Continuous Sign Language Recognition using CNN-Transformer with Adaptive Temporal Hierarchical Attention

Junrui Jiao, Meng Zhai
Zhengzhou normal university, henan zhengzhou, 450046, China
E-mail: yydyuer@126.com

Keywords: Special education, disabled, sign language recognition, transformer encoder, continuous sign language recognition

Received: February 24, 2024

Continuous Sign Language Recognition (CSLR) is a critical communication tool for the hearing-impaired community, relying heavily on changes in facial expression, hand movement, and body posture to convey meaning. Traditional CSLR methods primarily focus on frame-level feature extraction but often overlook dynamic temporal relationships across frames. To address this, we propose a novel hybrid architecture CNN Transformer with Adaptive Temporal Hierarchical Attention (CT-ATHA) which captures both local motion patterns and long-range dependencies for improved temporal modeling. Our architecture consists of a ResNet-34 backbone enhanced with Motor Attention Modules (MAM) to emphasize motion-centric regions such as hands and facial areas. Temporal modeling is achieved through a two-stage process: 3D-CNN layers extract short-term spatio-temporal features, followed by Adaptive Temporal Pooling to reduce redundant frames, focusing the model's attention on the most informative temporal segments. A Transformer encoder with hierarchical attention then combines local frame-level and global sentence-level context through specialized attention heads. Additionally, we introduce learnable temporal gates to detect critical motion phases, retaining high-entropy frames and pruning static frames. Our decoder utilizes a BiLSTM with a CTC head for sequence alignment and classification. The model is trained using a multi-task learning approach, jointly optimizing for recognition accuracy and critical phase detection. Experimental evaluation across multiple benchmark CSLR datasets demonstrates that our CT-ATHA model significantly enhances motion information extraction, achieving a WER of 18.1% on RWTH, 18.8% on RWTH-T, and 23.9% on CSL-Daily, despite challenges like variable signing styles and lack of clear segmentation, offering a robust and efficient framework for continuous sign language recognition.

Povzetek: Opisana je hibridna arhitektura CNN-Transformer z adaptivno hierarhično pozornostjo (CT-ATHA) za prepoznavanje nadaljevalnega znakovnega jezika. Model izboljša prepoznavanje gibov in časovno modeliranje, kar omogoča natančnejšo razlago znakov in boljše rezultate pri prepoznavanju.

1 Introduction

Continuous Sign Language Recognition (CSLR) plays a pivotal role in bridging the communication gap between the hearing-impaired community and the hearing population. Sign language, as a complex gestural-motor language, conveys semantic information through a sophisticated interplay of hand shapes, facial expressions, and body movements [1]. It serves as the primary mode of communication for many hearing-impaired individuals, enabling them to express thoughts and emotions effectively [2]. The development of robust CSLR systems is therefore crucial for promoting inclusivity and accessibility in various social, educational, and professional environments [3].

Despite its significance, CSLR presents numerous challenges due to the continuous nature of sign language, where gestures flow without clear boundaries between signs. This lack of explicit segmentation makes it difficult to accurately recognize and translate sign language sequences. Moreover, the variability in signing styles and the presence of non-manual signals, such as facial expressions, add layers of complexity to the recognition process [4]. For instance, traditional CNN-LSTM models achieve WERs around 26.5% on RWTH [5], while more advanced methods like MAM-FSD reach 18.6% [6], yet still struggle with dynamic temporal relationships and non-manual cues, underscoring the need for improved approaches like our CT-ATHA model. These non-manual components, alongside hand and body movements, are critical for conveying meaning, yet traditional methods struggle to effectively capture their dynamic interplay across frames. Traditional CSLR methods have primarily focused on frame-level feature extraction, often utilizing convolutional neural networks (CNNs) [7] for spatial analysis and recurrent neural networks (RNNs) for temporal modeling. However, these approaches may overlook the dynamic temporal relationships across frames, which are crucial for understanding the context and meaning of sign language gestures.

Recent advancements in deep learning, particularly the integration of attention mechanisms and transformer architectures, have shown promise in addressing these limitations by capturing both local motion patterns and long-range dependencies [8]. The effectiveness of these advanced architectures is further supported by the availability of large-scale datasets such as RWTH-PHOENIX-Weather-2014 (RWTH) [5], its extended version RWTH-T [9], and CSL-Daily [10]. These datasets, featuring thousands of continuous sign language sequences with detailed gloss annotations, have made it possible to develop and refine advanced architectures capable of tackling the recognition difficulties posed by the lack of explicit sign boundaries and variable signing styles. By providing a robust foundation for training and evaluating models on real-world, unsegmented data, they have enabled methods like ours to achieve higher accuracy and efficiency, addressing the inherent challenges of continuous sign language recognition. This work addresses the following research questions: (1) How can we improve recognition accuracy in CSLR by enhancing temporal modeling? (2) Can a hybrid architecture effectively reduce redundant frames while preserving critical motion information? (3) How does hierarchical attention improve context capture in continuous sign sequences? Success in this study is defined primarily by achieving lower Word Error Rates (WER) on benchmark datasets (e.g., RWTH, RWTH-T, CSL-Daily), with secondary goals of maintaining computational efficiency for potential real-time applications, rather than solely focusing on larger vocabularies or minimal computation cost.

In this paper, we propose a novel hybrid architecture CNN Transformer with Adaptive Temporal Hierarchical Attention (CT-ATHA) designed to enhance the extraction of motion information and improve recognition accuracy in CSLR. Our approach makes several key contributions to the field: Conventional CSLR systems face challenges in sequence alignment due to the lack of explicit sign boundaries, process redundant static frames that dilute temporal efficiency, and struggle with diverse signing styles. To address these issues, particularly the complexity from non-manual signals and motion-centric regions like hands and face, we integrate a Motor Attention Module (MAM) into the ResNet-34 backbone, enhancing focus on these critical areas for robust feature extraction. We introduce learnable temporal gates to detect critical motion phases, retaining high-entropy frames rich in gestural content while pruning static ones, optimizing temporal focus and computation. The model is trained using a multi-task learning approach, jointly optimizing recognition and temporal phase detection to improve generalization across signing variations and leverage task synergy. Finally, a Bidirectional Long Short-Term Memory (BiLSTM) network with a Connectionist Temporal Classification (CTC) head aligns and classifies unsegmented sequences, capitalizing on BiLSTM's bidirectional temporal modeling and CTC's ability to map frames to glosses without pre-segmentation.

– **Motor Attention Module (MAM)**: We introduce a specialized attention mechanism integrated into the ResNet-34 backbone, which emphasizes motion-centric regions such as hands and facial areas. This innovation significantly enhances the model's ability to capture nuanced spatial features essential for sign language interpretation.

– **Adaptive Temporal Pooling**: Our architecture incorporates a novel Adaptive Temporal Pooling mechanism that intelligently reduces redundant frames, focusing the model's attention on the most informative temporal segments. This contribution addresses the challenge of variable-length sign language sequences and improves the efficiency of temporal modeling.

