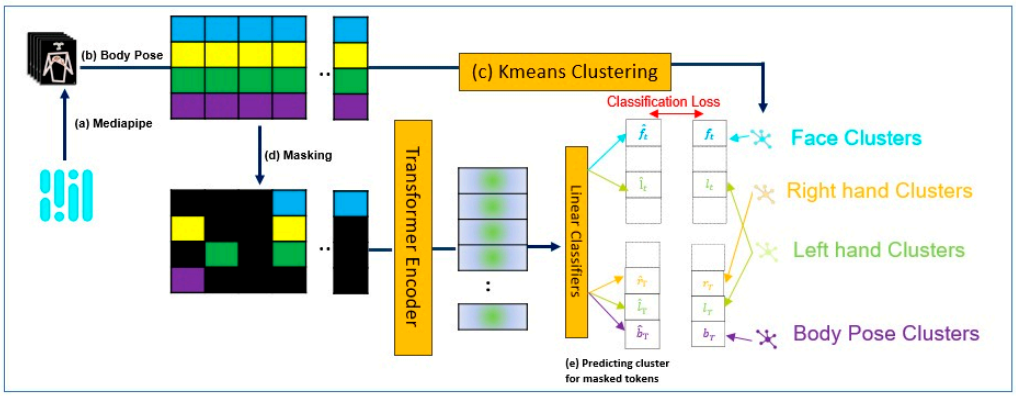


Figure 8. Pre-training of Sign Hidden-Unit BERT.

To formally define the self-supervised multi-stream cluster prediction objective, the loss function can be expressed as

$$\mathcal{L}_{cluster} = -\sum_{s=1}^{S}\sum_{i=1}^{N} q_i^{(s)} \log p_i^{(s)} \tag{10}$$

where S denotes the number of modality streams, $q_i^{(s)}$ represents the pseudo-label assignment probability for sample i in stream s, and $p_i^{(s)}$ corresponds to the predicted cluster distribution produced by the model.

This self-supervised objective enables the learning of shared latent representations across multiple feature streams without requiring extensive manual annotation. In the context of video-based SLR, such learning strategies are particularly valuable for low-resource sign languages where large-scale labeled datasets are difficult to obtain. However, cluster-based self-supervised objectives primarily capture global semantic structure and may insufficiently model fine-grained temporal gesture dynamics and subtle finger articulation patterns required for continuous SLR tasks.

$$\mathcal{L}_{consistency} = \sum_{s=1}^{S}\sum_{t=1}^{S} ||\mathbf{z}^{(s)} - \mathbf{z}^{(t)}||_2^2 \tag{11}$$

where $\mathbf{z}^{(s)}$ and $\mathbf{z}^{(t)}$ denote latent feature embeddings extracted from different modality streams.

The objective minimizes the Euclidean distance between embeddings from different streams corresponding to the same video segment. This encourages the model to generate gesture representations that are independent of a particular modality, reducing its depen-

dence on a single source of information. Maintaining consistency between different data streams ensures consistent SLR across different writing methods, changing illumination and partial occlusion. By storing additional information appropriate for the model's modality, it creates alignment through hidden embeddings across multiple threads to provide the same semantic identity for all. Although the dependence on the sign language datasets identified in the self-directed learning model has been reduced, there are limitations to the methods for clustering across multiple streams. Specifically, they cannot adequately capture the temporal dynamics of sign language movements during task performance, which involves static clustering, which is more commonly used in these methods. Furthermore, cluster-based self-supervised learning typically focuses on the semantics of the entire video, which may not be able to adequately model fine-grained finger movements or short-term movement patterns.

Additionally, cross-stream consistency constraints often assume that the importance of all modality streams is uniform, which may not reflect real-world sign language scenarios where certain modalities dominate depending on the gesture type. Therefore, an adaptive multi-stream self-supervised learning framework incorporating temporal attention mechanisms and fine-grained geometric feature modeling remains an open research topic.

Research in [76] proposed an integrated Isolated Sign Language Recognition (ISLR) pipeline that combines target data augmentation, regression heads, and an Intersection-Over-Union (IoU)-balanced classification loss. To improve generalizability and explainability, research in [77] developed an American Sign Language Knowledge Graph (ASLKG) from 12 expert sources. A neuro-symbolic model trained on the ASLKG achieved an ISR accuracy of 91%, an unseen semantic feature prediction accuracy of 14%, and a YouTube–ASL topic classification accuracy of 36%.

Subspace-based and multimodal fusion methods [78,79] improved training accuracy, but non-visual generalization and interlingual translation accuracy reveal persistent issues with domain shift, class confusion, and dependence on high-quality multimodal inputs. Although transfer learning and self-referential strategies offer advantages, their effectiveness depends on the quality of the learning data, which requires careful validation and dataset management that includes accurate and contextually relevant key assumptions.

Transfer learning and self-supervised approaches represent the most promising direction for addressing data scarcity in low-resource SLR. Models such as SHuBERT demonstrate that pre-training on large unlabeled sign language video corpora can achieve competitive performance with significantly less labeled data. However, the effectiveness of transfer learning depends critically on the domain proximity between the source and target sign languages: features learned from ASL may not transfer effectively to structurally distinct languages such as KSL, which differs in handshape inventory, spatial grammar, and non-manual marker usage. Furthermore, cluster-based self-supervised objectives typically optimize for global video-level semantics, which may inadequately capture the fine-grained temporal dynamics of individual phonological features. These observations highlight the need for language-specific pre-training corpora and temporally aware self-supervised objectives as priorities for future research in low-resource SLR.

4. Overview and Contribution to Datasets

The WLASL dataset has been expanded to include over 2000 words performed by more than 100 sign artists. As one of the largest publicly available resources for word-level gesture recognition in ASL, it enables experimentation with diverse deep learning approaches, including holistic visual perception and two-dimensional human pose estimation. However, despite its scale, the dataset may still be limited in participant diversity and environmental conditions, which can impact model generalization. The proposed Pose-

based Time Graph Convolution Networks (Pose-TGCNs) have been shown to improve recognition performance [80].

A scalable method for collecting data for gesture recognition in continuous videos is proposed. Using substandard broadcast subtitles in conjunction with a keyword detection methodology, gesture exemplars were automatically localized, resulting in a lexicon comprising 1000 characters from 1000 h of video footage. The British Sign Language 1000-class (BSL-1K) dataset, a substantial collection of BSL data, is accessible to the public and has been demonstrated to serve as a robust foundation for pre-learning models of other SLs, the Multilingual Sign Language Dataset (MSASL), and WLASL [81]. Due to its public accessibility and structured organization, this dataset is valuable for replicating experiments and conducting comparative analyses, but it may still lack detailed spatiotemporal annotations.

Two Arabic Sign Language datasets, ArASL2018 and ArASL2021, were presented in the study in [82], which cover gestures in simple and complex situations, respectively. Both datasets are publicly available and contribute to research in the field of assistive technologies. Nevertheless, the relatively small size compared to ASL datasets may limit deep learning model scalability.

The largest database of Saudi SL (SSL) has been created, comprising 293 signs, 33 SLs, and 145,035 samples across 10 different thematic areas [83]. However, the paper does not mention dataset availability, which significantly limits reproducibility and the broader scientific impact.

A large-scale dataset of ASL with annotations of six Recognizing Phonological Properties (RPPs) has been partially released [84]. Graph neural networks applied to skeletal features demonstrate promising recognition of these properties, including unseen patterns. However, restricted data access limits external validation and benchmarking.

A large-scale dataset of Turkish Sign Language (TSL) is available to the public. This dataset contains more than 22,000 individual videos, including 744 unique Turkish texts created by six native sign language specialists [85]. Its structured design supports cross-linguistic pre-training, but the limited number of performers may constrain the dataset's variability.

A multimodal corpus has been created to improve the understanding of religious aspects by deaf people using ArSL. The dataset includes audio signals and 262 Arabic Sign Language videos with text written by two experts [86], but it is not publicly available.

The Temporal Lift Pool (TLP) method has been improved in [87] for the aggregation of time series features. Experiments show that it has higher accuracy compared to traditional methods; however, public access to the datasets used is limited, which limits its general applicability. The ASL Citizen dataset [88] is a public and valuable research tool, while the Slovo dataset for Russian Sign Language [89] contains 20,000 Full HD videos covering 1000 classes created by 194 performers. Although this dataset allows for high accuracy in gesture classification, it is necessary to consider possible errors due to the homogeneity of participant selection and recording conditions.

Extending 3D sign language resources, SignAvatars provides a multi-narrative dataset of 70,000 videos and 8.34 million frames [90]. The main advantage of including ASL and HamNoSys gestures with the multimodal structure of the dataset is that they facilitate the interpretation and creation of 3D sign language. However, the high complexity of the data can make it difficult to train standard models and increase storage requirements.

The AUTSL dataset contains 226 gestures performed by 43 participants and contains 38,336 RGB-D images [91]. The diversity of backgrounds and the complexity of poses make it a valuable benchmark dataset, but a significant decrease in accuracy to 62.02% is observed in user-independent estimation, which indicates the difficulty of generalizing the model. Despite the limitations of a widely used dataset such as AUTSL, this issue highlights the

need to develop specialized corpora for specific sign languages with limited resources. In this regard, the creation of our own dataset is an important step to improve the recognition quality and increase the generalization ability of the models. The next section describes the development process of our developed KSL dataset.

