



Motion-temporal calibration network for continuous sign language recognition

Hongguan Hu¹ · Jianjun Peng¹ · Zhidong Xiao² · Li Guo¹ · Yi Hu^{3,4} · Di Wu^{3,4}

Received: 23 June 2025 / Accepted: 27 October 2025 / Published online: 9 December 2025
© The Author(s) 2025

Abstract

Continuous Sign Language Recognition (CSLR) is fundamental to bridging the communication gap between hearing-impaired individuals and the broader society. The primary challenge lies in effectively modeling the complex spatial-temporal dynamic features in sign language videos. Current approaches typically employ independent processing strategies for motion feature extraction and temporal modeling, which impedes the unified modeling of action continuity and semantic integrity in sign language sequences. To address these limitations, we propose the Motion-Temporal Calibration Network (MTCNet), a novel framework for continuous sign language recognition that integrates dynamic feature enhancement and temporal calibration. The framework consists of two key innovative modules. First, the Cross-Frame Motion Refinement (CFMR) module implements an inter-frame differential attention mechanism combined with residual learning strategies, enabling precise motion feature modeling and effective enhancement of dynamic information between adjacent frames. Second, the Temporal-Channel Adaptive Recalibration (TCAR) module utilizes adaptive convolution kernel design and a dual-branch feature extraction architecture, facilitating joint optimization in both temporal and channel dimensions. In experimental evaluations, our method demonstrates competitive performance on the widely-used PHOENIX-2014 and PHOENIX-2014-T datasets, achieving results comparable to leading unimodal approaches. Moreover, it achieves state-of-the-art performance on the Chinese Sign Language (CSL) dataset. Through comprehensive ablation studies and quantitative analysis, we validate the effectiveness of our proposed method in fine-grained dynamic feature modeling and long-term dependency capture while maintaining computational efficiency.

Keywords Continuous sign language recognition · Dynamic feature enhancement · Spatial-temporal modeling · Deep learning

Introduction

Sign language, as the primary means of communication for the deaf community, plays a crucial role in facilitating the social integration of deaf individuals. However, the specialized nature and high learning cost of sign language make it difficult for most hearing people to master, severely hindering effective communication between the deaf community and mainstream society [1]. Continuous Sign Language Recognition (CSLR) technology offers a viable solution to break this communication barrier by automatically converting sign language video sequences into text. Sign language recognition research can be broadly categorized into two directions: Isolated Sign Language Recognition (ISLR) [2–4] and Continuous Sign Language Recognition (CSLR) [5–7]. ISLR focuses on recognizing individual sign language vocabulary, with early approaches relying on handcrafted

✉ Jianjun Peng
pengjj@dpu.edu.cn

✉ Zhidong Xiao
zxiao@bournemouth.ac.uk

¹ School of Information Science and Engineering, Dalian Polytechnic University, Dalian 116034, China

² National Centre for Computer Animation, Bournemouth University, Bournemouth BH12 5BB, UK

³ Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China

⁴ Shenyang CASNC Technology Co., Ltd., 110168 Shenyang, China

features for classification, such as color-based hand segmentation [8] and shape-based gesture feature extraction. The advent of deep learning has led to significant performance improvements, with CNN-based methods [9] and skeleton-point-based methods [10] showing promising results. However, these approaches often assume pre-segmented input of single sign language vocabulary, making them difficult to apply in real-world continuous sign language communication scenarios [11, 12]. In contrast, CSLR presents a more complex set of challenges. Firstly, sign language actions exhibit high spatiotemporal correlations, requiring accurate capture and understanding of features such as gestures, facial expressions, and body postures in continuous video sequences [13]. Secondly, sign language often involves collaborative expressions, where certain semantic information can only be accurately conveyed through a combination of multiple action components [14]. Furthermore, the temporal dependencies between different sign language actions are complex, with long-distance semantic associations [15]. These characteristics make CSLR an exceptionally challenging sequence recognition task. Existing CSLR approaches predominantly encounter two fundamental limitations. First, in terms of dynamic feature extraction, current methods [16] utilize shared 2D CNNs to process individual frames independently, inherently constraining their capacity to capture temporal dynamics in sign language sequences. Although 3D CNNs [17] and their (2 + 1)D variants [18] partially address this limitation, their fixed receptive field architectures demonstrate insufficient adaptability in modeling motion patterns across multiple temporal scales [19]. Second, regarding long-term temporal modeling, existing approaches [20] typically optimize temporal and channel-dimensional features separately, thus limiting the model's capability to comprehend long-range dependencies in complex sign language sequences [21]. Specifically, certain semantic concepts require the analysis of gestural evolution over extended temporal windows for accurate interpretation, imposing rigorous demands on temporal modeling capabilities. While Transformer architectures [22] and temporal convolutional networks [23] have shown advancement in modeling long-term dependencies, these approaches often treat all feature channels uniformly, failing to adaptively emphasize channels that are more informative for the current sign context. In CNN feature representations, different channels encode varying levels of discriminative information—some channels may capture more task-relevant patterns while others contain redundant or noisy activations. This channel heterogeneity is particularly pronounced in sign language recognition, where the discriminative features vary significantly across different gesture types: fine-grained finger movements require different feature emphasis compared to broad arm

motions. Traditional approaches apply identical processing to all channels, missing the opportunity to dynamically reweight channels based on their information content and relevance to the current sign gesture [24]. To address these challenges, we propose Motion-Temporal Calibration Network (MTCNet), a comprehensive framework that unifies dynamic feature enhancement with temporal-channel adaptive recalibration for continuous sign language recognition. The key innovation lies in our joint modeling approach that explicitly addresses the channel correlation limitation identified above: while existing methods treat feature channels uniformly, our framework recognizes that different channels encode distinct semantic components in sign language—hand morphology, motion trajectories, facial expressions, and body postures—whose relative importance varies dynamically across different signs. MTCNet achieves this through two synergistic modules: Cross-Frame Motion Refinement (CFMR) captures precise inter-frame correspondences via feature-level attention mechanisms, while Temporal-Channel Adaptive Recalibration (TCAR) employs entropy-driven adaptive kernels to dynamically emphasize semantically relevant channels while modeling multi-scale temporal dependencies. This design enables the model to adaptively correlate, for instance, hand shape channels with motion trajectory channels to distinguish between similar gestures that differ in execution dynamics. We summarize our main contributions as follows:

- A calibration-driven paradigm that bridges motion correspondence and temporal-channel semantics, providing a principled alternative to uniform channel treatment and complementing Transformer-based global modeling.
- A Cross-Frame Motion Refinement module that estimates feature-level displacement via inter-frame differential attention and residual enhancement, enabling fine-grained motion correspondence without optical flow.
- A Temporal-Channel Adaptive Recalibration module that employs adaptive kernels guided by entropy and temporal variation, together with a dual-branch weighting scheme, to explicitly model channel semantics and temporal dependencies across multiple scales.
- A comprehensive empirical study on PHOENIX-2014/2014-T and CSL, with ablations showing complementary gains from CFMR and TCAR and state-of-the-art performance on CSL under RGB-only input.

Related work

Continuous sign language recognition

Continuous Sign Language Recognition (CSLR), a challenging sequence recognition task, has undergone significant development from traditional approaches to deep learning-based methods. Early studies primarily employed Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) approaches incorporating hand-crafted visual features [25]. However, these manual features demonstrated limited robustness and precluded end-to-end optimization within the recognition system.

To leverage the powerful feature extraction capabilities of deep learning, researchers developed hybrid architectures that integrated deep neural networks with HMM [26]. Although these approaches utilized deep networks for visual feature extraction and HMM for sequence dependency modeling, they were constrained by complex training procedures requiring frequent prior probability estimation. Subsequently, CTC-based methods [27–29] attracted considerable research interest due to their end-to-end training advantages. Notable contributions include CMA [27], which leveraged pseudo video-text pairs through cross-modal enhancement strategies and VAC [30], which introduced visual alignment constraints to enhance feature extractor expressiveness. Nevertheless, these methods exhibited limitations in fine-grained motion feature processing and long-term sequence modeling. Contemporary approaches can be categorized into unimodal [31] and multimodal [32] frameworks based on their input modality requirements. Multimodal methods [33] achieve enhanced feature representations by integrating multiple modalities, including RGB video, depth information, and skeletal data. For instance, STMC [34] and C2SLR [35] demonstrated substantial performance improvements through the incorporation of RGB video and human key-point information. However, multimodal approaches typically necessitate additional feature extraction modules and sophisticated fusion strategies, resulting in increased computational overhead and deployment complexity. In contrast, our proposed inter-frame attention-based method operates independently of external cues [34, 36]. Through end-to-end training, our model adaptively extracts motion features from video sequences, enabling automatic capture of hand motion trajectories and facial dynamics without additional computational burden.