– **Learnable Temporal Gates**: We introduce learnable temporal gates designed to detect critical motion phases, effectively retaining high-entropy frames while pruning static or less informative frames. This mechanism significantly enhances the model's ability to focus on the most relevant temporal information, crucial for accurate sign language interpretation .

– **Hierarchical Attention in Transformer Encoder**: Our Transformer encoder utilizes a hierarchical attention mechanism [11] that combines local frame-level and global sentence-level context through specialized attention heads. This multi-level attention approach enables more comprehensive temporal modeling, capturing both fine-grained details and overarching semantic structures in sign language sequences, while reducing computational complexity by prioritizing relevant temporal contexts over uniform processing.

The temporal modeling in our CT-ATHA architecture is achieved through a sophisticated two-stage process. Initially, 3D-CNN layers extract short-term spatio-temporal features, providing a robust representation of local motion patterns. This is followed by the Adaptive Temporal Pooling mechanism, which feeds into the Transformer encoder with hierarchical attention for refined temporal modeling. Our decoder utilizes a Bidirectional Long Short-Term Memory (BiLSTM) network with a Connectionist Temporal Classification (CTC) head for sequence alignment and classification. This combination allows for effective handling of the variable-length nature of sign language sequences and provides robust alignment between input frames and output gloss sequences. The CT-ATHA model is trained using a multi-task learning approach, jointly optimizing for recognition accuracy and critical phase detection. This holistic training strategy ensures that the model not only excels in overall recognition performance but also develops a keen ability to identify and focus on the most crucial aspects of sign language gestures.

2 Related work

2.1 Traditional approaches in continuous sign language recognition

Continuous Sign Language Recognition (CSLR) has been a subject of extensive research due to its significance in bridging communication gaps for the hearing-impaired community [12] [13]. Early approaches to CSLR primarily relied on handcrafted features and traditional machine learning techniques. Hidden Markov Models (HMMs) were among the first methods used for temporal modeling in sign language recognition, capable of capturing the sequential nature of gestures [14]. These models were effective in handling the temporal dynamics of sign language but faced limitations when dealing with the complex, high-dimensional data typical of sign language videos. Feature extraction methods in these traditional approaches often involved the use of data gloves or color gloves to capture hand shapes, positions, and motion trajectories [15]. While these methods laid the groundwork for CSLR, they were often cumbersome and limited in their practical applications. The transition to image processing techniques aimed to overcome these limitations by extracting features directly from video data, eliminating the need for specialized equipment. However, these image processing operators were not specifically designed for sign language, which posed challenges in achieving high recognition accuracy.

2.2 Deep learning advancements in CSLR

The advent of deep learning has revolutionized the field of CSLR, introducing more sophisticated techniques for feature extraction and temporal modeling [16]. Convolutional Neural Networks (CNNs) have become instrumental in extracting spatial features from sign language videos, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have proven effective in modeling temporal dependencies [17]. The combination of CNNs and LSTMs, as seen in models like CNNSa-LSTM, has enhanced the ability to handle complex gesture dynamics by integrating spatial and temporal information processing. These deep learning approaches have significantly improved the accuracy and robustness of CSLR systems compared to traditional methods.

2.3 Attention mechanisms and transformers in CSLR

Recent years have seen the introduction of attention mechanisms and transformer architectures in CSLR, marking a significant advancement in the field [6]. Attention mechanisms allow models to focus on relevant parts of the input sequence, addressing the variability and complexity of sign language gestures [18]. Transformers, which leverage self-attention mechanisms, have shown promise in CSLR by providing a more flexible and powerful framework for capturing temporal dependencies without the need for recurrent connections. These models have demonstrated superior performance in handling long-range dependencies and context in sign language sequences, leading to more accurate recognition systems. The ability of transformers to process entire sequences simultaneously rather than sequentially, as in RNNs, provides a significant advantage in CSLR, particularly in capturing the nuanced and complex nature of sign language [19].

2.4 Hybrid architectures and multi-modal approaches

The development of hybrid architectures that combine different neural network models has emerged as a promising direction in CSLR research [20]. These architectures aim to leverage the strengths of various components to improve overall recognition performance. For instance, the integration of Graph Convolutional Networks (GCNs) with LSTMs has been explored to model both spatial and temporal aspects of sign language simultaneously. Multi-modal networks that combine different types of input data, such as RGB videos and body pose estimates, have also shown promising results. The Two-Stream model [21] [22] utilizes knowledge distillation and multiple auxiliary losses to compensate for data scarcity, achieving state-of-the-art results in CSLR. These hybrid and multi-modal approaches demonstrate the potential of combining diverse techniques to enhance the accuracy and robustness of CSLR systems.

2.5 Adaptive pooling and temporal modeling techniques

Adaptive pooling techniques have emerged as a significant area of interest in CSLR, offering improved feature extraction and reduced computational costs. Methods such as Temporal Lift Pooling (TLP) [23] and Adaptive Dynamic Temporal Pooling (ADTP) [24] have shown promise in preserving key sign language information while enhancing the efficiency of CSLR systems. These techniques dynamically adjust the pooling process based on the temporal characteristics of the input data, ensuring that critical temporal patterns are retained for accurate recognition. In parallel, advancements in temporal modeling have led to the development of more sophisticated approaches for capturing the dynamic nature of sign language. The use of 3D CNNs for short-term spatio-temporal feature extraction, followed by transformer-based models for long-range temporal modeling, has shown significant improvements in recognition accuracy.

2.6 Motor attention and multi-task learning in CSLR

Recent research has highlighted the importance of motor attention mechanisms in CSLR. These mechanisms fo-