Collected Dataset for KSL

The Kazakh Sign Language Video Dataset (KSLVD) [92] was developed to address the critical absence of sentence-level resources for KSL, which remains one of the most underrepresented sign languages in the deep learning literature. The dataset comprises 18,000 augmented video sequences representing 200 linguistically validated sentence-level phrases, recorded by three native KSL signers under naturalistic conditions with intentional variability in background, lighting, and signer position. Three native speakers of KSL, all early bilinguals and graduates of specialized education programs for the deaf, conducted the recordings in accordance with ethical requirements after obtaining written informed consent and granting copyright agreements. The feature extraction was performed using MediaPipe, yielding 225 skeletal keypoints per frame; sequences were normalized to 50 frames. In comparison to the FluentSigners-50 dataset [93], which contains 173 lexical items, KSLVD provides broader vocabulary coverage and greater syntactic diversity at the sentence level (see Table 2).

Table 2. Comparative analysis of selected sign language datasets.

Reference	Dataset	Classes	Videos	Strengths	Limitations
[80]	WLASL	2000+	21,083	Large-scale, diverse signers, ASL	Limited environmental diversity
[91]	AUTSL	226	38,336	RGB-D, multimodal, varied backgrounds	62% signer-independent accuracy
[81]	BSL-1K	1000	1000 h	Scalable co-articulated SL, mouthing cues	Limited spatiotemporal annotations
[85]	Bosphorus Sign22k	744	22,000+	Large Turkish SL, cross-lingual pre-training	Limited number of signers (6)
[88]	ASL Citizen	2731	83,399	Community-sourced, high signer diversity	Isolated signs only
[89]	Slovo	1000	20,000	Full HD, Russian SL, 194 signers	Participant homogeneity
[93]	FluentSigners-50	173	43,250	Signer-independent KSL benchmark	No sentence-level, limited vocabulary
[92]	KSLVD	200	18,000	Sentence-level KSL, linguistic validation	Not publicly available, limited size

The dataset is currently not publicly available; however, it represents a structured linguistic resource for KSL research, and its detailed development methodology is described in a companion publication [92].

5. Evaluation Metrics

A comprehensive assessment of the performance of SLR systems consists of a set of indicators that quantitatively characterize the recognition accuracy, error rate, generalization ability of the model, and its performance in real time. Accuracy is the main indicator that reflects the proportion of correctly classified gestures [94,95]. In addition, to determine

the practical applicability of the system, its generalization ability, tolerance to input data distortion, and the quality of contextual interpretation are additionally taken into account.

SLR methods have shown high accuracy rates in various language contexts. For example, a South Indian SL classification method based on optical flow and an Inflated-3D model achieved an average accuracy of 0.8709 when subsequently translated into Kannada SL in real time [96]. A model for ArSL using two-dimensional (2D) keypoints achieved 98.39% accuracy for dynamic gestures and 88.89% accuracy for static gestures [97]. Using a CNN with the CBAM module for Malaysian Sign Language has achieved an accuracy rate exceeding 90% [98], while for Egyptian Sign Language, CNN and CNN-LSTM models achieved 90% and 72% accuracy, respectively [99]. To develop comparative analysis, new datasets for Indian Sign Language, INSIGNVID for ISL [100], as well as specialized corpora for Hong Kong SL [101], KSL [102], Bengali SL [103], and Bangla SL [104], have been proposed. These resources have contributed to the development of robust models. However, differences in dataset size, annotation quality, and writing conditions lead to heterogeneity, which can reduce the robustness of comparative analysis.

Modern architectures have shown significant improvements in recognition performance. In particular, 3D CNN-based models have achieved 97.5%, 99.75%, and 98% accuracy for Tamil, ISL, and ASL, respectively [105]. The CNN-LSTM architecture achieved 97% accuracy in a study [106], while the SKResNet-TCN model achieved 100% accuracy on the LSA64 dataset [107]. ViT and Transformer achieved 99.53% accuracy [108]. An ArSL recognition system with a CNN-LSTM–Self-Attention Multilayer Perceptron (CNN-LSTM-SelfMLP) achieved 87.69% on a segmented video dataset in [109]. An event-based dataset for event-based cameras was presented in [110], where a convolutional neural network achieved up to 77% accuracy. In [111], an LSTM model for Indian SL achieved 96% accuracy on the training set and 87% on the test set, while in [112], a ResNet-LSTM model for Argentine SL achieved 86.25% accuracy. Despite such high performance, many models show significant performance degradation in user-independent settings, which highlights the need for further research into generalization issues in this area.

The work in [113] proposed a method for dynamically adjusting the sampling rate to improve recognition accuracy in mobile environments, increasing performance to 83.64% (Top 5) and 66.54% (Top 1).

Other regional studies confirmed this trend. An ASL recognition system achieved 94% accuracy [114]. Moroccan SL recognition with a 3D CNN reached 99.6% [115]. ASL gestures and non-manual markers from RGB-D achieved 92.88% accuracy [116]. Experiments on the Mexican SL recognition keypoint sign signal dataset showed that the proposed arm-movement approach achieved 85.78% accuracy using six keypoints [117]. Chinese SL with optimized CNN-CB reached 94.88% [118]. Filipino SL with LSTM and MediaPipe reached 94% [119]. Arabic SL recognition with CNN-RNN reached 98% [120]. The study in [121] proposed a Mexican SL recognition system based on an RNN using spatial tracking of hands and facial expressions, achieving 0.93 accuracy in offline mode. Finally, multilingual systems are also emerging. A Visual Geometry Group (VGG)-19 network-based model achieved 99.9% accuracy for English signs and supported translation into Indian regional languages [122]. However, such high performance often depends on dataset-specific conditions and may not transfer effectively across languages or domains.

Overall, modern ISLR systems report recognition rates ranging from 87% to 100%. These results confirm the effectiveness of deep learning and multimodal technologies for both isolated and regional SL.

In continuous recognition tasks where it is necessary to interpret gestures as a sequence of glosses, the WER becomes a critical indicator of the average error per word, reflecting the discrepancy between the predicted and reference sequences. In [123], a new method for

detecting semantic boundaries based on reinforcement learning was proposed for poorly controlled CSLR, which allows for explicit alignment of video frames with gesture words. The method achieves a WER of 25.5% on the CSL Split II and RWTH-PHOENIX-Weather 2014 datasets. In [124], the Self-Mutual Knowledge Distillation (SMKD) method for CSLR was proposed, which simultaneously optimizes visual and contextual modules, avoiding the disadvantages of error backpropagation and addressing the CTC "spike effect" problem. In [125], the Visual Alignment Constraint (VAC) was introduced, improving the feature extractor, reducing overfitting, and making networks competitive in end-to-end setups. Despite these improvements, WER remains sensitive to alignment errors and the quality of gloss annotations, which limits robustness in real-world scenarios.

The study in [126] suggests an automatic annotation algorithm based on a Bi-Gated Recurrent Unit (BGRU) and Connectionist Temporal Classification (CTC), which reduces WER by 0.3% compared to modern SLR methods. Building on Transformer-based CSLR models, two additional constraints—spatial attention guided by keypoints and sentence embedding consistency—enabled strong performance on PHOENIX-2014, PHOENIX-2014-T, and SLR datasets [127]. The research in [128] introduces TransASL, an end-to-end real-time recognition and translation system using a wearable device to capture manual and non-manual markers, reaching a WER of 8.3% at the word level and 7.1% at the sentence level for 80 ASL words and 40 sentences.

A Self-Emphasizing Network (SEN) was presented in [129], highlighting informative spatial regions and discriminative frames, and achieving top accuracy on PHOENIX14, PHOENIX14-T, CSL-Daily, and CSL with minimal overhead. Two additional CSLR constraints were introduced in [130], a keyword-driven spatial attention module and sentence embedding consistency, which improved results on PHOENIX-2014, PHOENIX-2014-T, and CSL. The study in [131] proposes F2DCNet, a fully two-dimensional convolutional network for CSLR, relying only on 2D CNNs to extract spatiotemporal features and showing competitive performance on large datasets.

Article [132] suggests a multisensory CSLR framework combining RGB video, inertial measurement unit (IMU), and Surface Electromyography (sEMG) signals with BiLSTM and CTC, achieving a WER of 10.3% when modalities are fused. In [133], TCN is proposed as a hybrid network modeling spatiotemporal information from trajectories and correlated regions, yielding strong results on PHOENIX14, PHOENIX14-T, CSL, and CSL-Daily. By combining a gesture dictionary, an isolated recognition model, and a sliding window method, the study advanced the first steps toward online CSLR [134]. Finally, ref. [135] introduces a cross-modal complementation approach that generates pseudo video–text pairs by cross-modal editing, helping to resolve discrepancies between CTC and WER losses. However, multimodal approaches introduce additional challenges related to data synchronization, sensor availability, and system complexity.

Modern CSLR research demonstrates a steady downward trend in WER through the use of Transformers, bi-recurrent networks, attention mechanisms, and multimodal integration. These advances improve the alignment of video sequences with language units, mitigate CTC limitations, and enhance semantic interpretation of gestures, allowing recent systems to achieve WER below 10%, which moves CSLR closer to practical interactive communication.

The mean Average Precision (mAP) metric is actively used in tasks where it is necessary not only to classify but also to localize gestures in the video time interval. Lightweight and real-time architectures are increasingly used in SLR tasks, especially for mobile and embedded applications. An Android application was developed for two-way communication in Bahasa Isyarat Malaysia (BIM). It recognizes BIM letters using MobileNet, achieving 99.75% accuracy, and BIM gestures using SSD-MobileNet-V2 Feature Pyra-

mid Network Lite (FPNLite), achieving 61.60% accuracy [136]. An automatic Bangla SL (BSL) recognition system was developed for Jetson Nano, where YOLOv7-Tiny achieved the best real-time performance compared to Detectron2 and EfficientDet-D0 [137]. Another study introduced a model for both classification and correctness evaluation of gestures, combining flow-controlled features, 3D Conv, 2D deformable convolution, and attention mechanisms [138].