Extracting motion features

Motion feature extraction is fundamental to sign language recognition. Traditional approaches primarily employed optical flow [37] for explicit motion modeling; however,

these methods exhibit substantial computational overhead and sensitivity to noise. The development of deep learning has facilitated the emergence of 3D CNN-based approaches [38], which implicitly learn spatiotemporal representations through temporal dimension receptive field extension. Nevertheless, these approaches typically incur significant computational costs, and their rigid convolutional structures demonstrate limited adaptability to the rapid motion patterns inherent in sign language. To enhance feature extraction efficiency, attention-based methodologies have emerged as promising alternatives. The Non-local [39] mechanism enriches spatiotemporal feature representations through global pixel-level correlations, although its dense attention computations present challenges for processing extended sequences. Wang et al. [40] introduced explicit temporal information extraction through inter-frame feature map matching and subsequent fusion with appearance cues. MotionSqueeze [41] implements efficient motion feature extraction through internal learning mechanisms, circumventing external optical flow computation; however, its feature decomposition strategy exhibits insufficient adaptability to complex sign language gestures. Recent advances include STM [42], which employs spatiotemporal-motion attention modules for joint optimization of spatial, temporal, and motion features, demonstrating the potential of attention mechanisms in dynamic feature modeling. Despite these advancements, existing approaches exhibit notable limitations for sign language recognition. TEA's [43] motion excitation relies on channel-wise transformations and direct feature subtraction, focusing primarily on enhancing motion-sensitive channels rather than modeling precise spatial correspondences. This approach, while effective for general action recognition, lacks the spatial precision required for fine-grained gesture analysis in sign language. Similarly, ACTION-Net's [44] motion excitation compresses spatial information through average pooling, losing critical details for sign language interpretation. Both methods employ fixed enhancement strategies that cannot adaptively respond to varying motion magnitudes and predominantly focus on either channel-level motion sensitivity or global patterns, inadequately modeling the precise spatial correspondences essential for hand shape transitions and motion trajectories. Furthermore, a critical distinction lies in their motion representation mechanisms: TEA's approach lacks explicit spatial correspondence modeling between frames, while ACTION-Net's parallel processing architecture fails to capture the sequential nature of sign language gestures. These limitations become particularly evident when processing rapid hand movements and subtle finger articulations that characterize meaningful linguistic elements in sign language.

To address these limitations, we propose the Cross-Frame Motion Refinement (CFMR) module, which implements precise feature-level motion correspondence through an attention-based displacement calculation mechanism. Unlike previous approaches, CFMR operates on feature-level normalized coordinate systems rather than raw pixel positions, achieving resolution invariance while focusing on semantic motion relationships. Our module incorporates an adaptive enhancement strategy that dynamically adjusts based on calculated motion magnitudes, enabling nuanced representation of diverse motion patterns. This combination of precise spatial correspondence modeling, adaptive enhancement, and feature-level processing establishes a robust foundation for capturing the rapid and fine-grained hand gestures that carry critical linguistic information in sign language.

Spatiotemporal modeling methods

Spatiotemporal modeling constitutes a critical challenge in sign language recognition, emphasizing the effective capture of temporal dependencies and spatial semantic information within sign language sequences. Although BiLSTM [45] achieves comprehensive contextual understanding through bidirectional information flow, its sequential processing paradigm inherently constrains parallel computation capabilities and exhibits vulnerability to the vanishing gradient problem in processing extended sequences. While VSL [34] facilitates the modeling of global spatiotemporal dependencies through self-attention mechanisms, its computational complexity exhibits quadratic scaling with sequence length, thereby introducing significant efficiency constraints. TDN [46] implements a two-level differential modeling paradigm for capturing long-term structured motion features, whereas VideoSwin [47] effectively integrates spatial and temporal information in video data by extending the Swin Transformer's [48] spatial locality inductive bias to the spatiotemporal domain. MeMViT [24] enhances the processing of long-term contextual information through memory-augmented mechanisms, while V4D [49] employs 4D convolution operations to capture long-range spatiotemporal feature evolution while maintaining the robust representation capabilities of 3D convolutions. TCN [50] achieves long-term dependency modeling through the implementation of dilated convolutions and residual connections, with VideoSwin further extending this framework to enable multi-scale spatiotemporal feature learning.

In contrast to existing approaches, our method employs adaptive convolution to modulate feature weights across both temporal and channel dimensions, thereby effectively capturing salient action representations in sign language sequences.

Method

Overall

This paper presents MTCNet, a novel end-to-end framework for continuous sign language recognition that effectively models motion dynamics and long-term temporal dependencies in sign language sequences. Given a video sequence comprising T consecutive frames, the input is formulated as: $\mathbf{X} = \{\mathbf{x}^{(t)} \mid t = 1, 2, \dots, T\}$, $\mathbf{x}^{(t)} \in \mathbb{R}^{3 \times H \times W}$, where each frame $\mathbf{x}^{(t)}$ denotes an RGB image with spatial dimensions $H \times W$. MTCNet learns a nonlinear mapping function: $\mathcal{F} : \mathbb{R}^{T \times 3 \times H \times W} \rightarrow \mathbb{R}^N$, to transform the video sequence $\mathbf{X}$ into the target sign language word sequence: $\mathbf{Y} = \{\mathbf{y}^{(i)} \mid i = 1, 2, \dots, N\}$, $\mathbf{y}^{(i)} \in \mathcal{V}$, where $\mathcal{V}$ denotes the sign language vocabulary and N represents the target sequence length.

As depicted in Fig. 1, MTCNet employs ResNet [51] as the backbone architecture for hierarchical visual feature extraction, wherein CFMR and TCAR modules are integrated between hierarchical layers to optimize feature representations. The extracted features are processed through a dual-stream temporal modeling framework, where short-term temporal patterns are captured via 1D CNN architecture, while long-term dependencies are modeled through Transformer-based multi-head attention mechanisms. Importantly, the TCAR module and Transformer serve complementary rather than redundant functions: TCAR operates at the feature recalibration level, performing adaptive weight adjustment across channel and temporal dimensions within individual ResNet layers to enhance the quality of hierarchical feature representations, while the Transformer operates at the sequence modeling level, establishing global semantic dependencies across the entire temporal sequence using the enhanced features. This hierarchical design ensures that the Transformer receives optimally calibrated hierarchical features rather than raw CNN outputs, leading to more effective sequence-level dependency modeling than either component could achieve independently.

Following the dual-stream temporal modeling framework, MTCNet employs Connectionist Temporal Classification (CTC) decoding to transform frame-level predictions into coherent sign language glosses. As illustrated in the CTC beam search process, the decoder maintains multiple hypothetical paths simultaneously, where each path represents a potential sequence of sign language tokens including blank symbols for temporal alignment. At each time step t , the decoder expands existing beam paths by calculating the cumulative probability of appending each vocabulary token, effectively modeling the non-monotonic alignment between

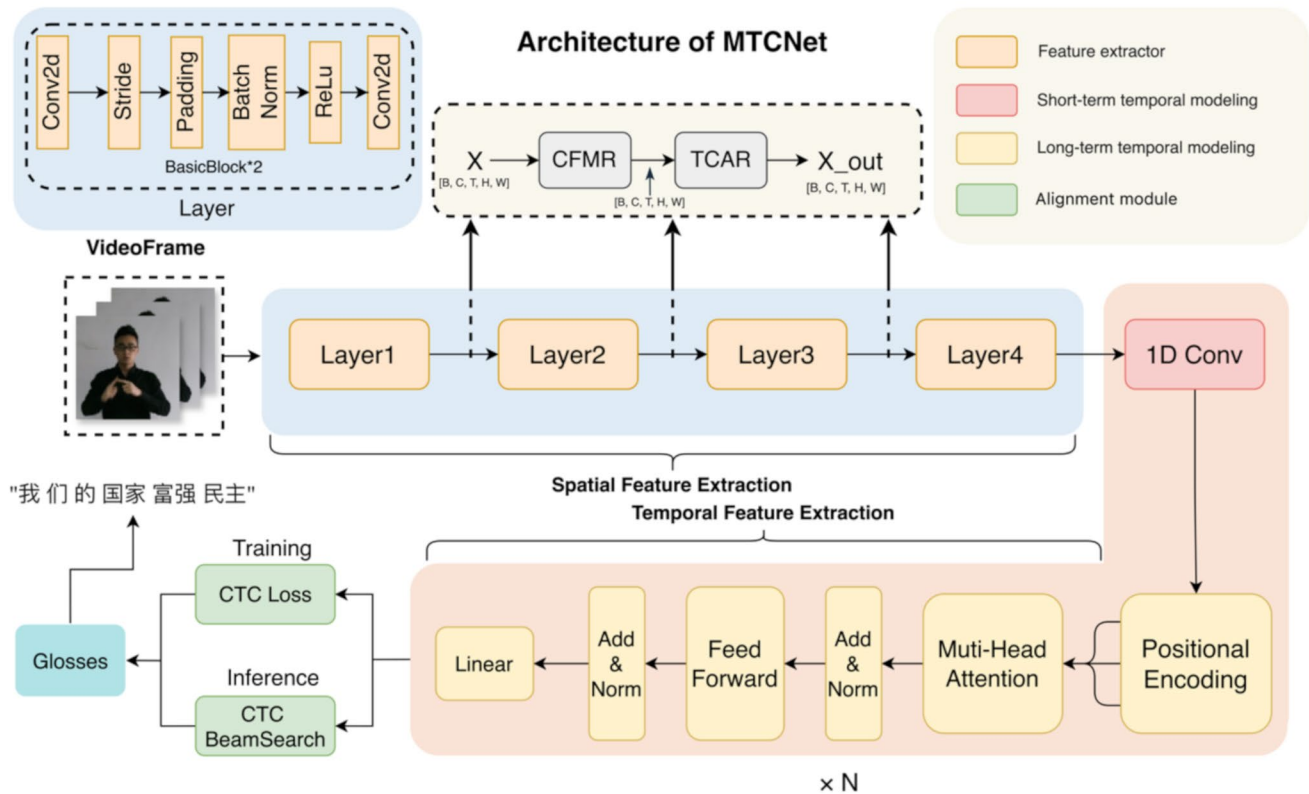


Fig. 1 Architecture of MTCNet. The framework consists of a feature extractor, short-term and long-term temporal modeling components, and an alignment module. The feature extractor enhances hierarchical features via CFMR and TCAR modules between ResNet layers. The

temporal modeling incorporates 1D convolution for local temporal patterns and a Transformer for global semantic dependencies. Finally, the CTC decoder maps the frame-level predictions to sign language glosses

dense video frames and sparse sign language glosses inherent in continuous sign language sequences.