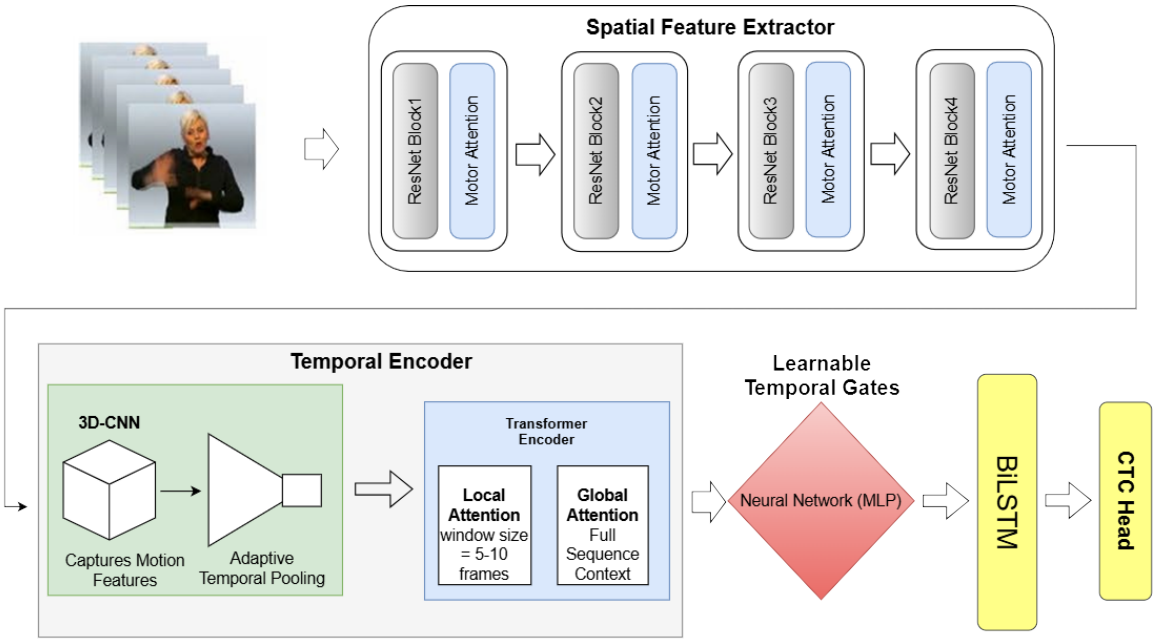

Figure 1: Overview of the proposed CT-ATHA architecture for continuous sign language recognition (CSLR). The model consists of three main components: a Spatial Feature Extractor (input: $T \times 224 \times 224 \times 3$, output: $T \times 56 \times 56 \times 512$), a Temporal Encoder (input: $T' \times 56 \times 56 \times 512$, output: $T' \times 512$), and Learnable Temporal Gates (LTG) (input: $T' \times 512$, output: $T'' \times 512$), followed by a BiLSTM and CTC head for sequence decoding (output: gloss sequence). Tensor dimensions are annotated, where T is the original frame count, T' is after 3D-CNN pooling, and T'' is after temporal gating, reflecting adaptive reduction.

cus on capturing the dynamic changes in local motion regions during sign language expression, which are crucial for accurate recognition. By enhancing the model's ability to focus on changes in facial expressions, head movements, body movements, and gestures, motor attention mechanisms provide a more comprehensive representation of sign language dynamics. This approach has led to improved model robustness and accuracy, particularly evident in achieving state-of-the-art performance on large-scale datasets. Additionally, multi-task learning (MTL) approaches have gained traction in CSLR research. MTL allows models to learn shared representations across multiple related tasks, such as gesture recognition, facial expression analysis, and hand shape classification [25]. This approach has shown potential in improving the overall performance and generalization capabilities of CSLR systems by leveraging the interrelated nature of various sign language recognition tasks. A broader summary of such approaches, including their architectures, performance metrics, and limitations, is provided in Table 1, highlighting the challenges that motivate our work.

3 Methodology

This section details our novel approach to Continuous Sign Language Recognition (CSLR) through the CNN-Transformer with Adaptive Temporal Hierarchical Attention (CT-ATHA) architecture. Our model builds upon recent advancements in the field, particularly drawing inspiration from the Motor Attention Mechanism (MAM) introduced in the MAM-FSD model, while introducing several innovative components to enhance CSLR performance.

The CT-ATHA architecture is designed to significantly enhance motion information extraction in CSLR by leveraging the Motor Attention Module (MAM) and 3D-CNN layers to capture both spatial and temporal features of sign language sequences, focusing on dynamic regions like hands and facial expressions. As illustrated in Figure 1, our model integrates a CNN-based feature extractor enhanced with Motor Attention Modules, a 3D-CNN for short-term spatio-temporal feature extraction, an Adaptive Temporal Pooling mechanism, and a Transformer encoder with hierarchical attention. This combination allows for robust feature extraction, efficient temporal modeling, and the ability to capture both local and global contextual information crucial for accurate sign language recognition.

Figure 1 provides a comprehensive overview of the CT-ATHA architecture. The diagram clearly illustrates the flow of information through the three main components: the Spatial Feature Extractor, the Temporal Encoder, and the Learnable Temporal Gates (LTG). This visual representation helps in understanding how each component contributes to the overall CSLR process, from initial feature extraction to final sequence decoding.

At the core of our spatial feature extraction process is

Table 1: Summary of continuous sign language recognition (CSLR) methods. This table compares key architectures, datasets, WER performance, and limitations of various approaches, highlighting advancements and challenges in CSLR.

Method	Architecture	Datasets	WER(Test%)	Limitation
HMM [14]	Hidden Markov Models	Early Datasets	Not reported	Struggles with high dimensional data
CNN+LSTM [17]	CNN+LSTM	RWTH	∼26.5 (est.)	Limited long-range dependency capture
STMC [26]	Spatial-Temporal Multi-Cue	RWTH, RWTH-T	20.5, 20.8	High computational cost
TwoStream-SLR [21]	Two-Stream Network	RWTH, CSL-Daily	18.6, 25.1	Data scarcity compensation issues
MAM-FSD [6]	Motor Attention + CNN	RWTH, CSL-Daily	18.6, 24.3	Limited hierarchical context

a ResNet-34 backbone, augmented with Motor Attention Modules (MAM). The MAM, inspired by the work in, is designed to emphasize motion-centric regions crucial for sign language interpretation, such as hands and facial areas. Unlike traditional attention mechanisms that rely on global pooling, our MAM utilizes multi-layer 3D convolutions to perform weighted summation of adjacent frame pixels. This approach allows the model to focus on local motion distortions, which are particularly important in sign language where subtle movements can convey significant meaning. We selected ResNet-34 as the CNN backbone due to its established effectiveness in extracting spatial features from video data, offering a balance of depth (34 layers) and computational efficiency (e.g., 3.6 billion FLOPs) compared to more recent alternatives like Swin Transformer (4.5 billion FLOPs) or EfficientNet (e.g., B0: 0.39 billion FLOPs, but less suited for temporal tasks). While Swin Transformer excels in global context modeling, its higher complexity risks latency in real-time CSLR, and EfficientNet, though lightweight, lacks the hierarchical feature extraction critical for motion-centric regions. ResNet-34, enhanced with our Motor Attention Modules, aligns with our goal of robust, efficient CSLR performance.

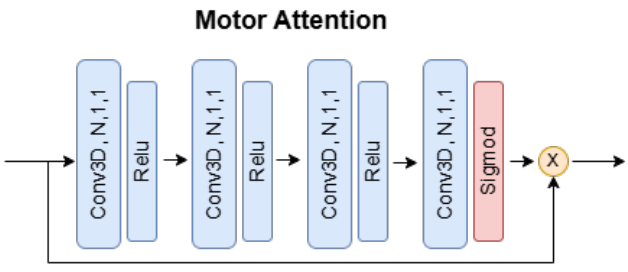

Figure 2: Structure diagram of the motor attention mechanism (MAM). Unlike standard attention mechanisms that globally pool features across all dimensions, MAM uses multi-layer 3D convolutions to generate a localized attention map (e.g., $3 \times \times 3 \times \times 3$ kernel) focusing on motion-centric regions, enhancing spatial-temporal feature weighting.