Real-time recognition of ASL letters (A–Z) was achieved using a pre-trained YOLOv8 model [139]. For Bengali gestures, an end-to-end system with YOLOv4-Tiny (99.7% mAP for 49 characters) and LSTM text generation (99.12% accuracy) was proposed [140]. A YOLOv5-based framework for Telugu SL recognition achieved strong performance with F1 of 90.5% and mAP of 98.1%, balancing accuracy and computational efficiency [141]. Finally, a comparative study on hand detection highlighted YOLOv8n and YOLOv8s, achieving 86.7% mAP on Oxford Hand and 98.9% on EgoHand after 100 epochs of training [142]. These studies highlight that the integration of lightweight architectures and optimization mechanisms can achieve a balance between accuracy (up to 99.75% mAP) and speed, which is particularly important in mobile applications and real-time systems focused on comprehensive communication and gesture learning.

The Spatial–Temporal Multi-Cue (STMC) Transformer [143] significantly improved glossary-to-text and video-to-text translation on PHOENIX-Weather-2014T (BLEU +5 and +7, respectively) and ASLG-PC12 (BLEU +16), but also demonstrated the limitations of glossary-based representations and suggested the need for end-to-end learning. The task-aware instruction network for sign language translation (TIN-SLT) network [144] introduced task-based learning and feature integration to the Transformer, surpassing previous best results with 1.65 BLEU-4 on PHOENIX-2014T and 1.42 BLEU-4 on ASLG-PC12. The large-scale OpenASL dataset [145] was introduced, along with techniques for handling real-world conditions, such as gesture search and multi-task learning. Rule-based pseudo-vocabulary generation was proposed in [146] as a simple method to improve SLT, achieving advanced results on PHOENIX-Weather 2014T and ASLG-PC12. An initial SLT benchmark using a dilated 3D-based Transformer on the broad-coverage How2Sign dataset was presented in [147].

Further research focused on multimodal and transfer learning approaches. The gloss attention SLT network (GASLT) model [148] combined RGB video and 3D pose data and improved BLEU by 18.39%. Transfer learning was explored in [149], where a Twin Residual Net outperformed multiple ResNet variants, achieving an average BLEU of 0.46. A comparative study [150] showed that ResNet-101 was the most effective backbone for video-based SLT in an LSTM framework. The study in [151] described that the quality of translation has a significant impact on the selection of algorithm optimization, activation function, and label smoothing, which is a key consideration in careful learning design. Performance on the PHOENIX-Weather 2014T dataset was improved by a graph-based dynamic multimodal SLT model proposed in [152] that integrates additional semantic knowledge at the word level. However, BLEU-based evaluation may not fully capture the semantic correctness and naturalness of sign language translation, especially in complex linguistic contexts.

Therefore, the selection of evaluation metrics must be carefully aligned with task-specific requirements, the characteristics of the captured body movements, and the desired balance between recognition accuracy and computational efficiency.

6. Comparative Analysis of Methods and Models

This section analyzes modern methods of SLR based on video data. Special attention is paid to the study of key models and approaches, their architectural features, accuracy,

performance, and computational efficiency, structured within the framework of the proposed classification.

According to the proposed classification, the main random-access schemes are summarized for each category in Tables S1 and S2. To make a clear comparison of the various methods, we highlight the main goal, method, performance improvements, and limitations of each approach. Table 1 shows modern models and their performance in SLR tasks. Most of the methods demonstrate high accuracy and efficiency and solve specific tasks such as sentence translation or multimodal feature fusion. CBAM-3DResNet-based approaches achieve over 90% accuracy [27]; LSTM [34,35] models have shown great performance gains on specific data. The limitations of these methods include the lack of datasets, the complexity of processing hidden features, and the need for complex, large datasets.

Table S3 provides modern approaches to sign language tasks, including performance analysis, creation of new datasets, and architecture development. Most of the methods demonstrate high accuracy, for example, greater than 96% recognition on AUTSL [91], and successfully solve specialized tasks such as modeling phonological properties or developing multimodal systems. However, most solutions face limitations, such as the difficulty of distinguishing visually similar gestures, a lack of data, or the impact of uncontrolled conditions. These methods require further consideration of cultural aspects, improvement of interpretability of models, and ensuring equity in access to data.

Table S4 compares methods for providing complexity and individualization in SLR tasks. These methods, such as PoseNet-LSTM [33], have shown high accuracy. Many approaches take into account specific user needs, such as developing energy-efficient SLR applications or ensuring equal access to knowledge using ResNet-50 [30]. However, some methods face challenges such as the complexity of processing noisy data or the need to improve performance with limited computational resources. Ignoring these aspects may limit the development of complex and robust systems that can effectively operate in real-world situations.

Table S5 provides an analysis of the performance and robustness of various SLR methods, evaluating them in terms of accuracy, processing speed, real-time applicability, coverage, and noise immunity. Despite the high robustness of multimodal methods to noise, most are computationally intensive, which limits their practical application in real time. The Multi-Stream Graph-Based Deep Neural Network (SL-GDN) [46] model achieves high performance with relatively low computational load even on complex backgrounds, but its generalizability across different scenarios remains limited.

HA-GCN and multi-branch GCN architectures show more efficient representation of skeletal data in dynamic situations, improving spatiotemporal modeling, but may face scaling issues on large datasets. In practical applications, the Time–Frequency Network (TFNet) can handle complex scenarios such as the Chinese Continuous Sign Language Dataset (CE-CSL), but its performance is highly dependent on the quality and diversity of the training data. The AUTSL dataset demonstrates the benefits of multimodal approaches and contextual diversity, but also shows that existing methods still need improvement to fully generalize to new users and survey situations.

Table S5 presents the results of the evaluation of various methods aimed at improving the performance of SLR and SLT systems. The methods were compared on criteria such as accuracy, processing speed, real-time capability, coverage, and noise immunity. The analysis included the development of practical and inclusive solutions for 3D CNNs, cross-attention Transformers, and multimodal approaches that combine information from gestures, posture, and facial expressions. The methods under study may represent technological progress in terms of efficiency, noise immunity, and adaptability. However, challenges remain in

reducing computational requirements and improving model interpretation, which represent key areas for future research.

7. Discussions

Recent advances in the field of video-based SLR have seen significant progress [153–163], largely due to developments in deep learning, multimodal data analysis, and the creation of specialized datasets. The systematic analysis conducted in this review across six architectural categories reveals several overarching patterns that transcend individual method comparisons and deserve explicit discussion.

A clear evolutionary trajectory emerges from the reviewed literature. Early CNN and 3D-CNN approaches established strong spatial feature extraction baselines but were fundamentally limited by their inability to model long-range temporal dependencies. Recurrent architectures such as LSTM and GRU addressed sequential modeling but introduced information bottlenecks through fixed-size hidden states, which caused performance degradation on longer and more complex signing sequences. Transformer-based models subsequently resolved the temporal dependency problem through attention mechanisms but introduced new challenges related to data requirements and computational cost. GCN-based approaches offered geometric robustness by explicitly modeling skeletal structure but remained vulnerable to pose estimation failures in uncontrolled environments. The hybrid architectures currently represent the state-of-the-art by combining the complementary strengths of these paradigms, yet they simultaneously amplify the data requirements of their constituent components. This pattern suggests that architectural progress in SLR has been largely driven by addressing the limitations of the preceding dominant paradigm rather than by a principled unified framework, and that a more productive direction would be the development of architectures that simultaneously address spatial precision, temporal modeling, geometric structure, and computational efficiency within a single coherent design.

It is important to note that this evolutionary progression also reflects a broader shift in SLR research from isolated benchmark optimization towards attempts to achieve robustness, scalability, and practical deployment feasibility under real-world communication conditions. As architectures become increasingly sophisticated, the central challenge is no longer limited to maximizing recognition accuracy but rather to achieving an effective balance between temporal modeling capacity, geometric precision, computational efficiency, and cross-domain generalization.

As shown in Table 3, a clear performance–complexity trade-off emerges across the six architectural categories reviewed. CNN- and RNN-based approaches offer the best balance of computational efficiency and real-time applicability, but their robustness to real-world variation remains limited, particularly in cross-dataset and signer-independent scenarios. Transformer-based models achieve higher accuracy on large benchmarks but require substantial annotated data and computational resources, limiting their direct applicability to low-resource sign languages. GCN-based approaches provide structural robustness through skeletal modeling but remain vulnerable to pose estimation failures in uncontrolled environments. Hybrid architectures represent the current state-of-the-art in terms of accuracy but amplify the data and computational requirements of their constituent components, creating a significant barrier to deployment in resource-constrained settings. Transfer learning and self-supervised approaches offer the most promising path for low-resource SLR, but their effectiveness depends critically on domain proximity between source and target languages. These trade-offs suggest that no single architectural paradigm currently satisfies all requirements for robust, generalizable, and computationally efficient real-world SLR (see Figure 9), and that future progress will depend on princi-

pled integration strategies rather than continued optimization of individual components in isolation.

Table 3. Cross-architectural comparison of deep learning approaches for video-based SLR.