The beam search mechanism with beam width $K=3$ demonstrates the decoder's capacity to handle the temporal variability characteristic of sign language expressions, where individual signs may span variable frame durations and exhibit natural coarticulation effects. As shown in Fig. 2, the beam width $K=3$ serves as an illustrative example, with the optimal value being determined empirically based on the complexity of the target dataset and computational constraints. The blank token insertion strategy addresses the fundamental challenge of CTC-based sequence modeling in sign language recognition: accommodating the temporal misalignment between frame-level feature representations and word-level semantic boundaries. This decoding paradigm proves particularly effective for continuous sign language recognition, as it inherently accounts for the variable-length nature of sign language gestures while maintaining computational efficiency through dynamic programming. The final sequence selection based on cumulative probability ensures optimal global alignment between the enhanced spatiotemporal features extracted by CFMR and TCAR modules and

the target sign language vocabulary, thereby completing the end-to-end recognition pipeline.

Cross-frame motion refinement module

Sign language recognition, as a challenging spatial-temporal sequence understanding problem, necessitates accurate capture of continuous motion dynamics. This paper introduces an inter-frame attention-based feature enhancement paradigm that enhances the representation of continuous sign gestures through explicit modeling of temporal dynamic correlations. As illustrated in Fig. 3, for consecutive frames t and $t+1$, we employ attention mechanisms to compute cross-frame similarity matrices and leverage the resulting weights for motion displacement estimation. The methodology is formulated as follows:

Initially, the module extracts query features Q and key features K from frames t and $t+1$, respectively. This architecture enables temporal feature correlation modeling. For spatial position (i, j) , the query feature Q is formulated as:

$$Q_{i,j} = A_{i,j}^t W_Q \quad (1)$$

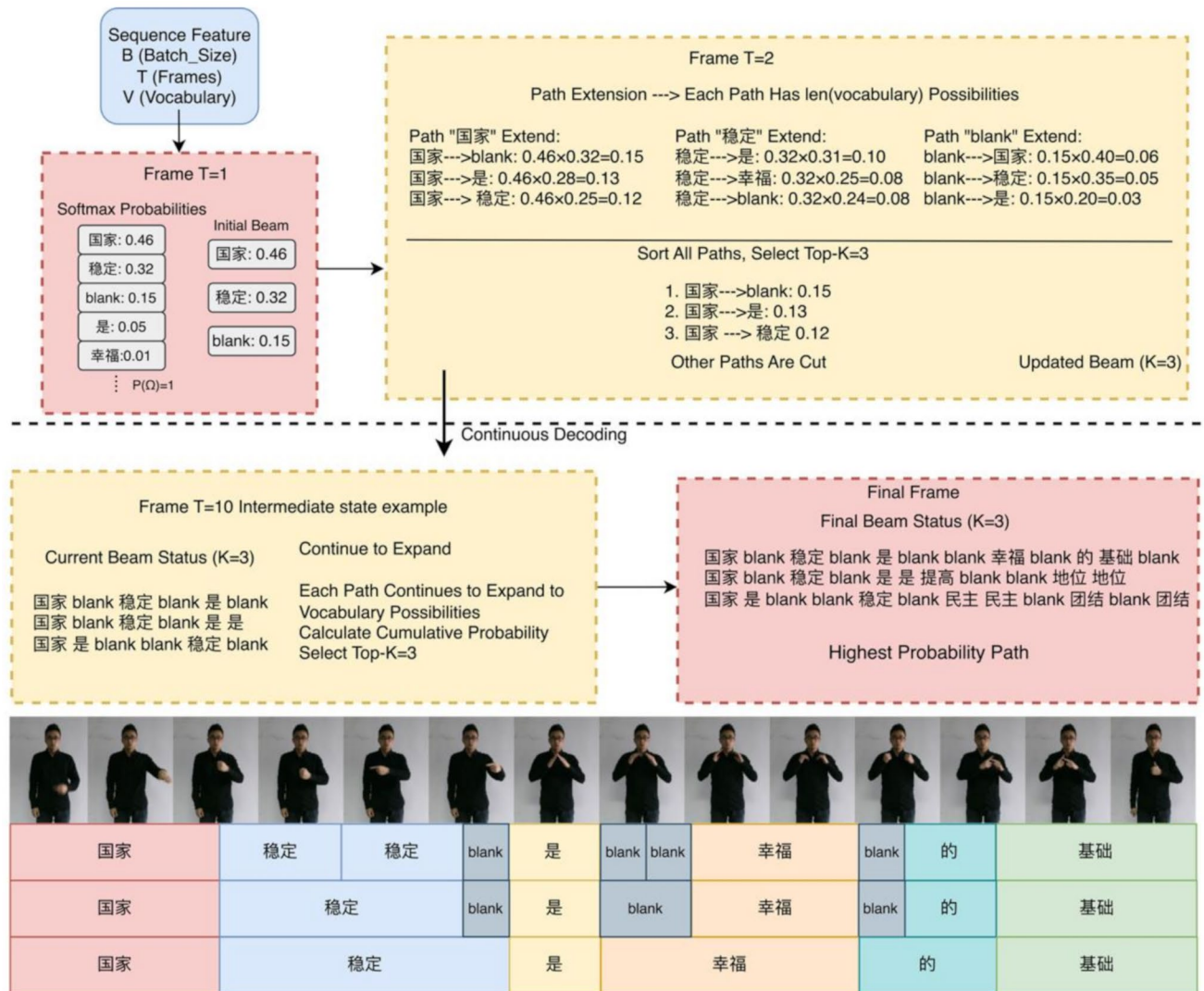


Fig. 2 Illustration of CTC beam search decoding process in MTCNet

where $A_{i,j}^t$ denotes the feature at position (i, j) in frame t , and W_Q represents the query feature mapping matrix. Similarly, the key feature K is extracted from frame $t+1$:

$$K_{i,j} = A_{i,j}^{t+1} W_K \quad (2)$$

where $A_{i,j}^{t+1}$ denotes the feature at position (i, j) in frame $t+1$, and W_K represents the key feature mapping matrix. This formulation enables cross-temporal spatial information extraction and establishes inter-frame feature correspondences.

The attention weights are subsequently computed to quantify spatial correlations between frames t and $t+1$. Specifically, the weight matrix $S_{i,j}^{t \rightarrow t+1}$ is derived through the scaled dot-product attention:

$$S_{i,j}^{t \rightarrow t+1} = \text{softmax} \left(\frac{Q_{i,j} K_{i,j}^T}{\sqrt{d}} \right) \quad (3)$$

where d denotes the feature dimension, and the softmax operation normalizes attention weights to $[0, 1]$. This weight matrix $S_{i,j}^{t \rightarrow t+1}$ encodes pixel-level semantic correlations between frames, enabling adaptive focus on motion-salient regions and facilitating subsequent motion feature enhancement.

In the inter-frame attention mechanism, motion displacement computation serves as an integral component for modeling sign language continuity. The generation of motion displacement enables quantitative characterization of pixel-wise movements between frames t and $t+1$, providing precise motion dynamics for the model. Specifically, we compute inter-frame motion displacement estimates using attention-weighted methods based on the attention weight matrix $S_{i,j}^{t \rightarrow t+1}$ and coordinate map B^{t+1} of frame $t+1$.

It is important to note that the coordinate maps B^t and B^{t+1} are not based on raw pixel positions but are

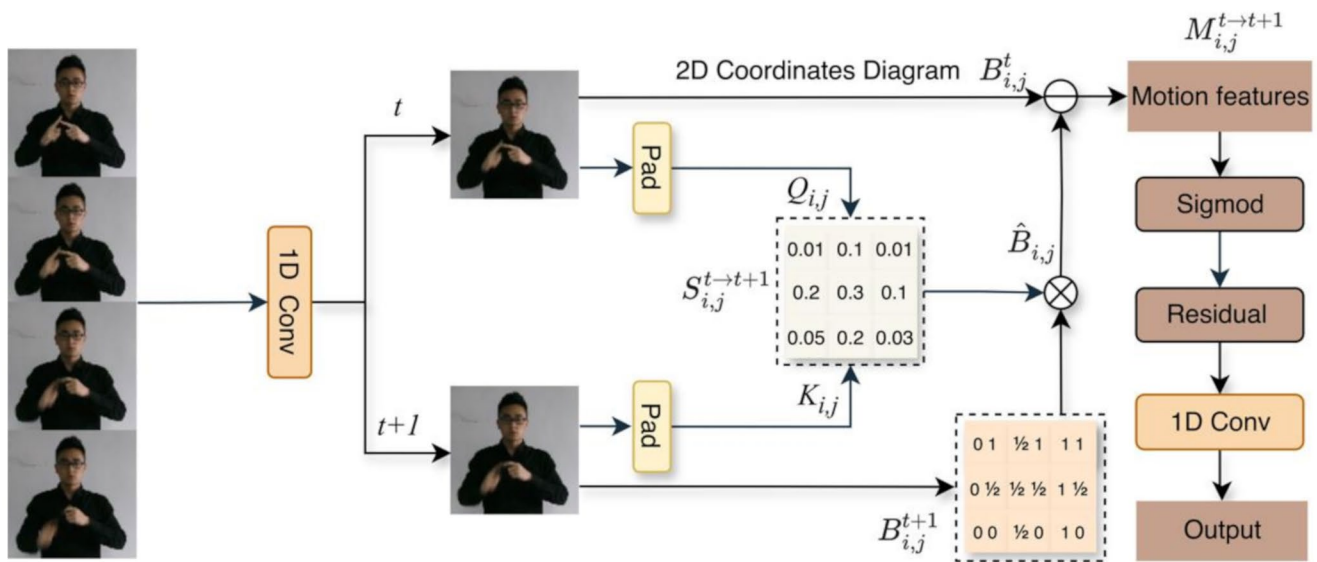


Fig. 3 Illustration of the CFMR module. The proposed module initially extracts query (Q) and key (K) vectors from adjacent frames and computes an attention weight matrix for motion displacement estimation. Furthermore, through an adaptive enhancement strategy coupled with

normalized feature-level positional encodings that represent the relative spatial locations within the feature maps. These coordinate maps are identical for both frames t and $t+1$, with values ranging from (0,1) at the top-left corner to (1,0) at the bottom-right corner, creating a normalized coordinate system across the feature space. This normalization ensures that the motion modeling is resolution-invariant and focuses on semantic relationships rather than absolute pixel displacements.