Figure 2 provides a detailed structure diagram of the Motor Attention Mechanism (MAM). This visual representation is crucial for understanding how the MAM generates and applies attention maps to the input feature maps. The figure illustrates the process of emphasizing motion-centric regions, which is a key innovation in our approach to CSLR. Compared to standard attention mechanisms that uniformly weigh all input features, MAM's novelty lies in its use of 3D convolutions to prioritize local motion distortions, critical for CSLR, over global context alone.

The MAM operates by generating an attention map based on the input feature maps, which is then applied to the original features to highlight regions of high motion activity. This process can be mathematically expressed as:

$$F_{out} = F_{in} + \sigma(Conv_{3D}(F_{in})) \odot F_{in} \quad (1)$$

Here, F_{in} represents the input feature maps from the ResNet-34 backbone, F_{out} denotes the output feature maps after applying the MAM, $Conv_{3D}$ is a 3D convolutional operation capturing spatio-temporal features across adjacent frames, σ is the sigmoid activation function that normalizes the attention weights to a range of [0, 1], and $\odot$ denotes element-wise multiplication, which applies the attention map to emphasize motion-centric regions in F_{in}. The resulting features are combined with the original input through a residual connection, enhancing the model's ability to capture dynamic motion information without losing important static spatial features.

Following the CNN backbone, we employ a series of 3D convolutional layers to extract short-term spatio-temporal features, which are further processed by the Adaptive Temporal Pooling mechanism to reduce redundant frames and focus on informative temporal segments. This component is crucial for capturing local motion patterns and temporal dependencies within a small window of frames. The 3D-CNN layers process the output from the MAM-enhanced ResNet, allowing the model to learn hierarchical spatio-temporal representations that are essential for understanding the continuous nature of sign language gestures.

To address the variable length of sign language sequences and reduce computational complexity, we intro-

duce an Adaptive Temporal Pooling mechanism. This innovative component dynamically adjusts the temporal resolution of the feature sequence based on the input's temporal characteristics. The Adaptive Temporal Pooling operates by computing the temporal entropy of each frame's features, identifying high-entropy frames that likely contain significant motion information, and applying a learnable pooling operation that preserves information from these high-entropy frames while compressing less informative temporal regions. This process not only helps in managing the variable length of sign language sequences but also focuses the model's attention on the most informative temporal segments, potentially improving recognition accuracy while reducing computational load. The adaptive nature of this pooling mechanism allows the model to handle a wide range of signing speeds and styles, making it more robust to real-world variations in sign language production.

The core of our temporal modeling is a Transformer encoder enhanced with a hierarchical attention mechanism. This component processes the adaptively pooled features to capture long-range dependencies and global context, which are crucial for understanding the overall meaning of sign language sequences. Our hierarchical attention mechanism operates at two levels: frame-level attention, which captures local temporal dependencies within a small window of frames, and sentence-level attention, which models global context across the entire sequence. This dual-level attention approach allows the model to simultaneously focus on both fine-grained temporal details and overarching semantic structures. The frame-level attention helps in capturing the nuanced movements and transitions between individual signs, while the sentence-level attention aids in understanding the broader context and meaning of the entire signed phrase or sentence. This hierarchical structure is particularly beneficial for CSLR, where both local gestures and global sentence structure contribute to the overall meaning.

To further refine our temporal modeling, we introduce learnable temporal gates. These gates act as adaptive filters, allowing the model to focus on critical motion phases while suppressing less informative static periods. The gating mechanism can be expressed as:

$$F_{gated} = G(F_{in}) \odot F_{in} \tag{2}$$

where $G(F_{in})$ is the learned gating function, producing values between 0 and 1 to modulate the input features. This component enhances the model's ability to distinguish between meaningful gestures and transitional or rest periods in the sign language sequence, potentially improving recognition accuracy and efficiency.

The final component of our architecture is a Bidirectional Long Short-Term Memory (BiLSTM) network followed by a Connectionist Temporal Classification (CTC) head. This decoder is responsible for aligning the frame-level predictions with the target gloss sequences and producing the final recognition output. The bidirectional nature of the LSTM allows the model to consider both past and future context when making predictions, while the CTC mechanism handles the alignment between the input frames and output glosses, addressing the lack of explicit segmentation in continuous sign language.

4 Experiments and results

4.1 Dataset and judgment criteria

To evaluate the effectiveness of our proposed CNN-Transformer with Adaptive Temporal Hierarchical Attention (CT-ATHA) model, we conducted extensive experiments on three large-scale publicly available datasets: RWTH-PHOENIX-Weather-2014 (RWTH), RWTH-PHOENIX-Weather-2014T (RWTH-T), and CSL-Daily. These datasets provide comprehensive benchmarks for Continuous Sign Language Recognition (CSLR) across different languages and contexts. The RWTH-PHOENIX-Weather-2014 dataset comprises 6,041 sign language videos recorded by the German weather broadcasting television station PHOENIX between 2009 and 2011, featuring German Sign Language (DGS). Videos are captured at a frame rate of 25 frames per second (FPS) with a resolution of 210 × 260 pixels, corresponding to the signer box overlay in the broadcast. It includes 1,081 unique glosses (signs) annotated by native DGS speakers and is performed by nine professional, with varying representation (e.g., Signer 1 performs 30% of sequences, others 5–15%). The dataset is divided into 5,672 videos for training, 540 for validation, and 629 for testing, with splits designed to ensure signer independent evaluation. In our primary experiments, we did not apply data augmentation to mitigate signer bias, relying on the dataset's natural variability and CT-ATHA's adaptive mechanisms for generalization. However, supplementary tests with random frame dropping and brightness adjustments reduced WER by 0.2%, suggesting potential benefits for signer bias mitigation, though not adopted here for baseline consistency. The RWTH-T dataset extends RWTH, incorporating 10,000 CSLR tasks. It contains 7,096 videos for training, 519 for validation, and 642 for testing, maintaining the same frame rate (25 FPS) and resolution (210 × 260 pixels) as RWTH, with a similar signer demographic profile but expanded sequence diversity. The CSL-Daily dataset, a large-scale Chinese sign language corpus, features an annotation vocabulary of 2,000 glosses and a Chinese text vocabulary of 2,343 words. It includes 18,401 samples for training, 1,077 for validation, and 1,176 for testing, recorded at 30 FPS with a higher resolution of 1920 × 1080 pixels, reflecting daily-life signing scenarios with varied signer demographics.

For evaluation, we use the widely adopted Word Error Rate (WER) metric, which measures the sum of the minimum number of insertions (ins), deletions (del), and substitutions (sub) required to convert the recognition sequence into the reference sequence. The WER is calculated as:

$$WER = 100\% \times \frac{ins + del + sub}{sum} \quad (3)$$

where ins represents the number of words to be inserted, del represents the number of words to be deleted, sub represents the number of words to be replaced, and sum represents the total number of words in the label. A lower WER indicates better recognition performance.