Architecture	Accuracy Range	Computational Complexity	Robustness to Real-World Variation	Data Requirements	Suitable for Low-Resource SL	Real-Time Applicability
CNN/3D-CNN	83–99%	Medium	Low—sensitive to background and lighting	Medium	Limited	Yes (lightweight variants)
RNN/ LSTM/GRU	88–99%	Low–Medium	Medium—degrades on long sequences	Low–Medium	Moderate	Yes
Transformer/ Attention	85–98%	High	Medium–High	High	Limited without pre-training	Partial
GCN-based	88–100%	Medium–High	Medium—depends on pose estimation quality	Medium	Moderate	Partial
Hybrid End-to-End	90–99.97%	Very High	High	Very High	Low	Limited
Transfer/ Self-supervised	80–95%	Medium	Medium–High	Low (labeled)	High	Partial

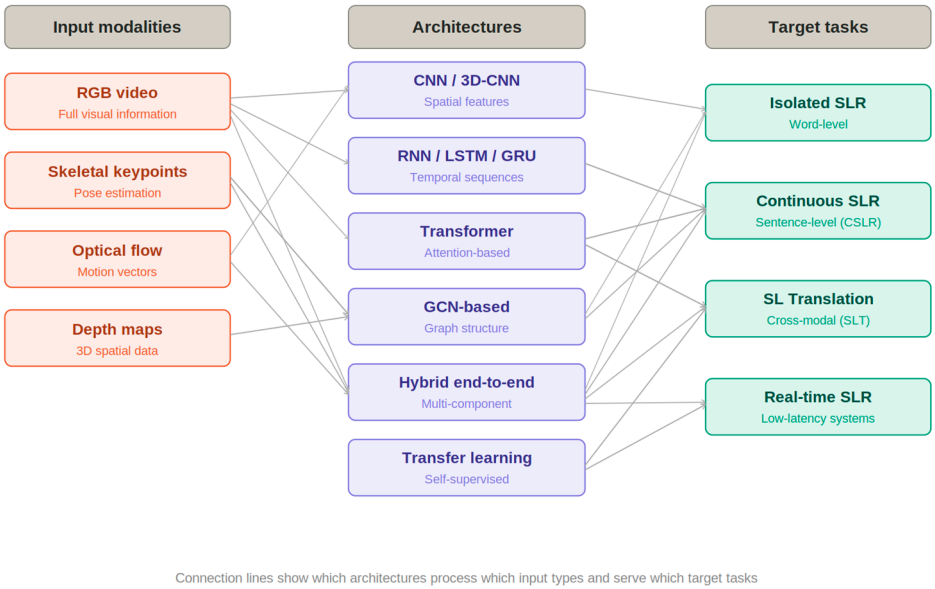


Figure 9. Unified system-level framework showing relationships between input modalities, deep learning architectural paradigms, and target tasks in video-based SLR. Connection lines indicate which architectures process which input types and serve which target tasks.

It is suggested by these observations that future SLR systems should be evaluated not only according to benchmark accuracy metrics but also according to deployment-oriented criteria. These include, but are not limited to, inference latency, memory efficiency, robustness to environmental variability, and adaptability across signer populations and recording conditions. The evaluation principles outlined here are of particular importance for assistive communication systems intended for continuous real-world interaction in environments external to laboratory settings.

A persistent and systematic gap exists between reported benchmark performance and real-world applicability. The majority of reviewed studies report accuracy figures above 90% on standard datasets, yet practical deployment of SLR systems in real assistive communication contexts remains limited. This gap arises from three structural factors that are consistently present across the reviewed literature: most benchmarks are collected

under controlled laboratory conditions that do not reflect natural signing environments; evaluation is typically performed in signer-dependent settings that mask the model's inability to generalize to unseen users; and dataset-specific fine-tuning inflates reported metrics in ways that do not transfer across languages or domains. Future evaluations should prioritize signer-independent protocols and cross-dataset generalization benchmarks as standard reporting requirements across the field.

The three core challenges identified throughout this review are data scarcity, generalization, and computational complexity that deserve deeper analysis than a simple enumeration. The data scarcity in sign language research is not merely a quantitative problem of insufficient samples, but a structural problem rooted in the requirements of the data collection process itself. Unlike speech recognition, where large-scale data collection is relatively straightforward, sign language data collection requires native signers, specialized recording infrastructure, linguistically informed annotation protocols, and ethical frameworks for working with deaf communities. These requirements create a compounding barrier that disproportionately affects low-resource sign languages, including Kazakh, Ethiopian, Azerbaijani, and hundreds of others, where the community of native signers is small and institutional support is limited. The transfer learning and self-supervised pre-training offer partial mitigation, but their effectiveness depends on domain proximity between source and target languages, which cannot be assumed across structurally distinct sign language families. The generalization problem is similarly more complex than it appears: sign language production is inherently variable across signers, regional dialects, and communicative contexts, and this variability is not noise to be filtered out but linguistically meaningful variation that must be modeled explicitly. The computational complexity challenge creates a direct tension with the primary use case of SLR systems, namely real-time assistive communication on resource-constrained devices, and model compression, knowledge distillation, and efficient architecture design remain underexplored directions in the SLR literature.

Based on the systematic analysis conducted in this review, we identify the following priority research directions organized by time horizon. In the short term (1–2 years), the field would benefit most from standardization of evaluation protocols, specifically the adoption of signer-independent splits and cross-dataset evaluation as minimum reporting requirements, and from the development of linguistically validated benchmark datasets for underrepresented sign languages, with particular priority given to sign languages of Central Asia, Africa, and Southeast Asia. Additionally, the systematic evaluation of existing state-of-the-art architectures under cross-domain conditions, specifically across datasets collected in different recording environments, with different signer demographics, and for different sign language families, should be established as a standard validation requirement to provide more reliable estimates of real-world generalization capability. In the medium term (2–4 years), the most impactful directions include the development of large-scale multilingual sign language pre-training corpora enabling cross-lingual transfer learning, the integration of geometric skeletal representations with Transformer-based temporal modeling to address fine-grained motion modeling gaps, and the development of adaptive multimodal fusion strategies that dynamically weight manual and non-manual features rather than applying static equal-weight concatenation. In the long term (4+ years), the field needs end-to-end SLR systems capable of handling the full linguistic complexity of natural sign language, including non-manual markers, spatial grammar, and prosodic features, as well as participatory design frameworks that involve deaf communities as active partners in dataset creation, model evaluation, and system deployment.

A cross-cutting theme emerging from this review is the structural underrepresentation of low-resource sign languages in the deep learning literature. Of the 103 reviewed studies,

the majority focus on ASL, CSL, or ArSLs, with only a small number addressing languages such as Kazakh, Ethiopian, or Azerbaijani Sign Language. This distributional imbalance reflects existing inequalities in access to assistive communication technology. The development of resources such as KSLVD, discussed in Section 4, represents one concrete step toward addressing this gap, but the broader challenge requires coordinated effort across linguistics, computer science, and deaf education communities. Future work should treat the geographic and linguistic distribution of studied sign languages as an explicit metric of field inclusivity, and funding bodies should prioritize research on underrepresented sign languages as a matter of equity in assistive technology development.

Deep learning methods have been shown to significantly improve SLR and translation performance, and the integration of multimodal architectures, Transformer-based models, and self-supervised learning strategies represents the most promising current direction. However, it is essential that proposed SLR methods increasingly take into account not only technical performance metrics but also real-world usability, linguistic validity, and equitable access across sign language communities. These considerations should guide the next generation of research in this field.

8. Conclusions

This review analyzes systematic deep learning approaches and their limitations. Furthermore, the lack of high-quality standardized datasets, especially for regional sign language variants, significantly complicates the development and validation of systems that handle large vocabularies and huge amounts of data. The KSLVD was developed as a standardized resource for sentence-level KSL recognition.

The obtained conclusions emphasize the need for further research aimed at improving algorithms, creating larger and more diverse datasets, as well as developing universal and user-difference-resistant models for the practical implementation of SLR technologies. We hope that the results of this review will be useful to future researchers, providing systematic information, references and inspiration for the development of new approaches and solutions in the field of SLR.

Supplementary Materials: The following supporting information can be downloaded at: https://docs.google.com/document/d/1wYwW0IOaziNB9W1ANisS2r9S0J_4rQ9Z/edit?usp=sharing&ouid=117959307128899551459&rtpof=true&sd=true.

Author Contributions: Conceptualization, U.B. and A.Y.; methodology, U.B.; formal analysis, M.M.; investigation, U.B. and B.S.; resources, B.S. and U.B.; data curation, E.D. and D.T.; writing—original draft preparation, U.B., A.Y. and M.M.; writing—review and editing, U.B., A.Y., D.O. and E.D.; project administration, U.B.; supervision, U.B. and A.Y.; project administration, A.Y.; funding acquisition, A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992875).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Traini, D.; Alcalde-Llergo, J.M.; Buenestado-Fernández, M.; Ursino, D.; Yeguas-Bolívar, E. Behavioral engagement in VR-based sign language learning: Visual attention as a predictor of performance and temporal dynamics. *Multimodal Technol. Interact.* **2026**, *10*, 23. [CrossRef]