Through weighted aggregation of coordinate information from adjacent pixel positions, we derive a comprehensive coordinate estimate $B_{i,j}$ that encodes multi-directional motion patterns:

$$B_{i,j} = S_{i,j}^{t \rightarrow t+1} \cdot B^{t+1} \quad (4)$$

This operation effectively utilizes attention weights of neighboring pixels to determine the potential position of the current pixel in frame $t+1$, thus modeling the pixel's temporal trajectory. The inter-frame motion displacement vector $M_{i,j}^{t \rightarrow t+1}$ is then computed as the differential between the weighted coordinates $B_{i,j}$ and the original coordinates B^t of frame t :

$$M_{i,j}^{t \rightarrow t+1} = B_{i,j} - B^t \quad (5)$$

Here, $M_{i,j}^{t \rightarrow t+1}$ denotes the pixel-wise displacement between frames t and $t+1$. This displacement vector characterizes motion dynamics in gesture sequences, facilitating the capture of inter-frame temporal information. In the context of

residual connections, the module achieves precise motion information capture and feature enhancement. $\otimes$ denotes matrix multiplication operation, while $\ominus$ represents element-wise subtraction operation

sign language recognition, such displacement information constitutes the fundamental representation of action features, explicitly encoding the continuity and directionality of sign language movements.

A key advantage of our approach is that the motion modeling occurs at the feature level rather than the raw pixel level. Since the CFMR module is integrated between hierarchical layers of the backbone network, it operates on high-level semantic feature maps that have already undergone multiple layers of convolutional processing. This design allows the module to capture semantic motion patterns (such as hand shape transitions and meaningful gestures) rather than being susceptible to low-level noise or background movements. The attention mechanism further enhances this capability by focusing on semantically relevant regions while suppressing irrelevant background information.

Following the computation of motion displacement vectors, we implement normalization and adaptive enhancement strategies to ensure robust integration of motion information across varying motion magnitudes. Given the inherent amplitude variations in sign language gestures, motion displacement normalization is essential. We derive an adaptive motion enhancement weight $\alpha_{i,j}$ by mapping displacement $M_{i,j}^{t \rightarrow t+1}$ through a sigmoid function to the interval (0,1):

$$\alpha_{i,j} = \sigma(M_{i,j}^{t \rightarrow t+1}) \quad (6)$$

The proposed weighting mechanism not only regulates the magnitude of motion features but also constrains them

within a reasonable range, ensuring computational stability across varying motion amplitudes. The normalized weight coefficient $\alpha_{i,j}$ serves as a motion enhancement factor. Through element-wise multiplication with the current frame features $A_{i,j}^t$, we obtain the modulated representation. Subsequently, a residual connection is employed to combine this with the original frame features $A_{i,j}^t$, yielding the final motion-enhanced features $A_{i,j}^{t*}$:

$$A_{i,j}^{t*} = A_{i,j}^t + \alpha_{i,j} \cdot A_{i,j}^t \quad (7)$$

This formulation implements adaptive motion enhancement, emphasizing salient motion patterns while preserving the stability of background regions. The residual connection ensures that both static contextual information and dynamic motion features are preserved, which is crucial for sign language recognition where both the posture and its changes carry semantic meaning.

Temporal-channel adaptive recalibration

While temporal attention mechanisms effectively capture local variations between adjacent frames in continuous sign language recognition, they exhibit substantial limitations in modeling complex gestural transitions and long-range dependencies. To address this limitation, we propose the TCAR module, illustrated in Fig. 4, which enhances the model's spatial-temporal representation capacity through adaptive feature weighting across temporal and channel dimensions.

Given the input video features $\mathbf{X} \in \mathbb{R}^{B \times C \times T \times H \times W}$, where B, C, T denote the batch size, number of channels, and temporal length respectively, and H, W represent the spatial dimensions, we first apply global spatial average pooling to obtain a compact temporal-channel representation:

$$X_{\text{pool}} \in \mathbb{R}^{B \times C \times T}, X_{\text{pool}}(B, C, T) = \frac{1}{H \cdot W} \sum_{h=1}^H \sum_{w=1}^W X(B, C, T, h, w) \quad (8)$$

This dimensionality reduction strategy preserves essential global contextual information while significantly reducing computational overhead. To address the fundamental challenge of capturing multi-scale features in the complex feature space of sign language recognition, we introduce an adaptive convolution kernel enhancement mechanism. This approach jointly models channel and temporal dimensions through:

$$X'_C(B, C, T) = \text{Conv1D}(X_{\text{pool}}(B, :, T); \text{kernel_size} = k_c) \quad (9)$$

$$X'_T(B, C, T) = \text{Conv1D}(X_{\text{pool}}(B, C, :); \text{kernel_size} = k_t) \quad (10)$$

where X'_C and X'_T denote the features enhanced by adaptive 1D convolution operations along channel and temporal dimensions, respectively. The adaptive kernel design addresses the fundamental challenge that fixed receptive fields cannot adequately capture the multi-scale temporal dynamics and channel-wise semantic correlations inherent in sign language sequences.

The logarithmic scaling factors $\alpha_c \cdot \log_2(C)$ and $\alpha_t \cdot \log_2(T)$ establish dimension-aware base receptive fields that grow sub-linearly with feature dimensions, ensuring computational tractability while providing sufficient coverage for inter-channel and inter-temporal correlations. The complexity-adaptive terms $1 + \beta_c \cdot Sc$ and $1 + \beta_t \cdot St$ enable content-dependent kernel modulation, allowing expansion for semantically rich regions with high information entropy or rapid temporal variations, and contraction for simpler patterns. The learnable bias terms Bc and Bt provide dataset-specific calibration to prevent kernel saturation and ensure odd-integer constraints for symmetric convolution operations. This formulation theoretically guarantees that kernel sizes adapt proportionally to both intrinsic feature complexity and semantic content density:

$$k_c = |\alpha_c \cdot \log_2(C) \cdot (1 + \beta_c \cdot Sc) + Bc|_{\text{odd}} \quad (11)$$

$$k_t = |\alpha_t \cdot \log_2(T) \cdot (1 + \beta_t \cdot St) + Bt|_{\text{odd}} \quad (12)$$

where α_c, α_t are learnable base scaling factors, β_c, β_t are learnable dynamic adjustment weights, and Bc, Bt are learnable bias terms. $|\cdot|_{\text{odd}}$ denotes the operation of obtaining the nearest odd integer.

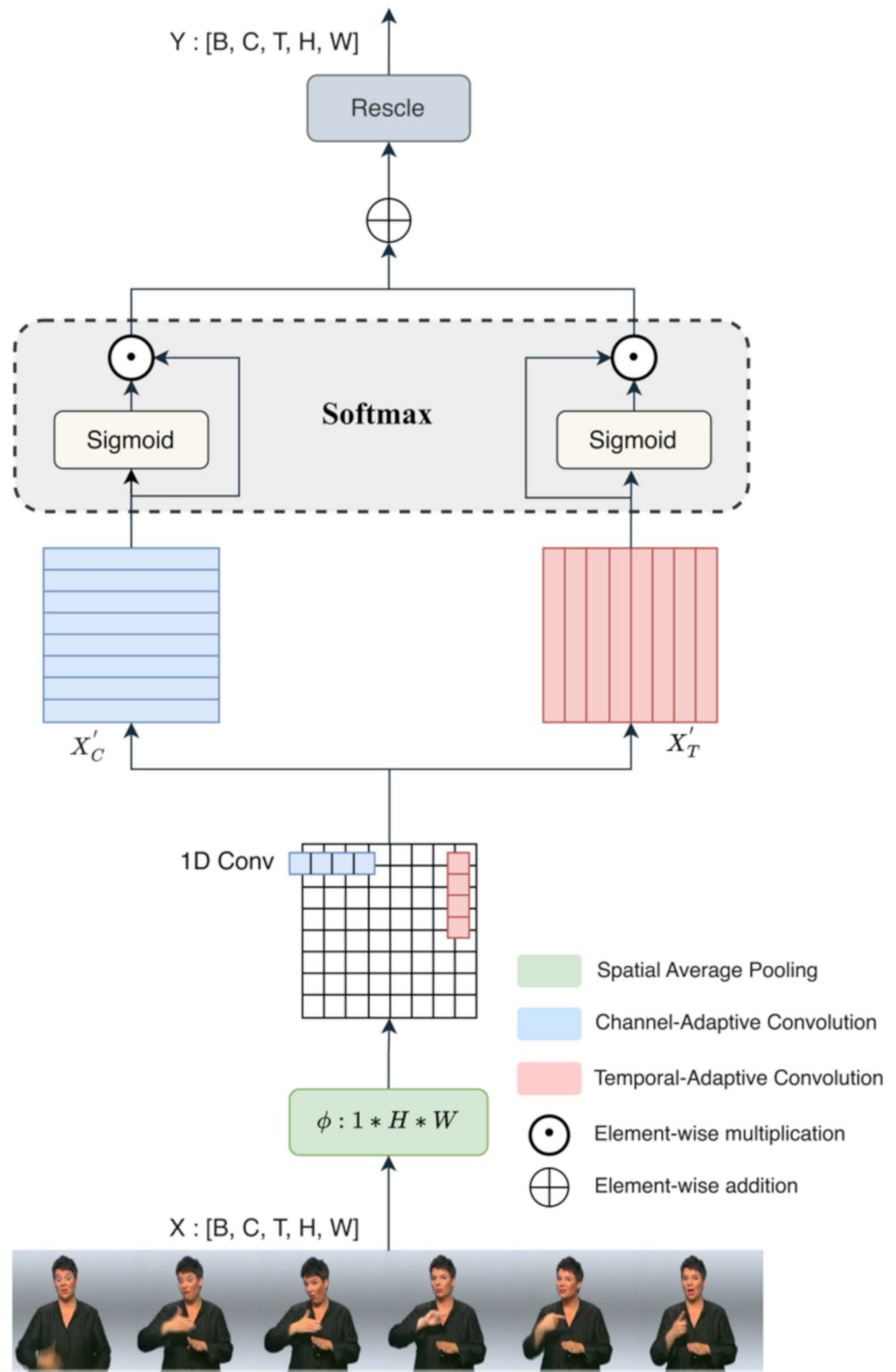
The key innovation in our adaptive kernel design lies in the introduction of sign language-specific dynamic indicators Sc and St , which quantify the complexity and variation characteristics unique to sign language actions:

$$Sc = \sigma \left(\frac{H(X_{\text{pool}})}{\log_2(C)} \right) \quad (13)$$

$$St = \sigma \left(\frac{\|X_t - X_{t-1}\|_F}{\|X_{t-1}\|_F} \right) \quad (14)$$

where St represents the temporal action variation rate calculated using the Frobenius norm ($\|\cdot\|_F$) to measure the relative change between adjacent frames, normalized by the sigmoid function $\sigma(\cdot)$. Sc quantifies the channel complexity using Shannon entropy $H(\cdot)$ of the pooled features,

Fig. 4 Illustration of the TCAR module. The module implements a dual-stream architecture that performs parallel feature recalibration along both temporal and channel dimensions. Multi-scale feature extraction is achieved through adaptive kernels, while dynamic weights are generated via a self-gating mechanism integrated with softmax normalization



normalized by the theoretical maximum entropy $\log_2(C)$. Here, X_{pool} represents dimensionality-reduced features that preserve essential global contextual information while mitigating computational complexity. The Shannon entropy effectively captures the information distribution across

channels, with higher entropy indicating greater variability in channel activations. By normalizing against $\log_2(C)$, we obtain a scale-invariant measure bounded between 0 and 1, enabling consistent complexity assessment across network architectures with varying channel dimensions. This

information-theoretic approach provides a principled metric for quantifying channel-wise feature complexity that directly reflects the intrinsic characteristics of sign language gestures, where certain channels capture critical morphological features while others encode more contextual information. This design enables the convolution kernels to dynamically expand when capturing complex hand gestures or rapid motion transitions, and contract when processing simpler or more static sign components, thereby optimizing the receptive field according to the semantic density of the current sign language element. The differential treatment of channels based on information content further facilitates the extraction of linguistically relevant features while suppressing redundant or noise-dominated channels, addressing the unique challenges of sign language recognition where semantic information is unevenly distributed across feature channels.

This logarithmic adaptive design with sign language-specific modulation dynamically adjusts convolution kernel sizes according to both feature dimensions and semantic characteristics, ensuring an adequate receptive field while maintaining computational efficiency. The design prevents over-modeling of non-informative channels, establishing both theoretical and methodological foundations for analyzing temporal dynamic features. Sign language recognition inherently involves temporal evolution where instantaneous features encode semantic flow of continuous actions. Our architecture captures both short-term dynamics and global dependencies across temporal and channel dimensions.

To enhance feature representation capacity, we integrate a self-gating mechanism. The gating weights are computed by applying the sigmoid function to the input features, followed by element-wise multiplication:

$$\mathbf{X}_{sig_t} = \sigma(\mathbf{X}'_t) \odot \mathbf{X}'_t, \mathbf{X}_{sig_c} = \sigma(\mathbf{X}'_c) \odot \mathbf{X}'_c \quad (15)$$

The temporal features $\mathbf{X}_{sig_t}$ and channel features $\mathbf{X}_{sig_c}$ are enhanced through an adaptive self-gating mechanism, where $\sigma(\cdot)$ denotes the sigmoid activation function and $\odot$ represents the Hadamard product. This self-gating mechanism selectively emphasizes salient features while attenuating redundant information, establishing an adaptive foundation for feature weighting. To ensure robust feature representation while maintaining adaptive capabilities, we implement a “preserve-enhance-balance” concatenation strategy. This mechanism maintains information integrity by preserving original pooled features alongside their adaptively processed counterparts, creating an expanded feature space that accommodates both baseline representations and enhanced dynamics. Specifically, for each processing branch (temporal and channel), the self-gated features are concatenated with the original pooled features along their respective

dimensions, enabling the model to dynamically balance between conservative feature retention and aggressive feature enhancement based on the semantic complexity of the current sign language content. This concatenation strategy ensures joint participation of original and dynamically modulated features in the subsequent weighting process:

$$\mathbf{X}_{cat_t} = \text{Concat}(\mathbf{X}_{pool}, \mathbf{X}_{sig_t}, \text{dim} = T) \quad (16)$$

$$\mathbf{X}_{cat_c} = \text{Concat}(\mathbf{X}_{pool}, \mathbf{X}_{sig_c}, \text{dim} = C) \quad (17)$$

where $\mathbf{X}_{cat_t}$ and $\mathbf{X}_{cat_c}$ denote the concatenated features along temporal and channel dimensions, respectively. The concatenated features are then transformed through a softmax operation to generate normalized weight matrices:

$$\begin{aligned} \mathbf{W}_t &= \text{Softmax}(\mathbf{X}_{cat_t}, \text{dim} = T), \\ \mathbf{W}_c &= \text{Softmax}(\mathbf{X}_{cat_c}, \text{dim} = C) \end{aligned} \quad (18)$$

where $\mathbf{W}_t$ and $\mathbf{W}_c$ represent the normalized weight matrices for temporal and channel dimensions. This normalization ensures unit sum constraints across each temporal step and channel, facilitating dynamic feature balancing. The final feature fusion integrates information from both dimensions through weighted combination:

$$\begin{aligned} \mathbf{X}_{fused} &= (\mathbf{W}_t[:, :, :T] \odot \mathbf{X}_{pool} + \mathbf{W}_t[:, :, T:] \odot \mathbf{X}_{sig_t}) \\ &\quad + (\mathbf{W}_c[:, :C, :] \odot \mathbf{X}_{pool} + \mathbf{W}_c[:, C:, :] \odot \mathbf{X}_{sig_c}) \end{aligned} \quad (19)$$

$\mathbf{X}_{fused}$ denotes the fused feature representation incorporating both temporal and channel-wise information. The module performs element-wise multiplication between the fused features and the original input for adaptive feature recalibration:

$$\mathbf{Y} = \mathbf{X} \odot \mathbf{X}_{fused} \quad (20)$$

where $\mathbf{Y} \in \mathbb{R}^{B \times C \times T \times H \times W}$ represents the enhanced feature output. The module factorizes temporal and channel modeling into independent subtasks, employing lightweight 1D convolutions for feature enhancement while integrating the results at the spatial resolution level to capture fine-grained motion dynamics. In conjunction with inter-frame attention mechanisms, the module optimizes global representations across temporal and channel dimensions beyond local dynamic modeling, enabling more precise comprehension of action-specific semantic transitions.

Experiments

Datasets

The PHOENIX-2014 [25] dataset, established for continuous sign language recognition research, comprises weather forecast videos featuring nine professional sign language interpreters using German Sign Language. The dataset contains 6,841 annotated sentences with a vocabulary size of 1,081, with 5,672 training, 540 validation, and 629 test samples, respectively.

As an extension, PHOENIX-2014-T [52], introduced in 2018, serves as a comprehensive German Sign Language corpus oriented towards weather broadcasts. This dataset supports dual objectives: sign language translation and recognition tasks. The corpus encompasses 2,887 sentences with a vocabulary of 1,066 signs, with 7,096 training, 519 validation, and 642 test samples, respectively.

The CSL dataset [53], developed by the University of Science and Technology of China, employs Microsoft Kinect cameras for data acquisition. This benchmark dataset comprises 100 systematically designed sign language sentences encompassing daily communications, with a vocabulary of 178 Chinese lexical items. The collection features 25,000 high-quality video sequences captured from 50 native signers, recorded at a resolution of 1280×720 pixels and a frame rate of 30 frames per second. The dataset follows a standard partition protocol with an 8:2 ratio for training and testing splits.

Evaluation metric

Word Error Rate (WER) serves as a standard evaluation metric for quantifying the performance of sign language recognition systems. The metric is computed by measuring the minimum edit distance between the predicted sequence and the ground truth sign language sequence. Specifically, WER is defined as the ratio between the total number of edit operations (comprising insertions, deletions, and substitutions) and the total number of words in the reference sequence. The performance of the recognition system exhibits an inverse relationship with WER, where lower values indicate superior recognition accuracy.