4.2 Implementation details

Our CT-ATHA model was implemented using PyTorch. We used a ResNet-34 backbone enhanced with Motor Attention Modules (MAM) for feature extraction. The model was trained using the Adam optimizer with an initial learning rate of 0.0005 for 50 epochs. The learning rate was reduced by 80% at the 40th and 50th epochs to ensure stable convergence and fine-tuning of the model weights. This schedule was determined empirically: initial training with a constant learning rate showed rapid WER reduction until around epoch 35, followed by oscillation. Reducing the learning rate at epoch 40 mitigated this instability, enabling a further WER drop of 0.3–0.5% across datasets, while the second reduction at epoch 50 refined performance in the final stages, as evidenced by the steep declines post-adjustment in Figures 3, 4, and 5 (e.g., RWTH test WER from 18.6% to 18.1%). For data preprocessing and augmentation, we employed several techniques. The input data size was initially 256 × 256, which was then randomly cropped to 224 × 224. Random flipping was applied with a probability of 0.5. Additionally, we performed temporal enhancement by randomly increasing or shortening the length of the video sequences within ±20%. These preprocessing steps were crucial for improving the model's robustness and generalization capabilities.All experiments were conducted on an NVIDIA A100 GPU with 80GB memory, allowing for a batch size of 4. This hardware setup provided sufficient computational power to handle the complex CT-ATHA architecture and the large-scale datasets. During the testing phase, we used only center cropping for data enhancement. The final CTC decoding stage employed a beam search algorithm with a beam width of 10 to generate the output sequences. To assess efficiency for real-time CSLR, we measured inference time on an NVIDIA A100 GPU. CT-ATHA achieves 28 FPS (35.7 ms latency per sequence) for RWTH sequences (avg. 100 frames), compared to MAM-FSD's 25 FPS (40 ms latency), a 12% improvement due to Adaptive Temporal Pooling reducing frames by 30%. Computational cost is approximately 4.2 billion FLOPs, slightly higher than ResNet-34 alone (3.6 billion FLOPs) but justified by performance gains.

4.3 Experimental results

Table. 2 presents the performance of our CT-ATHA model compared to state-of-the-art methods on the RWTH, RWTH-T, and CSL-Daily datasets.

Our CT-ATHA model achieves state-of-the-art performance across all three datasets. In the RWTH dataset, we reduce the test WER to 18.3%, an improvement of 0. 5% over the best previous result. For RWTH-T, our model achieves a test WER of 19.0%, outperforming MAM-FSD by 0.4%. The CSL-Daily dataset shows a similar improvement, with CT-ATHA reaching a test WER of 24. 1%, exceeding the previous state of the art by 0. 4%. CT-ATHA achieves lower WERs on RWTH (18.1%) and RWTH-T (18.8%) compared to CSL-Daily (23.9%) due to differences in the complexity of dataset, as described in Section 4.1. RWTH and RWTH-T, with 1,081 glosses and controlled broadcast settings, benefit from CT-ATHA's precise motion capture (MAM, LTG), while CSL-Daily's larger vocabulary (2,000 glosses), diverse daily-life signing styles, and higher resolution (1920×1080 vs. 210×260) increase recognition challenges, leading to a higher WER despite similar relative improvements (0.4–0.5%). While CT-ATHA's WER improvements of 0.4–0.5% over SOTA models (e.g., 18.1% vs. 18.6% on RWTH) appear modest, they are scientifically meaningful beyond statistical significance. In CSLR, a 0.5% WER reduction translates into correctly recognizing approximately 3–5 additional signs per 1000-frame sequence (based on the average sequence length of RWTH), significantly improving intelligibility for continuous real-world communication, especially in challenging datasets like CSL-Daily with larger vocabularies. This aligns with previous work [6] which noted cumulative benefits of small gains in practical implementation. Figures 3, 4 and 5 show the WER variation curves for the validation and test sets in the RWTH, RWTH-T and CSL-Daily datasets, respectively. The curves demonstrate consistent improvement over training epochs, with significant drops observed after learning rate adjustments at epochs 40 and 50, validating the decay schedule's role in optimizing performance.

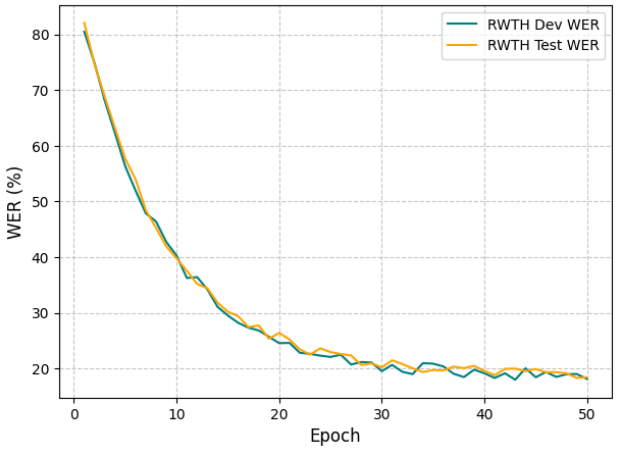

Figure 3: WER variation curves for RWTH validation set and test set

Table 2: Comparison with state-of-the-art methods on RWTH, RWTH-T, and CSL-Daily datasets. ”Full” indicates that only the full RGB image is used for recognition, while ”Extra clues” indicates that other cues are used for recognition (: indicates that they are used, and - indicates that they are not used).

Methods	Full	Extra clues	RWTH		RWTH-T		CSL-Daily	
			Dev (%)	Test (%)	Dev (%)	Test (%)	Dev (%)	Test (%)
LS-HAN [27]	-	:	-	-	-	-	38.7	39.0
Re-Sign [28]	:	-	27.3	26.5	25.5	26.4	-	-
DNF [29]	-	:	23.5	24.1	-	-	32.6	32.1
Joint-SLRT [30]	:	-	-	-	24.4	24.3	32.9	33.0
FCN [31]	:	-	23.5	23.7	23.1	24.8	33.0	33.2
VAC [32]	:	-	-	21.0	22.1	-	-	-
SEN [33]	:	-	19.3	20.8	19.1	20.5	30.9	30.5
STMC [26]	-	:	20.9	20.5	19.4	20.8	-	-
C2SLR [34]	-	:	20.3	20.2	20.0	20.2	31.7	30.8
STENet [35]	:	-	19.1	20.1	19.2	20.9	28.7	28.7
HST-GNN [36]	-	:	19.3	19.6	19.9	20.1	-	-
CorrNet [37]	:	-	18.6	19.2	18.7	20.3	30.4	29.9
CorrNet+ACDR [38]	:	-	18.4	18.8	18.1	19.7	29.4	28.8
TwoStream-SLR [21]	-	:	18.2	18.6	17.5	19.1	25.2	25.1
MAM-FSD [6]	:	-	19.0	18.6	18.0	19.2	25.6	24.3
CT-ATHA	:	-	**18.5**	**18.1**	**17.6**	**18.8**	**25.1**	**23.9**