2. Doan, H.G.; Nguyen, N.T. Fusion Machine Learning Strategies for Multi-modal Sensor-based Hand Gesture Recognition. *Eng. Technol. Appl. Sci. Res.* **2022**, *12*, 8628–8633. [CrossRef]
3. Huang, Z.; Xue, W.; Zhou, Y.; Sun, J.; Wu, Y.; Yuan, T.; Chen, S. Dual-stage temporal perception network for continuous sign language recognition. *Vis. Comput.* **2025**, *41*, 1971–1986. [CrossRef]
4. Badiger, R.M.; Yakkundimath, R.; Konnurmath, G.; Dhulavvagol, P.M. Deep Learning Approaches for Age-based Gesture Classification in South Indian Sign Language. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 13255–13260. [CrossRef]
5. Hashi, A.O.; Hashim, S.Z.M.; Asamah, A.B. Dynamic Adaptation in Deep Learning for Enhanced Hand Gesture Recognition. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 15836–15841. [CrossRef]
6. Jebali, M.; Dakhli, A.; Jemni, M. Vision-based continuous sign language recognition using multimodal sensor fusion. *Evol. Syst.* **2021**, *12*, 1031–1044. [CrossRef]
7. Tang, S.; Guo, D.; Hong, R.; Wang, M. Graph-based multimodal sequential embedding for sign language translation. *IEEE Trans. Multimed.* **2021**, *24*, 4433–4445. [CrossRef]
8. Nagavasista, B.; Datta, M.V.; Mounish, A.S.; Prachotan, U.; Anjali, T.; Abhishek, S. A robust hand gesture recognition model for assistive typing systems using YOLOv8. In Proceedings of the 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 6–8 November 2024; pp. 553–560. [CrossRef]
9. Vijitkunsawat, W.; Anunvrapong, P.; Chantngarm, P. Human joint coordinate sequencing in video-based Thai finger spelling recognition. In Proceedings of the 2024 8th International Conference on Information Technology (InCIT), Chonburi, Thailand, 14–15 November 2024; pp. 681–686. [CrossRef]
10. Khnyunin, M.V.; Grif, M.G. Sign language translation system using machine vision. In Proceedings of the 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russia, 15–17 November 2024; pp. 940–943. [CrossRef]
11. Katti, R.K.; Sujatha, C.; Desai, P.; Shankar, G. Character and word level gesture recognition of Indian Sign Language. In Proceedings of the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 7–9 April 2023; pp. 1–6. [CrossRef]
12. Mullapudi, N.; Suma, G.J. Video Level Sign Language Recognition with Key Frame Extraction Using Adaptive Convolution Neural Networks with a New Activation Function. *Eng. Technol. Appl. Sci. Res.* **2025**, *15*, 30356–30361. [CrossRef]
13. Liu, Z.; Pang, L.; Qi, X. MEN: Mutual enhancement networks for sign language recognition and education. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *35*, 311–325. [CrossRef]
14. Jiang, Y.; Li, F.; Li, Z.; Liu, Z.; Wang, Z. Enhancing continuous sign language recognition with self-attention and MediaPipe Holistic. In Proceedings of the 2023 8th International Conference on Instrumentation, Control, and Automation (ICA), Jakarta, Indonesia, 9–11 August 2023; pp. 97–102. [CrossRef]
15. Liu, K.; Hou, Y.; Guo, Z.; Yin, W.; Ren, Y. Visual context learning based on cross-modal knowledge for continuous sign language recognition. *Vis. Comput.* **2024**, *41*, 4017–4031. [CrossRef]
16. Hassan, E.; Shams, M.Y.; El-Hafeez, T.A.; Elseddik, M. A novel model for expanding horizons in sign language recognition. *Sci. Rep.* **2025**, *15*, 24358. [CrossRef]
17. Kındıroglu, A.A.; Özdemir, O.; Akarun, L. Aligning accumulative representations for sign language recognition. *Mach. Vis. Appl.* **2023**, *34*, 12. [CrossRef]
18. Al Abdullah, B.A.; Amoudi, G.A.; Alghamdi, H.S. Advancements in sign language recognition: A comprehensive review and future prospects. *IEEE Access* **2024**, *12*, 128871–128895. [CrossRef]
19. Feng, Y.; Chen, N.; Wu, Y.; Jiang, C.; Liu, S.; Chen, S. DFCNet+: Cross-modal dynamic feature contrast net for continuous sign language recognition. *Image Vis. Comput.* **2024**, *151*, 105260. [CrossRef]
20. Violet, I.M.M.; Sri, R.A.L. A comprehensive survey on recent advances and challenges in sign language recognition systems. *Discov. Artif. Intell.* **2025**, *5*, 419. [CrossRef]
21. Priyanto, D.; Hairani. A Hijaiyah Letters Sign Language Recognition Approach utilizing Deep Learning. *Eng. Technol. Appl. Sci. Res.* **2025**, *15*, 22452–22458. [CrossRef]
22. Buribayev, Z.; Aouani, M.; Zhangabay, Z.; Yerkos, A.; Abdirazak, Z.; Zhassuzak, M. Enhancing Kazakh Sign Language Recognition with BiLSTM Using YOLO Keypoints and Optical Flow. *Appl. Sci.* **2025**, *15*, 5685. [CrossRef]
23. Moryossef, A.; Tsochantaridis, I.; Dinn, J.; Camgoz, N.C.; Bowden, R.; Jiang, T.; Rios, A.; Muller, M.; Ebling, S. Evaluating the immediate applicability of pose estimation for sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 3434–3440. [CrossRef]
24. Janpirom, C.; Chunngam, B.; Phewkham, T.; Suksukont, A. A Two-Stream Convolutional Attention Network for Hand Gesture Recognition and Classification. *Eng. Technol. Appl. Sci. Res.* **2026**, *16*, 30963–30970. [CrossRef]
25. Al-Hammadi, M.; Muhammad, G.; Abdul, W.; Alsulaiman, M.; Bencherif, M.A.; Mekhtiche, M.A. Hand gesture recognition for sign language using 3DCNN. *IEEE Access* **2020**, *8*, 79491–79509. [CrossRef]
26. Sharma, S.; Kumar, K. ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. *Multimed. Tools Appl.* **2021**, *80*, 26319–26331. [CrossRef]

27. Khan, R.; Farooq, M.; Noordin, N.K. Evaluating the efficiency of CBAM-ResNet using Malaysian sign language. *Comput. Mater. Contin.* **2022**, *71*, 2755–2772. [CrossRef]
28. Louison, N.; Goodridge, W.; Khan, K. Learning sign language representation using CNN LSTM, 3DCNN, CNN RNN LSTM and CCN TD. *arXiv* **2024**, arXiv:2412.18187. [CrossRef]
29. Saxena, R.; Maggo, C.; Upmanyu, D.; Rathi, D.; Tanwar, A.S. A deep learning-based sign language recognition system using convolutional neural networks. In Proceedings of the 2025 2nd International Conference on Intelligent Systems for Cybersecurity (ISCS), Gurugram, India, 14–15 November 2025; pp. 1–6. [CrossRef]
30. Bedaso, M.; Hussein, A.; Diriba, C. Developing sign language recognition model for Afaan Oromoo words using deep learning techniques. *Preprint* **2024**. [CrossRef]
31. Kothadiya, D.; Patel, J.; Gupta, N.; Shah, R. DeepSign: Sign language detection and recognition using deep learning. *Electronics* **2022**, *11*, 1780. [CrossRef]
32. Mejía-Peréz, K.; Córdova-Esparza, D.; Terven, J.; Herrera-Navarro, A.; García-Ramírez, T.; Ramírez-Pedraza, A. Automatic recognition of Mexican sign language using a depth camera and recurrent neural networks. *Appl. Sci.* **2022**, *12*, 5523. [CrossRef]
33. Shetty, S.; Hegde, R.; Prabhu, A. Gesture-to-text: A real-time Indian sign language translator with pose estimation and LSTMs. *Procedia Comput. Sci.* **2024**, *235*, 2684–2692. [CrossRef]
34. Mariappan, S.; Murugesan, P.; Selvan, H. Real-time interpreter for short sentences in Indian sign language using MediaPipe and deep learning. *Inf. Technol. Control* **2024**, *53*, 888–898. [CrossRef]
35. Oulad-Naoui, S.; Ben-Abderrahmane, H.; Chagha, A.; Cherif, A. An LSTM-based system for dynamic Arabic sign language recognition. In *Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS), Aflou, Algeria, 26–27 June 2024*; Atlantis Press: Dordrecht, The Netherlands, 2024; pp. 313–323. [CrossRef]
36. Bhadouria, A.; Bindal, P.; Khare, N.; Singh, D.; Verma, A. LSTM-based recognition of sign language. In *Proceedings of the 16th International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2024*; ACM: New York, NY, USA, 2024; pp. 508–514. [CrossRef]
37. Camgoz, N.C.; Hadfield, S.; Koller, O.; Bowden, R. Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10023–10033. [CrossRef]
38. Zhang, H.; Sun, Y.; Liu, Z.; Liu, Q.; Liu, X.; Jiang, M.; Fang, H. Heterogeneous attention-based transformer for sign language translation. *Appl. Soft Comput.* **2023**, *144*, 110526. [CrossRef]
39. Javaid, S.; Rizvi, S. A novel action transformer network for hybrid multimodal sign language recognition. *Comput. Mater. Contin.* **2023**, *75*, 523–537. [CrossRef]
40. Shin, J.; Miah, A.S.M.; Hasan, M.A.M.; Hirooka, K.; Suzuki, K.; Lee, H.S.; Jang, S.W. Korean sign language recognition using transformer-based deep neural network. *Appl. Sci.* **2023**, *13*, 3029. [CrossRef]
41. Hakim, Z.I.A.; Swargo, R.M.; Adnan, M.A. Attention-driven multi-modal fusion: Enhancing sign language recognition and translation. *arXiv* **2023**, arXiv:2309.01860. [CrossRef]
42. Zheng, J.; Wang, Y.; Tan, C.; Li, S.; Wang, G.; Xia, J.; Li, S. CVT-SLR: Contrastive visual-textual transformation for sign language recognition with variational alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 23141–23150. [CrossRef]
43. Ranjbar, H.; Taheri, A. Continuous sign language recognition using intra-inter gloss attention. *arXiv* **2024**, arXiv:2406.18333. [CrossRef]
44. Aloysius, N.; Geetha, M.; Nedungadi, P. Incorporating relative position information in transformer-based sign language recognition and translation. *IEEE Access* **2021**, *9*, 145929–145942. [CrossRef]
45. Tunga, A.; Nuthalapati, S.V.; Wachs, J. Pose-based sign language recognition using GCN and BERT. *arXiv* **2021**, arXiv:2012.00781. [CrossRef]
46. Miah, A.S.M.; Hasan, M.; Jang, S.-H.; Lee, H.; Shin, J. Multi-stream graph-based deep neural networks for skeleton-based sign language recognition. *Electronics* **2023**, *12*, 2841. [CrossRef]
47. Miah, A.S.M.; Hasan, M.; Shin, J. Dynamic hand gesture recognition using multi-branch attention-based graph and general deep learning model. *IEEE Access* **2023**, *11*, 4703–4716. [CrossRef]
48. Naz, N.; Sajid, H.; Ali, S.; Hasan, O.; Ehsan, M. SignGraph: An efficient and accurate pose-based graph convolution approach toward sign language recognition. *IEEE Access* **2023**, *11*, 19135–19147. [CrossRef]
49. Song, J.; Lee, M.; Zhang, Q.; Chen, T. Hand-aware graph convolution network for skeleton-based sign language recognition. *J. Inf. Intell.* **2024**, *3*, 36–50. [CrossRef]
50. Miah, A.; Hasan, M.M.; Nishimura, S.; Shin, J. Sign Language Recognition Using Graph and General Deep Neural Network Based on Large Scale Dataset. *IEEE Access* **2024**, *12*, 34553–34569. [CrossRef]