Training details

During preprocessing, all input sequences are resized to a fixed size of 256×256 , followed by random cropping to 224×224 during training. The data augmentation pipeline consists of random horizontal flipping ($p=0.5$) and temporal scaling augmentation ($p=0.2$) to enhance model generalization. For inference, center cropping (224×224) is applied to ensure consistent predictions. The backbone architecture comprises ResNet18 for visual feature extraction and a 4-layer Transformer with multi-head attention for temporal modeling, where $d_{\text{model}}=786$ and $d_{\text{ff}}=1536$. The sequence decoder implements CTC-based BeamSearch. Following VAC [30], we optimize the model using a combination of Visual Encoding Loss and Visual Alignment Loss, weighted at 1.0 and 25.0 respectively. The optimization protocol employs AdamW [54] with a two-phase learning rate schedule: a linear warm-up phase for the first 3 epochs ($1e-5$ to $1e-4$), followed by cosine annealing [55] over 50 epochs, decaying to $1e-6$. The learning rate remains constant for the remaining epochs, with training conducted for a total of 80 epochs. This scheduling strategy demonstrates improved convergence stability and model performance.

Ablation experiments

To systematically validate the effectiveness of our proposed modules, we conduct extensive ablation studies on the PHOENIX-2014-T dataset. Specifically, we investigate: the contribution of CFMR and TCAR modules to model performance; the optimal placement of CFMR and TCAR modules within the ResNet backbone; and the sensitivity of model performance to various convolution kernel sizes in the TCAR module.

Table 1 presents the ablation study results for the CFMR and TCAR modules in sign language recognition, evaluated using WER. The baseline model, without CFMR and TCAR modules, yields WERs of 21.2% and 22.9% on the development and test sets, respectively. Upon incorporating the CFMR module, the WERs decrease to 20.1% and 21.5% on the respective sets, indicating enhanced capability in specific feature capture. Similarly, the independent integration of the TCAR module reduces WERs to 19.7% and 21.1% on the development and test sets, validating its efficacy in temporal and channel-wise dynamic feature modeling.

The joint implementation of CFMR and TCAR modules yields optimal performance, achieving WERs of 19.0% and 20.2% on the development and test sets, respectively. These results suggest complementary functionality between the modules, where CFMR facilitates more effective information modeling during feature fusion, while TCAR enhances spatiotemporal feature representation through dynamic

Table 1 Ablation studies of CFMR and TCAR modules on PHOENIX-2014-T dataset

CFMR	TCAR	Dev(%)	Test(%)
×	×	21.2	22.9
√	×	20.1	21.5
×	√	19.7	21.1
√	√	19.0	20.2

convolution. The experimental results substantiate both the individual effectiveness of CFMR and TCAR modules and their synergistic benefits when combined, highlighting their crucial role in improving model performance.

To systematically evaluate the effectiveness of the CFMR module across different network hierarchies, we conducted comprehensive ablation studies, with results presented in Table 2. In single-level deployment experiments, the CFMR module achieved optimal performance when positioned at Layer2 (Dev/Test WER: 20.3%/21.7%), suggesting that mid-level features encompass rich spatial-temporal information conducive to inter-frame attention modeling. Conversely, Layer1 deployment exhibited suboptimal performance (Dev/Test WER: 20.7%/22.3%), attributable to the predominance of local texture information at lower levels, which limits global motion pattern capture. Layer3 implementation yielded intermediate results (Dev/Test WER: 20.4%/21.9%), indicating that high-level semantic features contribute moderately to inter-frame motion modeling.

In dual-level configurations, the Layer2+Layer3 combination demonstrated superior performance (Dev/Test WER: 20.1%/21.7%), validating the efficacy of leveraging both mid and high-level features for hierarchical spatial-temporal information capture in sign language recognition. While deploying CFMR modules across all levels achieved marginally better results (Dev/Test WER: 20.1%/21.5%), the performance gain relative to the optimal dual-layer configuration was modest despite a substantial increase in model parameters. These findings indicate that the Layer2+Layer3 configuration offers an optimal trade-off between performance and computational efficiency for practical applications. Our ablation analysis demonstrates favorable computational cost-benefit relationships. The optimal Layer2+Layer3 configuration adds 1.93 M parameters (7.2% increase) while achieving 1.2% WER improvement, yielding high efficiency. Importantly, full deployment requires additional 0.9 M parameters but provides only marginal 0.2% further improvement, confirming optimal efficiency at the dual-layer configuration and validating our architectural design's computational effectiveness.

Table 3 presents the ablation study on the positioning of TCAR modules across different stages. Experimental results demonstrate that deploying the TCAR module after Layer2 yields optimal single-layer performance (Test WER 21.6%), suggesting that mid-level features with adequate feature dimensionality and spatial-temporal representations are more conducive to channel-temporal joint modeling. The relatively inferior performance at Layer1 (Test WER 21.9%) and Layer3 (Test WER 21.7%) can be attributed to their inherent limitations: Layer1's channel dimension remains insufficient for complex spatial-temporal feature modeling, while Layer3, despite its rich channel

Table 2 Multi-level ablation studies of CFMR module on PHOENIX-2014-T dataset

Configuration	Params(M)	Dev(%)	Test(%)
Baseline	26.884	21.2	22.9
Layer1	27.784	20.7	22.3
Layer2	27.817	20.3	21.7
Layer3	27.882	20.4	21.9
Layer1+2	28.716	20.3	21.8
Layer2+3	28.815	20.1	21.7
Layer1+3	28.782	20.5	22.7
All Layers	29.714	20.1	21.5

Table 3 Ablation study on TCAR module positioning using the PHOENIX-2014-T dataset

Layer1	Layer2	Layer3	Dev(%)	Test(%)
×	×	×	21.2	22.9
√	×	×	20.8	21.9
×	√	×	20.0	21.6
×	×	√	20.1	21.7
×	√	√	19.7	21.1

Table 4 Analysis of Temporal and channel-wise Convolution kernel configurations in TCAR module on PHOENIX-2014-T dataset

Temporal dimension	Channel dimension	Dev(%)	Test(%)
1 × 1	1 × 1	21.8	22.9
1 × 1	3 × 3	21.3	22.4
1 × 1	5 × 5	21.0	22.1
1 × 1	Adaptive kernel	20.2	21.4
3 × 3	1 × 1	21.2	22.3
5 × 5	1 × 1	20.8	21.9
Adaptive kernel	1 × 1	20.0	21.4
Adaptive kernel	Adaptive kernel	19.7	21.1

representation, exhibits constrained capacity in capturing fine-grained temporal dynamics due to reduced spatial resolution. Moreover, concurrent deployment of TCAR modules at both Layer2 and Layer3 leads to substantial performance improvement (Test WER 21.1%). This dual-layer configuration leverages the complementary characteristics of both positions: the Layer2 TCAR facilitates preliminary spatial-temporal modeling while preserving spatial fidelity, whereas the Layer3 TCAR synthesizes and refines these features at a higher semantic level, thereby synergistically enhancing the model's capacity to comprehend sign language sequences across multiple levels of abstraction.

In our ablation studies on the TCAR module, we conducted a systematic investigation of how kernel configurations in both temporal and channel dimensions affect model performance. The results, presented in Table 4, reveal consistent patterns. With the temporal dimension fixed at 1 × 1 convolution kernels, expanding the kernel size in the channel dimension (1 × 1 → 3 × 3 → 5 × 5) yields consistent performance gains, reducing the Test WER from 22.9% to 22.1%. This improvement suggests that expanded receptive fields

enhance the capture of inter-channel correlations. The introduction of adaptive kernels further reduces the Test WER to 21.4%, demonstrating the efficacy of dynamic channel-wise receptive field adjustment.

Analogously, with channel dimensions constrained to 1×1 , increasing the temporal kernel size ($1 \times 1 \rightarrow 3 \times 3 \rightarrow 5 \times 5$) results in improved performance, with Test WER decreasing to 21.9%. This observation confirms that expanded temporal receptive fields enhance sequence dependency modeling. The implementation of adaptive kernels in the temporal dimension leads to a substantial reduction in Test WER to 21.4%, indicating that dynamic temporal receptive fields enable more precise multi-scale temporal pattern capture.

Notably, the simultaneous application of adaptive kernels in both temporal and channel dimensions yields optimal performance (Dev/Test WER: 19.7%/21.1%). This result demonstrates a significant synergistic effect in bi-dimensional dynamic feature modeling: temporal adaptive kernels modulate receptive fields according to sign language motion duration, while channel-wise adaptive kernels dynamically optimize feature combination strategies based on semantic complexity. This dual-dimensional adaptive mechanism facilitates flexible feature extraction across various semantic levels and temporal scales, ultimately achieving superior sign language recognition performance.

The baseline model for continuous sign language recognition exhibits a deletion-dominant error profile, with deletion errors comprising 54.3% of all recognition failures, representing the most critical failure mode since deletion errors constitute complete information loss and are inherently unrecoverable. Figure 5 presents a comprehensive error distribution analysis that elucidates the mechanisms underlying our modules' performance improvements on the PHOENIX-2014-T dataset.