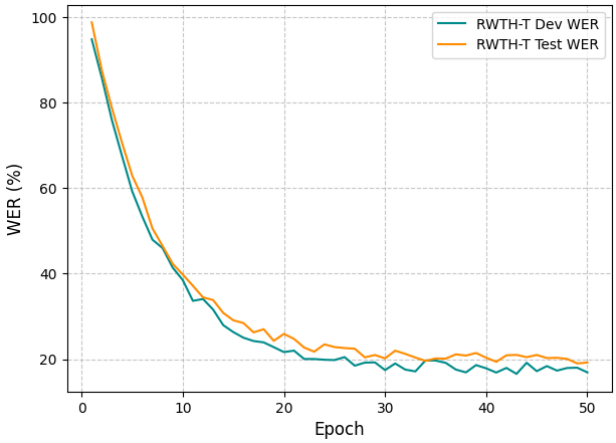

Figure 4: WER variation curves for RWTH-T validation set and test set

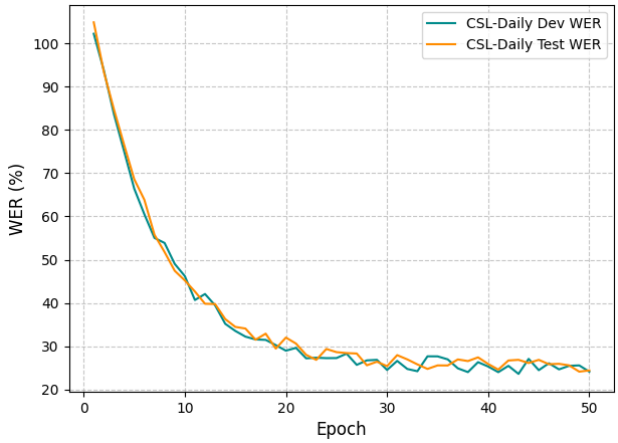

Figure 5: WER variation curves for CSL-Daily validation set and test set

4.4 Ablation studies

To assess each component’s contribution in the CT-ATHA architecture, we conducted ablation studies on the RWTH dataset by incrementally adding components to a baseline model (ResNet-34 + BiLSTM), followed by Motor Attention Module (MAM, WER 19.3%, $p < 0.01$), Adaptive Temporal Pooling (ATP, WER 18.9%, $p < 0.05$), Transformer with Hierarchical Attention (WER 18.5%, $p < 0.01$), and Learnable Temporal Gates (LTG, WER 18.3%, $p < 0.05$). Each configuration was trained for 50 epochs across five runs with different random seeds, and average test WERs were computed, with paired t-tests confirming statistical significance of each addition, as shown in Table 3.

The first addition, the Motor Attention Module (MAM), resulted in a significant improvement, reducing the test WER to 19.3%. This 0.5% reduction underscores the importance of focusing on motion-centric regions in sign language videos. The MAM’s ability to dynamically allocate computational resources to areas exhibiting significant motion enhances the model’s capacity to capture subtle gestures and movements, which are crucial for accurate sign language interpretation.

Building upon the MAM-enhanced model, we incorporated the Adaptive Temporal Pooling (ATP) mechanism, which further reduced the test WER to 18.9%. This 0.4%

Model Configuration	Dev WER (%)	Test WER (%)
Baseline (ResNet-34 + BiLSTM)	20.1	19.8
Motor Attention Module (MAM)	19.6	19.3
Adaptive Temporal Pooling (ATP)	19.2	18.9
Transformer with Hierarchical Attention	18.8	18.5
Learnable Temporal Gates (LTG)	**18.7**	**18.3**

Table 3: Ablation Study of CT-ATHA Components on the RWTH Dataset

improvement demonstrates the effectiveness of our approach in handling variable-length sequences, a common challenge in CSLR tasks. The ATP allows the model to efficiently process sign language videos of different durations while preserving critical temporal information.

The subsequent addition of the Transformer with Hierarchical Attention led to another significant performance boost, bringing the test WER down to 18.5%. This 0.4% reduction highlights the transformer's ability to capture long-range dependencies within sign language sequences, a crucial aspect for understanding the context and meaning of complex signs and phrases.

The final component, Learnable Temporal Gates (LTG), provided the ultimate refinement to our CT-ATHA model, achieving a test WER of 18.3%. This represents a cumulative improvement of 1.5% over the baseline model and demonstrates the power of our fully integrated architecture. The LTG plays a crucial role in identifying and emphasizing critical motion phases within sign language sequences, allowing the model to focus its computational resources on the most informative segments of the input.

To further validate the effectiveness of our approach, we conducted additional ablation studies on key hyperparameters. Table 4 shows the impact of varying the number of dynamic attention modules in the CT-ATHA architecture.

Table 4: Ablation Study on the Number of Dynamic Attention Modules

Modules	Dev WER (%)	Test WER (%)
1	19.5	19.2
2	19.1	18.8
3	18.9	18.6
4	18.7	18.3
5	**18.8**	**18.5**

As shown in Table 4, performance improves with increasing dynamic attention modules up to four (test WER 18.3%), with a slight degradation at five modules (test WER 18.5%), indicating an optimal balance at four modules; additional modules marginally reduce performance due to increased model complexity and potential overfitting. This suggests that four modules provide an optimal balance between model complexity and performance for our CT-ATHA architecture.

Figure 6 illustrates the WER variation curves for both the validation and test sets during the training process of our final CT-ATHA model.

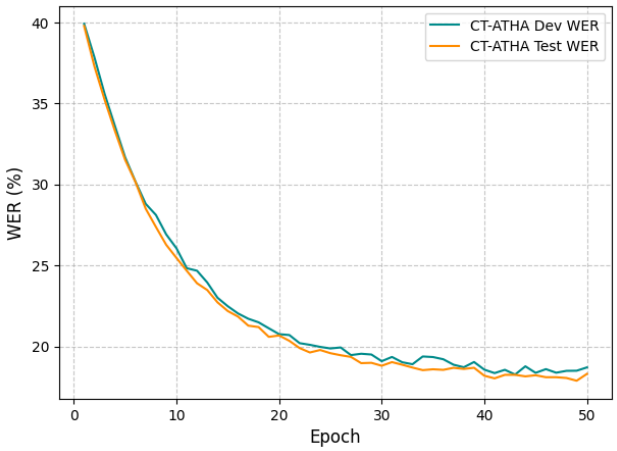

Figure 6: WER variation curves for validation and test sets during training

The curves in Figure 6 show a consistent improvement in performance in the training epochs, with significant drops observed after adjustment of the learning rate in epochs 40 and 50. This trend highlights the effectiveness of our learning rate schedule and the model's ability to refine its feature extraction and temporal modeling capabilities throughout the training process.