51. Parelli, M.; Papadimitriou, K.; Potamianos, G.; Pavlakos, G.; Maragos, P. Spatio-Temporal Graph Convolutional Networks for Continuous Sign Language Recognition. In Proceedings of the ICASSP 2022—IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; pp. 8457–8461. [CrossRef]
52. Naz, N.; Sajid, H.; Ali, S.; Hasan, O.; Ehsan, M. MIPA-ResGCN: A multi-input part attention enhanced residual graph convolutional framework for sign language recognition. *Comput. Electr. Eng.* **2023**, *112*, 109009. [CrossRef]
53. Aly, M.; Fathi, I.S. Recognizing American Sign Language gestures efficiently and accurately using a hybrid transformer model. *Sci. Rep.* **2025**, *15*, 20253. [CrossRef]
54. Kondo, T.; Narumi, S.; He, Z.; Shin, D.; Kang, Y. A performance comparison of Japanese sign language recognition with ViT and CNN using angular features. *Appl. Sci.* **2024**, *14*, 3228. [CrossRef]
55. Kumar, L.; Yadav, M.; Pandey, R.; Singh, A.; Sharma, R.; Meena, M. Deep learning based assistive technology on visual speech recognition for hearing impaired. *Int. J. Cogn. Comput. Eng.* **2022**, *3*, 24–30. [CrossRef]
56. Natarajan, B.; Saini, H.; Mehta, S.; Kumar, N.; Singh, R. Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation. *IEEE Access* **2022**, *10*, 104358–104374. [CrossRef]
57. Chen, Y.; Zuo, R.; Wei, F.; Wu, Y.; Liu, S.; Mak, B. Two-stream network for sign language recognition and translation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 17043–17056. [CrossRef]
58. Hugar, G.; Kagalkar, R.M.; Das, A. Comparative study of hybrid deep learning models for Kannada sign language recognition. *Int. J. Comput. Intell. Syst.* **2025**, *18*, 191. [CrossRef]
59. Xue, S.; Zhang, R.; Li, Y.; Wang, J.; Liu, X.; Chen, M. Multi-scale context-aware network for continuous sign language recognition. *Virtual Real. Intell. Hardw.* **2024**, *6*, 323–337. [CrossRef]
60. Dey, A.; Biswas, S.; Le, D. Recognition of Wh-question sign gestures in video streams using an attention driven C3D-BiLSTM network. *Procedia Comput. Sci.* **2024**, *235*, 2920–2931. [CrossRef]
61. Chaudhari, P.; Patel, N.; Sharma, R. Sign language recognition using spiking neural networks. *Procedia Comput. Sci.* **2024**, *235*, 2674–2683. [CrossRef]
62. Wijaya, F.; Ramadhan, A.; Hidayati, N. Quantitative analysis of sign language translation using artificial neural network model. *Procedia Comput. Sci.* **2024**, *245*, 998–1009. [CrossRef]
63. Li, J.; Zhang, W.; Zhao, M.; Liu, Y. Uyghur sign language recognition based on improved YOLOv7. *Procedia Comput. Sci.* **2024**, *242*, 512–519. [CrossRef]
64. Yehalem, T.A.; Belay, H.B.; Million, M. *End-to-End Continuous Ethiopia Sign Language Recognition*; Zenodo: Geneva, Switzerland, 2024. [CrossRef]
65. Hasanov, J.; Alishzade, N.; Nazimzade, A.; Dadashzade, S.; Tahirov, T. Development of a hybrid word recognition system and dataset for the Azerbaijani Sign Language dactyl alphabet. *Speech Commun.* **2023**, *153*, 102960. [CrossRef]
66. Bhuiyan, H.J.; Mozumder, M.F.; Khan, M.R.I.; Ahmed, M.S.; Nahim, N.Z. Enhancing bidirectional sign language communication: Integrating YOLOv8 and NLP for real-time gesture recognition & translation. In Proceedings of the 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI), Kyoto, Japan, 28–31 March 2025; pp. 168–174. [CrossRef]
67. Padmapriya, G.; Roy, S.; Das, A. Gesture Whisperer: Sign language to text conversion. In Proceedings of the 3rd International Conference on Optimization Techniques in the Field of Engineering, Chennai, India, 22–23 October 2024. [CrossRef]
68. Fan, D.; Yi, M.; Kang, W.; Wang, Y.; Lv, C. Continuous sign language recognition algorithm based on object detection and variable-length coding sequence. *Sci. Rep.* **2024**, *14*, 27592. [CrossRef]
69. Du, Y.; Peng, T.; Hu, X. Beyond granularity: Enhancing continuous sign language recognition with granularity-aware feature fusion and attention optimization. *Appl. Sci.* **2024**, *14*, 8937. [CrossRef]
70. Liu, T.; Tao, T.; Zhao, Y.; Zhu, J. A two-stream sign language recognition network based on keyframe extraction method. *Expert. Syst. Appl.* **2024**, *253*, 124268. [CrossRef]
71. Sumon, S.I.; Ali, M.B.; Bari, S.; Ohi, I.R.; Islam, M.; Rahman, S.M. Sign language word detection using LRCN. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2024; Volume 1305, p. 012023. [CrossRef]
72. Luqman, H.; El-Alfy, E. Towards hybrid multimodal manual and non-manual Arabic sign language recognition: MArSL database and pilot study. *Electronics* **2021**, *10*, 1739. [CrossRef]
73. Li, D.; Rodriguez, C.; Yu, X.; Li, H. Transferring cross-domain knowledge for video sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6205–6214. [CrossRef]
74. Sincan, O.; Junior, J.C.S.J.; Escalera, S.; Keles, H.Y. Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Virtual, 19–25 June 2021; pp. 3472–3481. [CrossRef]
75. Gueuwou, S.; Du, X.; Shakhnarovich, G.; Livescu, K.; Liu, A. SHuBERT: Self-Supervised Sign Language Representation Learning via Multi-Stream Cluster Prediction. *arXiv* **2024**, arXiv:2411.16765. [CrossRef]