The CFMR module's cross-frame attention mechanism, through motion displacement estimation and pixel-level

temporal correlation modeling, redistributes the error composition by reducing deletion error dominance from 54.3% to 48.1% of total errors. This systematic error redistribution provides quantitative validation of our motion modeling effectiveness. The 6.2% reduction in deletion errors directly quantifies the CFMR module's improved motion correspondence accuracy, as deletion errors predominantly occur when attention mechanisms fail to establish reliable inter-frame correspondences during rapid hand movements. The preserved error total with redistributed composition demonstrates that our inter-frame attention mechanism successfully maintains temporal continuity without introducing correspondence failures—if incorrect motion correspondences were frequent, we would observe increased rather than decreased deletion errors. This error pattern evidence serves as implicit motion consistency validation, indicating that the CFMR module's attention weights establish semantically meaningful correspondences that preserve gesture tracking integrity. This redistribution reflects the module's effectiveness in preserving gesture continuity through enhanced inter-frame feature correspondence. The TCAR module demonstrates superior restructuring capability, further reducing deletion error proportion to 44.2% through its adaptive convolution kernel design that enables multi-scale temporal dependency modeling via dynamic receptive field adjustment. Most significantly, the joint deployment achieves a synergistic error distribution transformation, fundamentally restructuring the error profile from deletion-dominant (54.3%) to substitution-dominant (55.2%). This represents a qualitative improvement in error characteristics, as the proportion reflects the most severe error type (deletion) decreases to 38.5%, while substitution errors, which preserve structural integrity and remain correctable through post-processing, become predominant. The modest increase in insertion error proportion (from 4.6% to 6.3%) enhanced feature sensitivity but constitutes an acceptable

Fig. 5 WER decomposition analysis on PHOENIX-2014 and PHOENIX-2014-T datasets

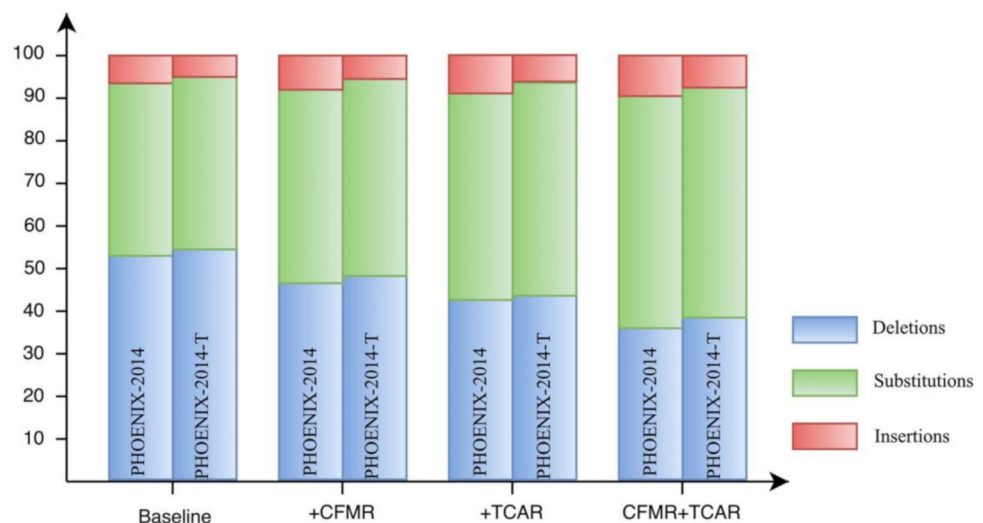


Table 5 Comparison of RGB-only and multi-modal approaches on the PHOENIX-2014 and PHOENIX-2014-T datasets

Method	Backbone	Pre	Phoenix-2014 Dev(%)	Test(%)	Phoenix-2014-T Dev(%)	Test(%)
RGB-based						
CNN-HMM [1]	GoogLeNet	√	31.6	32.5	-	-
SFL [56]	ResNet	√	26.2	26.8	25.1	26.1
Joint-SLRT [57]	Transformer	×	-	-	24.6	24.5
FCN [58]	CNN	×	23.7	23.9	23.3	25.1
SignBT [59]	CNN	√	-	-	22.7	23.9
MMTLB [60]	S3D	√	-	-	21.9	22.5
LCSA [61]	VGG	√	21.4	21.9	-	-
CMA [27]	GoogLeNet	√	21.3	21.9	-	-
VAC [30]	ResNet	√	21.2	22.3	-	-
SKMD [62]	ResNet	√	20.8	21.0	20.8	22.4
SSSLR [63]	ResNet	√	20.9	20.8	20.5	22.3
TLP [64]	ResNet	√	19.7	20.8	19.4	21.2
SEN [65]	ResNet	√	19.5	21.0	19.3	20.7
MSCA-Net [66]	ResNet	√	19.1	20.0	19.2	20.5
CorrNet [31]	ResNet	√	18.8	19.4	18.9	20.5
CVT-SLR [32]	ResNet	√	19.8	20.1	19.4	20.3
STNet [67]	ResNet	√	19.2	19.3	19.2	19.8
MTCNet(Ours)	ResNet	√	19.0	19.7	19.0	20.2
Additional Modalities						
EfficientConSignformer [68]	S3D	√	20.1	19.9	19.5	19.7
C2SLR [35]	ResNet	√	20.5	20.4	20.2	20.4
STMC [34]	VGG	√	21.1	20.7	19.6	21.0
TwoStreamSLR [36]	S3D	√	18.4	18.8	17.7	19.3
ConSignformer [69]	S3D	√	18.5	18.5	17.9	18.5
SignVTCL [70]	S3D	√	17.3	17.6	16.9	17.9

Table 6 Performance comparison with existing sign Language recognition methods on CSL dataset

Method	WER(%)
LS-HAN [53]	17.3
SubUnet [71]	11.0
SF-Net [72]	3.8
FCN [58]	3.0
CRKD [73]	2.6
STMC [34]	2.1
VAC [30]	1.6
C2SLR [35]	0.9
CorrNet [31]	0.8
AdaptSign [74]	0.7
MSTNet [75]	0.7
C2SLR+SRM [76]	0.68
MTCNet(Ours)	0.5

trade-off given the dramatic reduction in the most critical error category. This error distribution transformation demonstrates that our modules effectively address the primary failure mode in continuous sign language recognition by converting irrecoverable information loss into correctable misclassifications.

Comparison with state-of-the-art methods

To comprehensively evaluate the effectiveness of our proposed approach, we conduct extensive comparisons between MTCNet and state-of-the-art sign language recognition methods. Table 5 presents comparative results on the PHOENIX-2014 and PHOENIX-2014-T datasets, encompassing both RGB-based and multi-modal approaches, while Table 6 demonstrates the comparative evaluation on the CSL dataset.

The proposed MTCNet achieves highly competitive results on the challenging PHOENIX-2014-T dataset, delivering strong performance with WERs of 19.0% and 20.2% on the Dev and Test sets respectively. These results significantly outperform numerous established methods including SSSLR (Dev/Test WER: 20.5%/22.3%), and TLP (Dev/Test WER: 19.4%/21.2%), while closely rivaling the current leading RGB-only approaches such as STNet (Dev/Test WER: 19.2%/19.8%). Moreover, MTCNet directly addresses the challenging characteristics of the PHOENIX dataset, including inter-signer variability and background noise, through its CFMR module that dynamically tracks hand movements via inter-frame attention weights, making it particularly robust to signer-dependent variations.