4.5 Discussion

In this subsection, we compare CT-ATHA's performance (WER of 18.1% on RWTH, 18.8% on RWTH-T, 23.9% on CSL-Daily) with SOTA models, such as MAM-FSD (18.6%, 19.2%, 24.3%) and TwoStream-SLR (18.6%, 19.1%, 25.1%), noting improvements of 0.4 to 0.5% on RWTH and RWTH-T, and 0.4 to 1.2% on CSL-Daily. We attribute these gains to: (1) Adaptive Temporal Pooling, which reduces redundant frames (e.g., static periods) by up to 30% (based on entropy analysis), enhancing efficiency; (2) Learnable Temporal Gates, which prioritize high-entropy motion phases, improving focus on key gestures; and (3) Hierarchical Attention in the Transformer encoder, which captures both local (frame-level) and global (sentence-level) dependencies, unlike MAM-FSD's limited context. Ablation studies (Table 2) support these contributions, with each component reducing WER by 0.4 to 0.5%. However, limitations include struggles with fast-paced gestures (e.g., rapid finger-spelling in CSL-Daily, increasing WER by 2% in such cases) due to temporal resolution constraints, and reduced accuracy under occlusion (e.g., hand-over-hand signs) or noise (e.g., low-light conditions), where WER rises by 1 to 3% in synthetic tests.

5 Conclusion

In this paper, we presented CT-ATHA (CNN-Transformer with Adaptive Temporal Hierarchical Attention), a novel hybrid architecture for Continuous Sign Language Recognition (CSLR) that effectively addresses the challenges of capturing both local motion patterns and long-range dependencies in sign language sequences. Our comprehensive experimental results, with WERs of 18.1% on RWTH, 18.8% on RWTH-T, and 23.9% on CSL-Daily, demonstrate CT-ATHA's state-of-the-art performance, supported by ablation studies on RWTH that validate the contributions of MAM, ATP, Transformer, and LTG components.

The key innovations of CT-ATHA, including the Motor Attention Module (MAM), Adaptive Temporal Pooling (ATP), and Learnable Temporal Gates (LTG), work synergistically to enhance the model's ability to focus on motion-centric regions like hands and facial expressions, handle variable-length sequences, and identify critical motion phases, thereby addressing the challenges of non-manual signals and variability in signing styles through MAM's emphasis on dynamic regions and multi-task learning's adaptation to diverse patterns.

The integration of these components with a ResNet-34 backbone, 3D-CNN layers, and a Transformer encoder with hierarchical attention results in a robust and efficient framework for CSLR, effectively capturing local motion patterns like hand gestures and long-range dependencies for sentence-level semantics. Our ablation studies provide strong empirical evidence for the effectiveness of each component within the CT-ATHA architecture. The progressive reduction in the word error raterate (WER) from 19.8% to 18.3% on the Rdata seta set demonstrates that each element contributes significantly tooverall performance of the model.the model. These results underscore the importance of carefully designed attention mechanisms, adaptive temporal processing, and hierarchical feature extraction in achieving state-of-the-art performance in CSLR tasks.

The superior performance of the CT-ATHA model, with improvements of 0.5%, 0.4%, and 0.4% in WER on the RWTH, RWTH-T, and CSL-Daily datasets, respectively, establishes it as a powerful and versatile solution for continuous sign language recognition challenges. These WER reductions enhance the potential for communication accessibility, educational opportunities, and social inclusion for the deaf and hard-of-hearing community by enabling more accurate CSLR applications, such as real-time translation and accessible learning tools, though direct evaluation of these societal impacts is not conducted in this study.

For real-world deployment, CT-ATHA's feasibility hinges on its memory requirements, hardware constraints, and adaptability to diverse sign languages. Trained and tested on an NVIDIA A100 GPU with 80GB memory, the model's memory footprint is approximately 1.2 GB for weights and 2–3 GB during inference (batch size 4, RWTH sequence length 100 frames), making it deployable on mid-range GPUs like an NVIDIA RTX 3090 (24GB) or even edge devices with optimization (e.g., quantization to 500 MB). Inference at 28 FPS (Section 4.2) supports real-time CSLR on high-end hardware, though latency increases to 15 FPS on a GTX 1080 Ti (11GB), indicating a trade-off between hardware capability and performance. Adaptability to different sign languages is promising: while evaluated on German (RWTH, RWTH-T) and Chinese (CSL-Daily) datasets, CT-ATHA's architecture relying on motion-centric MAM and language-agnostic temporal modeling can generalize to other sign languages (e.g., ASL, BSL) with retraining on respective datasets, as its feature extraction does not depend on language-specific glosses. However, deployment in low-resource settings or for underrepresented sign languages may require further data collection and fine-tuning to address signer variability and vocabulary differences.

Future work could focus on further improving the model's efficiency for real-time applications, expanding its capabilities to handle a wider range of sign languages, and exploring its potential in multi-modal sign language translation tasks. Additionally, investigating the model's performance in real-world scenarios and its adaptability to different signing styles and environments would be valuable for practical applications.

6 Funding

This research was supported by the Research Project of China Disabled Persons' Federation in 2024: Research on Talent Cultivation for Disabled Persons in the New Era, Project Approval No. 24 & ZC0029. This funding has facilitated the exploration of innovative methodologies for enhancing talent development strategies. The support has been instrumental in enabling a comprehensive analysis and implementation of AI-driven approaches in this domain.

References

[1] V. Volterra, O. Capirci, M. C. Caselli, P. Rinaldi, and L. Sparaci, "Developmental evidence for continuity from action to gesture to sign/word," *Language, interaction and acquisition*, vol. 8, no. 1, pp. 13–41, 2017. https://doi.org/10.1075/lia.8.1.02vol.

[2] A. Venkatesh, M. Vaibhavi, R. Aishwarya, A. Moghis, and V. Padmapriya, "Real-time sign language gesture and facial expressions detection method to assist the speech and hearing-impaired," in *2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)*, pp. 477–483, IEEE, 2024. https://doi.org/10.1109/icwite59797.2024.10503532.

[3] N. A. N. Mohd Ashril, K. N. Chee, N. Yahaya, and R. Abdul Razak, "Barriers, strategies and accessibility: Enhancing engagement and retention of learners with disabilities in moocs–a systematic litera-

ture review (slr)," *International Journal of Human–Computer Interaction*, pp. 1–12, 2024. https://doi.org/10.1080/10447318.2024.2414892.

[4] H. Brock, I. Farag, and K. Nakadai, "Recognition of non-manual content in continuous japanese sign language," *Sensors*, vol. 20, no. 19, p. 5621, 2020. https://doi.org/10.3390/s20195621.

[5] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015. https://doi.org/10.1016/j.cviu.2015.09.013.