76. Kvanchiani, K.; Kalogeiton, M.; Dapogny, A. Training strategies for isolated sign language recognition. *arXiv* **2024**, arXiv:2412.11553. [CrossRef]
77. Kezar, L.; Munikote, N.; Zeng, Z.; Sehyr, Z.; Caselli, N.; Thomason, J. The American Sign Language knowledge graph: Infusing ASL models with linguistic knowledge. *arXiv* **2024**, arXiv:2411.03568. [CrossRef]
78. Sato, R.; Beleza, S.; Shimomoto, E.; de Lima, M.; Kato, N.; Fukui, K. Sign language recognition based on subspace representations in the spatio-temporal frequency domain. In Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM), Rome, Italy, 24–26 February 2024; pp. 152–159. [CrossRef]
79. Bird, J.; Ekárt, A.; Faria, D. British sign language recognition via late fusion of computer vision and Leap Motion with transfer learning to American sign language. *Sensors* **2020**, *20*, 5151. [CrossRef] [PubMed]
80. Li, D.; Rodriguez, C.; Yu, X.; Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1459–1469. [CrossRef]
81. Albanie, S.; Varol, G.; Momeni, L.; Afouras, T.; Chung, J.S.; Fox, N.; Zisserman, A. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 35–53. [CrossRef]
82. Alasmari, N.; Asiri, S. ASLDetect: Arabic sign language detection using ResNet and U-Net like component. *Sci. Rep.* **2025**, *15*, 18012. [CrossRef] [PubMed]
83. Alsulaiman, M.; Faisal, M.; Mekhtiche, M.; Bencherif, M.; Alrayes, T.; Muhammad, G.; Mathkour, H.; Abdul, W.; Alohali, Y.; Alqahtani, M.; et al. Facilitating the communication with deaf people: Building a largest Saudi sign language dataset. *J. King Saud Univ. Comput. Inf. Sci.* **2023**, *35*, 101642. [CrossRef]
84. Tavella, F.; Schlegel, V.; Romeo, M.; Cangelosi, A.; Galata, A. WLASL-LEX: A dataset for recognising phonological properties in American Sign Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; Volume 2: Short Papers, pp. 453–463. [CrossRef]
85. Özdemir, O.; Kındıroğlu, A.; Camgöz, N.C.; Akarun, L. BosphorusSign22k sign language recognition dataset. *arXiv* **2020**, arXiv:2004.01283. [CrossRef]
86. Abbas, S.; Alahmadi, D.; Al-Barhamtoshy, H. Establishing a multimodal dataset for Arabic Sign Language (ArSL) production. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102165. [CrossRef]
87. Hu, L.; Gao, L.; Liu, Z.; Feng, W. Temporal lift pooling for continuous sign language recognition. In *Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 511–527. [CrossRef]
88. Desai, A.; Berger, L.; Minakov, F.; Milano, N.; Singh, C.; Pumphrey, K.; Ladner, R.E.; Daumé, H., III; Lu, A.X.; Caselli, N.; et al. ASL Citizen: A community-sourced dataset for advancing isolated sign language recognition. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 76893–76907. [CrossRef]
89. Kapitanov, A.; Karina, K.; Nagaev, A.; Pavlova, E. Slovo: Russian sign language dataset. In *Proceedings of the International Conference on Computer Vision Systems, Vienna, Austria, 27–29 September 2023*; Springer: Cham, Switzerland, 2023; pp. 63–73. [CrossRef]
90. Yu, Z.; Huang, S.; Cheng, Y.; Birdal, T. Signavatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *Proceedings of the Computer Vision—ECCV 2024, Milan, Italy, 29 September–4 October 2024*; Springer: Cham, Switzerland, 2024; pp. 1–19. [CrossRef]
91. Sincan, O.; Keles, H. AUTSL: A large-scale multi-modal Turkish sign language dataset and baseline methods. *IEEE Access* **2020**, *8*, 181340–181355. [CrossRef]
92. Yerimbetova, A.; Sakenov, B.; Sambetbayeva, M.; Daiyrbayeva, E.; Berzhanova, U.; Othman, M. Creating a parallel corpus for the Kazakh sign language and learning. *Appl. Sci.* **2025**, *15*, 2808. [CrossRef]
93. Mukushev, M.; Ubingazhibov, A.; Kydyrbekova, A.; Imashev, A.; Kimmelman, V.; Sandygulova, A. FluentSigners-50: A signer independent benchmark dataset for sign language processing. *PLoS ONE* **2022**, *17*, e0273649. [CrossRef] [PubMed]
94. Aishwarya, A.; D'Souza, D.J. Sign language recognition. In Proceedings of the 2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Nitte, India, 19–20 November 2021; pp. 104–106. [CrossRef]
95. Chophuk, P.; Chamnongthai, K. Backhand-view-based continuous-signed-letter recognition using a rewound video sequence and the previous signed-letter information. *IEEE Access* **2021**, *9*, 40187–40197. [CrossRef]
96. Rathinagiriswaran, M.; Managaokar, S.; Jois, K.R.Y.; Suvarna, K.V.; Krupa, N. Inflated 3D architecture for South Indian sign language recognition. In Proceedings of the 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, India, 3–4 December 2021; pp. 1–6. [CrossRef]
97. Bencherif, M.A.; Algabri, M.; Mekhtiche, M.; Faisal, M.; Alsulaiman, M.; Mathkour, H. Arabic sign language recognition system using 2D hands and body skeleton data. *IEEE Access* **2021**, *9*, 59612–59627. [CrossRef]

98. Khan, R.; Khattak, H.; Wong, W.; AlSalman, H.; Mosleh, M.; Rahman, S. Intelligent Malaysian sign language translation system using convolutional-based attention module with residual network. *Comput. Intell. Neurosci.* **2021**, *2021*, 9023010. [CrossRef]
99. Elhagry, A.; Elrayes, R. Egyptian sign language recognition using CNN and LSTM. *arXiv* **2021**, arXiv:2107.13647. [CrossRef]
100. Mistree, K.; Thakor, D.; Bhatt, B. Towards Indian sign language sentence recognition using INSIGNVID: Indian sign language video dataset. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 881. [CrossRef]
101. Zhou, Z.; Lui, K.; Tam, V.; Lam, E. Applying (3+2+1)D residual neural network with frame selection for Hong Kong sign language recognition. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4296–4302. [CrossRef]
102. Aitim, A.; Sattarkhuzhayeva, D.; Khairullayeva, A. Development of a hybrid CNN-RNN model for enhanced recognition of dynamic gestures in Kazakh sign language. *East.-Eur. J. Enterp. Technol.* **2025**, *2*, 58–67. [CrossRef]
103. Mostavi, M.M.; Islam, R.; Ahsan, M. Phrase level Bangla sign language recognition using keypoints from hand gesture video. In Proceedings of the International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh, 16–17 June 2023; pp. 1–6. [CrossRef]
104. Reyazul, H.K.; Adnan, M.A. EMPATH: MediaPipe-aided ensemble learning with attention-based transformers for accurate recognition of Bangla word-level sign language. In *Proceedings of the International Conference on Pattern Recognition, Kolkata, India, 1–5 December 2024*; Springer: Cham, Switzerland, 2024; pp. 355–371. [CrossRef]
105. Taranya, E.; Annie, A.X.; Sundaresan, K.; Pradhakshya, J. Hand gesture recognition for sign languages using 3DCNN for efficient detection. In *Proceedings of the Computer Methods, Imaging and Visualization in Biomechanics and Biomedical Engineering: Proceedings International Symposium, Online, 7–9 September 2021*; Springer: Cham, Switzerland, 2021; pp. 215–233. [CrossRef]
106. Li, X.; Zhao, Q.; Song, S.; Shen, T. Sign language recognition based on residual network. In *Proceedings of the International Conference on Computer Engineering and Networks, Haikou, China, 4–7 November 2022*; Springer: Cham, Switzerland, 2022; pp. 1240–1249. [CrossRef]
107. Xu, X.; Meng, K.; Chen, C.; Lu, L. Isolated word sign language recognition based on improved SKResNet-TCN network. *J. Sens.* **2023**, *2023*, 9503961. [CrossRef]
108. Marais, M.; Brown, D.; Connan, J.; Boby, A. Spatiotemporal convolutions and video vision transformers for signer-independent sign language recognition. In Proceedings of the International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 3–4 August 2023; pp. 1–6. [CrossRef]
109. Podder, K.K.; Ezeddin, M.; Chowdhury, M.E.H.; Sumon, S.I.; Tahir, A.M.; Ayari, M.A.; Dutta, P.; Khandakar, A.; Bin Mahbub, Z.; Kadir, M.A. Signer-independent Arabic sign language recognition system using deep learning model. *Sensors* **2023**, *23*, 7156. [CrossRef]
110. Chen, X.; Su, L.; Zhao, J.; Qiu, K.; Jiang, N.; Zhai, G. Sign language gesture recognition and classification based on event camera with spiking neural networks. *Electronics* **2023**, *12*, 4786. [CrossRef]
111. Vyavahare, P.; Dhawale, S.; Takale, P.; Koli, V.; Kanawade, B.; Khonde, S. Detection and interpretation of Indian sign language using LSTM networks. *J. Intell. Syst. Control* **2023**, *2*, 132–142. [CrossRef]
112. Huang, J.; Chouvatut, V. Video-based sign language recognition via ResNet and LSTM network. *J. Imaging* **2024**, *10*, 149. [CrossRef]
113. Kim, T.; Kim, B. Techniques for detecting the start and end points of sign language utterances to enhance recognition performance in mobile environments. *Appl. Sci.* **2024**, *14*, 9199. [CrossRef]
114. Ruben, R.M.; Kambli, P.; Kumar, S. Sign language dynamic gesture recognition system leveraging deep learning and computer vision. In Proceedings of the IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 16–17 October 2022; pp. 1–9. [CrossRef]
115. Boukdir, A.; Benaddy, M.; Meslouhi, O.; Kardouchi, M.; Akhloufi, M. Moroccan sign language video recognition with deep learning. In *Proceedings of the 7th International Congress on Information and Communication Technology (ICICT), London, UK, 21–24 February 2022*; Springer: Cham, Switzerland, 2022; Volume 1, pp. 415–422. [CrossRef]
116. Vahdani, E.; Jing, L.; Huenerfauth, M.; Tian, Y. Recognizing American sign language manual signs from RGB-D videos. *Preprint* **2022**. [CrossRef]
117. Hilario-Acuapan, G.; Ordaz-Hernández, K.; Castelán, M.; Lopez-Juarez, I. Toward a recognition system for Mexican Sign Language: Arm movement detection. *Sensors* **2025**, *25*, 3636. [CrossRef]
118. Gao, Y.; Zhang, Y.; Jiang, X. An optimized convolutional neural network with combination blocks for Chinese sign language identification. *Comput. Model. Eng. Sci.* **2022**, *132*, 95–117. [CrossRef]
119. Montefalcon, M.; Padilla, J.; Rodriguez, R. Sign language recognition of selected Filipino phrases using LSTM neural network. In *Proceedings of the Seventh International Congress on Information and Communication Technology (ICICT), London, UK, 21–24 February 2022*; Springer: Singapore, 2022; Volume 4, pp. 633–641. [CrossRef]
120. Balaha, M.; El-Kady, S.; Balaha, H.; Salama, M.; Ezzat, E.; Hassan, M.; Saafan, M. A vision-based deep learning approach for independent-users Arabic sign language interpretation. *Multimed. Tools Appl.* **2023**, *82*, 6807–6826. [CrossRef]

121. Borges-Galindo, E.; Morales-Ramírez, N.; González-Lee, M.; García-Martínez, J.; Nakano-Miyatake, M.; Perez-Meana, H. Sign language interpreting system using recursive neural networks. *Appl. Sci.* **2024**, *14*, 8560. [CrossRef]
122. Ghadekar, P.; Dhame, P.; Dixit, S.; Patil, A.; Sanjekar, R.; Shinde, S.; More, P. Partial hand sign recognition and sentence construction for diverse age groups using VGG19 and Seq2Seq. In Proceedings of the 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), Vellore, India, 22–23 February 2024; pp. 1–9. [CrossRef]
123. Wei, C.; Zhao, J.; Zhou, W.; Li, H. Semantic boundary detection with reinforcement learning for continuous sign language recognition. *IEEE Trans. Syst. Video Technol.* **2021**, *31*, 1138–1149. [CrossRef]
124. Hao, A.; Min, Y.; Chen, X. Self-mutual distillation learning for continuous sign language recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 11303–11312. [CrossRef]
125. Min, Y.; Hao, A.; Chai, X.; Chen, X. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 11542–11551. [CrossRef]
126. Wang, J.; Xue, C.; Liu, Y.; Yu, M.; Jia, J. Sign language annotation algorithm based on BGRU-CTC. In *Proceedings of the 2021 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Chengdu, China, 17–19 December 2021*; IEEE: New York, NY, USA, 2021; pp. 1–6. [CrossRef]
127. Zuo, R.; Mak, B. C2SLR: Consistency-enhanced continuous sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5131–5140. [CrossRef]
128. Jin, Y.; Choi, S.; Gao, Y.; Li, J.; Li, Z.; Jin, Z. TransASL: A smart glass based comprehensive ASL recognizer in daily life. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23), Sydney, Australia, 27–31 March 2023; pp. 802–818. [CrossRef]
129. Hu, L.; Gao, L.; Liu, Z.; Feng, W. Self-emphasizing network for continuous sign language recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 854–862. [CrossRef]
130. Hu, L.; Gao, L.; Liu, Z.; Feng, W. Continuous sign language recognition with correlation network. In Proceedings of the IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 2529–2539. [CrossRef]
131. Zhu, Q.; Li, J.; Yuan, F.; Gan, Q. Fully 2D Convolutional Network for Continuous Sign Language Recognition. In *Fuzzy Systems and Data Mining IX*; IOS Press: Amsterdam, The Netherlands; pp. 244–254. [CrossRef]
132. Chen, Y.; Li, J.; Lin, S.; Xu, Y.; Yang, C. A BiLSTM and CTC-based multi-sensor information fusion framework for continuous sign language recognition. In Proceedings of the 10th International Conference on Electrical Engineering, Control and Robotics (EECR), Suzhou, China, 29–31 March 2024; pp. 310–315. [CrossRef]
133. Lu, H.; Salah, A.; Poppe, R. TCNet: Continuous sign language recognition from trajectories and correlated regions. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 3891–3899. [CrossRef]
134. Zuo, R.; Wei, F.; Mak, B. Towards online continuous sign language recognition and translation. *arXiv* **2024**, arXiv:2401.05336. [CrossRef]
135. Hu, H.; Pu, J.; Zhou, W.; Fang, H.; Li, H. Prior-aware cross-modality augmentation learning for continuous sign language recognition. *IEEE Trans. Multimed.* **2024**, *26*, 593–606. [CrossRef]
136. Bahri, I.; Saon, S.; Mahamad, A.; Isa, K.; Fadlilah, U.; Ahmadon, M.; Yamaguchi, S. Interpretation of Bahasa Isyarat Malaysia (BIM) using SSD-MobileNet-V2 FPNLite and COCO mAP. *Information* **2023**, *14*, 319. [CrossRef]
137. Siddique, S.; Islam, S.; Neon, E.; Sabbir, T.; Naheen, I.T.; Khan, R. Deep learning-based Bangla sign language detection with an edge device. *Intell. Syst. Appl.* **2023**, *18*, 200224. [CrossRef]
138. Zhang, M.; Yang, S.; Zhao, M. Deep learning-based standard sign language discrimination. *IEEE Access* **2023**, *11*, 125822–125834. [CrossRef]
139. Kalimuthu, S. Video captioning based on sign language using YOLOv8 model. In *Proceedings of the International Conference on Advances in Computing and Data Sciences, Dubai, United Arab Emirates, 25–27 April 2023*; Springer: Cham, Switzerland, 2023; pp. 221–232. [CrossRef]
140. Begum, N.; Rahman, R.; Jahan, N.; Khan, S.S.; Helaly, T.; Haque, A.; Khatun, N. Borno-Net: A real-time Bengali sign-character detection and sentence generation system using quantized YOLOv4-Tiny and LSTMs. *Appl. Sci.* **2023**, *13*, 5219. [CrossRef]
141. Reddy, P.V.; Reddy, B.V.V. Sign language recognition based on YOLOv5 algorithm for the Telugu sign language. *arXiv* **2024**, arXiv:2406.10231. [CrossRef]
142. Purnomo, H.; Dewi, C.; Kristanto, B.; Hartomo, K.; Hashim, S. Utilizing the YOLOv8 model for accurate hand gesture recognition with complex background. *SSRN Electron. J.* **2024**. [CrossRef]
143. Yin, K.; Read, J. Better sign language translation with STMC-transformer. *arXiv* **2020**, arXiv:2004.00588. [CrossRef]

144. Cao, Y.; Li, W.; Li, X.; Chen, M.; Chen, G.; Hu, L.; Huang, K. Explore more guidance: A task-aware instruction network for sign language translation enhanced with data augmentation. *arXiv* **2022**, arXiv:2204.05953. [CrossRef]
145. Shi, B.; Brentari, D.; Shakhnarovich, G.; Livescu, K. Open-domain sign language translation learned from online video. *arXiv* **2022**, arXiv:2205.12870. [CrossRef]
146. Peng, R.; Zeng, Y.; Zhao, J. Better sign language translation with monolingual data. *arXiv* **2023**, arXiv:2304.10844. [CrossRef]
147. Tarrés, L.; Gállego, G.; Duarte, A.; Torres, J.; Nieto, X.G.-I. Sign language translation from instructional videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 5625–5635. [CrossRef]
148. Nishimura, T.; Abbasi, B. Improving sign language recognition performance using multimodal data. In Proceedings of the IEEE International Conference on Information Reuse and Integration for Data Science (IRI), San Jose, CA, USA, 7–9 August 2024; pp. 184–189. [CrossRef]
149. Pol, A.; Patil, S. Twin residual network for sign language recognition from video. In Proceedings of the International Conference on Automation and Computation (AUTOCOM), Dehradun, India, 14–16 March 2024; pp. 580–585. [CrossRef]
150. Pol, A.; Patil, S. Unlocking cognitive insights: Leveraging transfer learning for caption generation in signed language videos. In Proceedings of the 11th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 14–15 March 2024; pp. 1–6. [CrossRef]
151. Roy, P.; Han, J.; Chouhan, S.; Thumu, B. American sign language video to text translation. *arXiv* **2024**, arXiv:2402.07255. [CrossRef]
152. Zheng, J.; Li, S.; Tan, C.; Wu, C.; Chen, Y.; Li, S. Leveraging graph-based cross-modal information fusion for neural sign language translation. *arXiv* **2022**, arXiv:2211.00526. [CrossRef]
153. Khanh, D. ViSL model: The model automatically generates sentences of Vietnamese sign language. *J. Sci. Technol. Inf. Technol. Mech. Opt.* **2024**, *157*, 779–787. [CrossRef]
154. Ezra, D.; Mastitz, S.; Rabaev, I. Signsability: Enhancing communication through a sign language app. *Software* **2024**, *3*, 368–379. [CrossRef]
155. Martins, G.V.; Quinaz, A.; Viegas, C.; Cavaco, S.; Magalhães, J. SLVideo: A sign language video moment retrieval framework. *arXiv* **2024**, arXiv:2407.15668. [CrossRef]
156. Qin, J.; Wang, M. Sign language recognition based on dual-channel star-attention convolutional neural network. *Sci. Rep.* **2025**, *15*, 27685. [CrossRef]
157. Coster, M.D.; Herreweghe, M.V.; Dambre, J. Isolated sign recognition from RGB video using pose flow and self-attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 3441–3450. [CrossRef]
158. Natarajan, B.; Ramasamy, E.; Prasad, M.L. Sentence2SignGesture: A hybrid neural machine translation network for sign language video generation. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 9807–9821. [CrossRef]
159. Yerimbetova, A.; Kaidina, D.; Sakenov, B.; Daiyrbayeva, E.; Turdalyuly, M.; Berzhanova, U. Recognising Kazakh Sign Language with Mediapipe. In Proceedings of the 2024 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Türkiye, 26–28 October 2024; pp. 828–833. [CrossRef]
160. Kumari, D.; Anand, R. Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism. *Electronics* **2024**, *13*, 1229. [CrossRef]
161. Varol, G.; Momeni, L.; Albanie, S.; Afouras, T.; Zisserman, A. Read and Attend: Temporal Localisation in Sign Language Videos. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16852–16861. [CrossRef]
162. Chen, S.; Situ, J.; Cheng, H.; Su, S.; Kirst, D.; Ming, L.; Huang, Y. Inclusive emotion technologies: Addressing the needs of d/Deaf and hard of hearing learners in video-based learning. *arXiv* **2024**, arXiv:2410.00199. [CrossRef]
163. Anand, K.; Urolagin, S.; Mishra, R.K. How does hand gestures in videos impact social media engagement—Insights based on deep learning. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100036. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.