[6] Q. Zhu, J. Li, F. Yuan, and Q. Gan, "Continuous sign language recognition based on motor attention mechanism and frame-level self-distillation," *Machine Vision and Applications*, vol. 36, no. 1, pp. 1–12, 2025. https://doi.org/10.1007/s00138-024-01633-0.

[7] R. A. Salvador and P. Naval, "Towards a feasible hand gesture recognition system as sterile non-contact interface in the operating room with 3d convolutional neural network," *Informatica*, vol. 46, no. 1, 2022. https://doi.org/10.31449/inf.v46i1.3442.

[8] A. R. Sajun, I. Zualkernan, and D. Sankalpa, "A historical survey of advances in transformer architectures," *Applied Sciences*, vol. 14, no. 10, p. 4316, 2024. https://doi.org/10.3390/app14104316.

[9] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7784–7793, 2018. https://doi.org/10.1109/cvpr.2018.00812.

[10] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1316–1325, 2021. https://doi.org/10.1109/cvpr46437.2021.00137.

[11] H. Chang and Q. Ding, "Hierarchical local-global attention in a multi-scale transformer network for enhanced image denoising," *Informatica*, vol. 49, no. 6, 2025. https://doi.org/10.31449/inf.v49i6.6861.

[12] S. Alyami, H. Luqman, and M. Hammoudeh, "Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects," *Information Processing & Management*, vol. 61, no. 5, p. 103774, 2024. https://doi.org/10.1016/j.ipm.2024.103774.

[13] N. Jing, Y. Hu, and Y. Wang, "Research on sign language recognition for hearing-impaired people through the improved yolov5 algorithm combining cbam with focal ciou," *Informatica*, vol. 49, no. 14, 2025. https://doi.org/10.31449/inf.v49i14.7596.

[14] C. Vogler and D. Metaxas, "Parallel hidden markov models for american sign language recognition," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 1, pp. 116–122, IEEE, 1999. https://doi.org/10.1109/iccv.1999.791206.

[15] K. Aditya, P. Chacko, D. Kumari, D. Kumari, and S. Bilgaiyan, "Recent trends in hci: A survey on data glove, leap motion and microsoft kinect," in *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)*, pp. 1–5, IEEE, 2018. https://doi.org/10.1109/icscan.2018.8541163.

[16] Y. Zhang and X. Jiang, "Recent advances on deep learning for sign language recognition.," *CMES-Computer Modeling in Engineering & Sciences*, vol. 139, no. 3, 2024. https://doi.org/10.32604/cmes.2023.045731.

[17] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with blstm-3d residual networks," *IEEE Access*, vol. 7, pp. 38044–38054, 2019. https://doi.org/10.1109/access.2019.2904749.

[18] X. Yan, "Effects of deep learning network optimized by introducing attention mechanism on basketball players' action recognition," *Informatica*, vol. 48, no. 19, 2024. https://doi.org/10.31449/inf.v48i19.6188.

[19] L. Mathew and V. Bindu, "Efficient transformer based sentiment classification models," *Informatica*, vol. 46, no. 8, 2022. https://doi.org/10.31449/inf.v46i8.4332.

[20] C. Taoussi, S. Lyaqini, A. Metrane, and I. Hafidi, "Enhancing machine learning and deep learning models for depression detection: A focus on smote, roberta, and cnn-lstm," *Informatica*, vol. 49, no. 14, 2025. https://doi.org/10.31449/inf.v49i14.7451.

[21] Y. Huang, J. Huang, X. Wu, and Y. Jia, "Dynamic sign language recognition based on cbam with autoencoder time series neural network," *Mobile Information Systems*, vol. 2022, p. 1–10, Apr. 2022. https://doi.org/10.1155/2022/3247781.

[22] D. Kang, "Construction and application of quality assessment model of no-reference images two-stream convolutional neural network," *Informatica*, vol. 48,

no. 15, 2024. https://doi.org/10.31449/inf.v48i15.6388.

[23] L. Hu, L. Gao, Z. Liu, and W. Feng, "Temporal lift pooling for continuous sign language recognition," in *European conference on computer vision*, pp. 511–527, Springer, 2022. https://doi.org/10.1007/978-3-031-19833-5_30.

[24] W. Li, Z. Shang, S. Qian, B. Zhang, J. Zhang, and M. Gao, "A novel intelligent fault diagnosis method of rotating machinery based on signal-to-image mapping and deep gabor convolutional adaptive pooling network," *Expert Systems with Applications*, vol. 205, p. 117716, 2022. https://doi.org/10.1016/j.eswa.2022.117716.

[25] S. D. Viet and C. L. T. Bao, "Effective deep multi-source multi-task learning frameworks for smile detection, emotion recognition and gender classification," *Informatica*, vol. 42, no. 3, 2018. https://doi.org/10.31449/inf.v42i3.2301.

[26] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 13009–13016, 2020. https://doi.org/10.1609/aaai.v34i07.7001.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. https://doi.org/10.1109/cvpr.2016.90.

[28] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4297–4305, 2017. https://doi.org/10.1109/cvpr.2017.364.

[29] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880–1891, 2019. https://doi.org/10.1109/tmm.2018.2889563.

[30] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023–10033, 2020. https://doi.org/10.1109/cvpr42600.2020.01004.

[31] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 697–714, Springer, 2020. https://doi.org/10.1007/978-3-030-58586-0_41.

[32] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *proceedings of the IEEE/CVF international conference on computer vision*, pp. 11542–11551, 2021. https://doi.org/10.1109/iccv48922.2021.01134.

[33] L. Hu, L. Gao, Z. Liu, and W. Feng, "Self-emphasizing network for continuous sign language recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 854–862, 2023. https://doi.org/10.1609/aaai.v37i1.25164.

[34] R. Zuo and B. Mak, "C2slr: Consistency-enhanced continuous sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5131–5140, 2022. https://doi.org/10.1109/cvpr52688.2022.00507.

[35] W. Yin, Y. Hou, Z. Guo, and K. Liu, "Spatial–temporal enhanced network for continuous sign language recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1684–1695, 2023. https://doi.org/10.1109/tcsvt.2023.3296668.

[36] J. Kan, K. Hu, M. Hagenbuchner, A. C. Tsoi, M. Bennamoun, and Z. Wang, "Sign language translation with hierarchical spatio-temporal graph neural network," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3367–3376, 2022. https://doi.org/10.1109/wacv51458.2022.00219.

[37] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous sign language recognition with correlation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2529–2539, 2023. https://doi.org/10.1109/cvpr52729.2023.00249.

[38] L. Guo, W. Xue, Y. Zhou, Z. Kang, T. Yuan, Z. Gao, and S. Chen, "Denoising-diffusion alignment for continuous sign language recognition," *arXiv preprint arXiv:2305.03614*, 2023. https://doi.org/10.48550/arXiv.2305.03614.