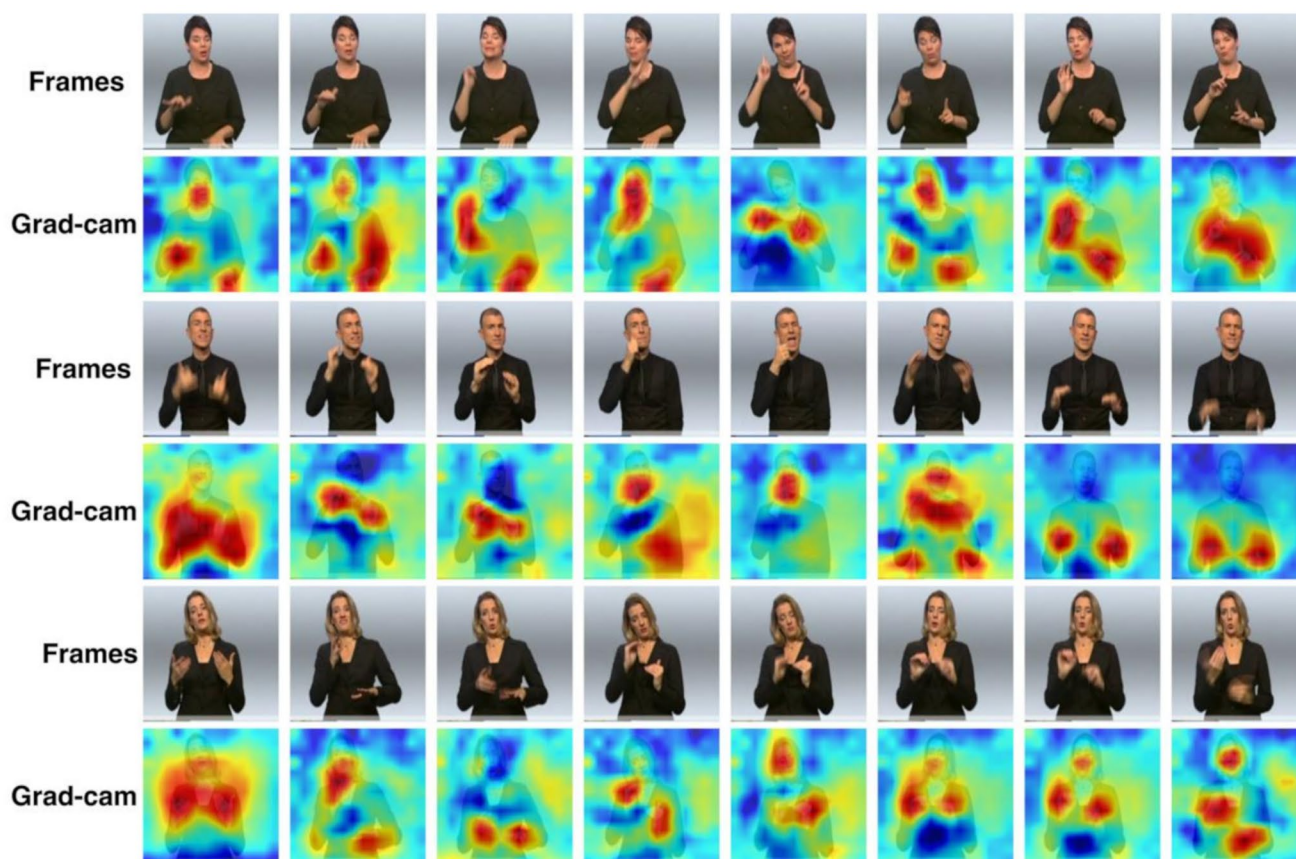


Fig. 6 Visualization of attention regions using Grad-CAM in MTCNet

When compared with recent temporal modeling frameworks like MSCA-Net (Dev/Test WER: 19.2/20.5%) and SEN (Dev/Test WER: 19.3/20.7%), our method demonstrates superior performance through the TCAR module's adaptive kernel design, which effectively captures multi-scale temporal dependencies critical for modeling the complex temporal dynamics in sign language. Figure 6 presents the visualization analysis of MTCNet. The attention heatmaps illustrate the model's focus on hand regions during continuous sign language sequence processing, where regions in red correspond to higher attention weights. Through the collaborative mechanism of CFMR and TCAR modules, the model exhibits robust capability in localizing and tracking critical regions of sign language gestures, thus validating its effectiveness in fine-grained dynamic feature extraction. Although multi-modal approaches such as SignVTCL achieve superior performance (Dev/Test WER: 16.9%/17.9%), they typically require additional computational modules for keypoint detection or feature extraction, leading to increased computational complexity and deployment challenges. In contrast, our method, utilizing solely RGB input while incorporating the proposed CFMR and TCAR modules, achieves comparable performance to

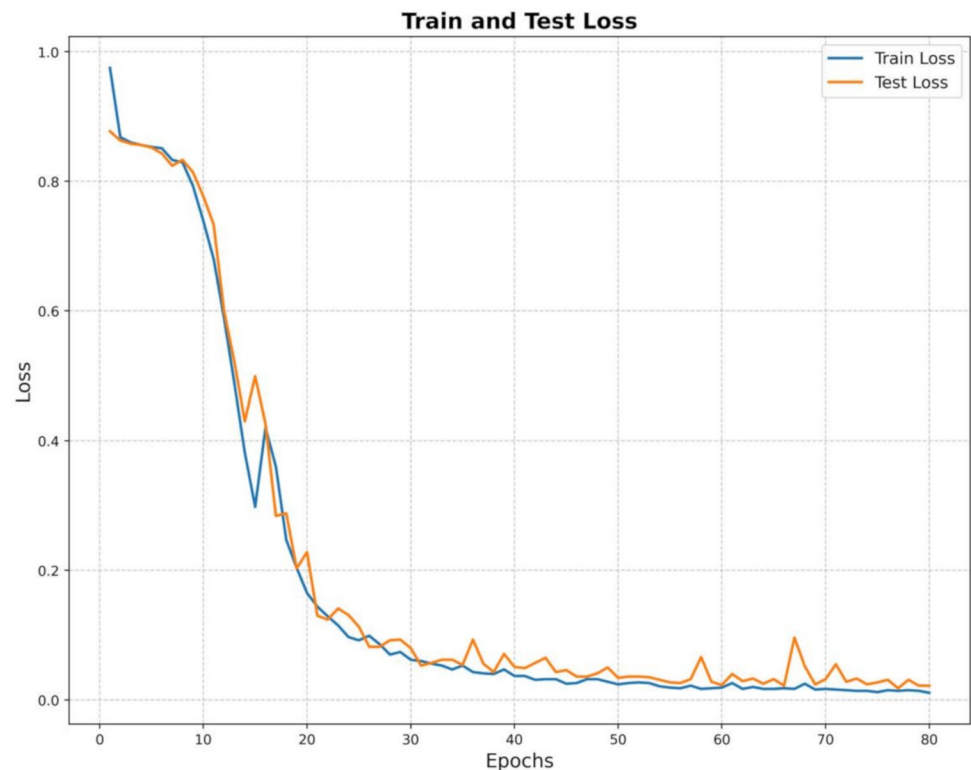
Table 7 Computational efficiency comparison with efficient consign-former

Metric	Efficient ConSignformer	MTCNet
Parameters	120	29.71
Inference Time (seconds)	0.16	0.04
Time for MHA (microseconds)	560	257
Gflops	> 300	183

multi-modal approaches, thereby demonstrating enhanced practical utility.

Our computational efficiency evaluation demonstrates MTCNet's significant advantages compared to Efficient ConSignformer across multiple performance metrics. As shown in Table 7, MTCNet achieves approximately 75% parameter reduction with only 29.91 M parameters versus Efficient ConSignformer's 120 M. This substantial difference stems primarily from our unimodal architecture that eliminates the need for separate modality-specific encoders and fusion networks. The efficiency extends to inference speed, where MTCNet (0.04s) operates four times faster than Efficient ConSignformer (0.16s), enabling more responsive real-time applications. Our streamlined attention mechanism further reduces multi-head attention computation time by 54% (257 μ s vs. 560 μ s). Most notably, MTCNet

Fig. 7 Evolution of training and test loss on the CSL dataset. The train loss trajectory is shown in blue, while the test loss is shown in the orange curve. Both metrics show consistent convergence behavior throughout the training process



requires only 183 GFLOPs of computation, approximately 39% lower than our conservative estimate of Efficient ConSignformer's computational requirements (>300 GFLOPs, based on architectural complexity analysis). We note that most existing works in sign language recognition literature do not explicitly report GFLOPs metrics, necessitating our careful estimation based on model architecture. These metrics quantitatively validate our claim that MTCNet achieves superior computational efficiency while maintaining competitive recognition accuracy, as evidenced by comparable WER scores on PHOENIX-2014-T dataset (MTCNet: Test 20.2%, Efficient ConSignformer: Test 19.7%), making it particularly suitable for deployment in resource-constrained environments.

On the CSL dataset, MTCNet achieves a WER of 0.5%, surpassing the previous state-of-the-art method C2SLR+SRM (0.68%) as shown in Table 6. Figure 7 depicts the evolution of training and testing losses over epochs, which reveals a consistent downward trend. The convergence behavior indicates the model's successful acquisition of discriminative feature representations and robust generalization capabilities on the test set. These results further validate the cross-dataset generalization efficacy of our proposed inter-frame dynamic feature enhancement and temporal channel adaptive recalibration strategies. Notably, MTCNet maintains reliable recognition performance on the CSL dataset, which encompasses diverse daily-life scenarios and complex sign

language expressions, thus demonstrating its significant potential for real-world applications.

The substantial performance variation between PHOENIX-2014-T (20.2% WER) and CSL (0.5% WER) reflects fundamental dataset characteristics rather than architectural limitations. PHOENIX-2014-T exhibits inherently higher recognition complexity due to its larger vocabulary (1,066 vs. 178 signs), natural broadcast conditions with variable lighting and backgrounds, and inter-signer stylistic variations among professional interpreters. Conversely, CSL's controlled laboratory environment, standardized gesture execution, and carefully curated vocabulary contribute to its lower baseline difficulty. Crucially, our proposed modules demonstrate consistent positive contributions across both datasets, validating their generalization capabilities. MTCNet achieves substantial WER reduction from 22.9% to 20.2% on PHOENIX-2014-T and establishes state-of-the-art performance on CSL with 0.5% WER. These consistent improvements across datasets with vastly different complexity profiles demonstrate that our architectural innovations capture fundamental spatiotemporal dynamics of sign language rather than dataset-specific patterns. The inter-frame attention mechanism in CFMR and adaptive kernel design in TCAR address universal challenges in continuous sign language recognition—motion continuity modeling and multi-scale temporal dependency capture—thereby ensuring robust generalization across diverse signing conditions and vocabulary complexities.

Conclusion

This paper presents a novel continuous sign language recognition framework that leverages the CFMR module and the TCAR module for precise sign language action modeling. The CFMR module incorporates an inter-frame differential attention mechanism to explicitly capture motion trajectories within sign language sequences, while preserving contextual information through residual connections. Additionally, the TCAR module implements adaptive convolution kernels with a dual-branch architecture, enabling dynamic feature selection across temporal and channel dimensions, thereby substantially enhancing the model's capacity for long-term dependency modeling. Comprehensive experimental evaluations demonstrate the framework's competitive performance across PHOENIX-2014-T, PHOENIX-2014-T, and CSL datasets. Visualization studies validate the effectiveness of the proposed modules in accurately localizing hand regions and tracking their motion trajectories. The proposed end-to-end framework exhibits significant advantages in both computational efficiency and recognition accuracy compared to existing approaches based on spatiotemporal modeling or multimodal fusion. Future directions include investigating contrastive learning-based feature representations, reducing annotation dependencies, and developing end-to-end real-time sign language recognition systems for mobile device deployment.

Acknowledgments This research project was partially funded by the International Science Partnerships Fund of the British Council (Grant No. 1668 XIAO Japan).

Author contributions HH: Algorithm innovation, Methodology, Experiments, Data analysis, Validation, Writing manuscripts, Visualization. JP: Data preprocessing, Technical guidance, Result analysis, Manuscript revision. ZX: Investigation, Supervision, Experimental validation. YH: Model evaluation, Comparative experiments, Literature review, Manuscript revision. DW: Model optimization, Experimental environment setup. LG: Project administration, Supervision, Technical guidance.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose in relation to this research. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter discussed in this manuscript. The authors declare that they have no competing interests that could have influenced the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this

article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Koller O, Zargaran S, Ney H et al (2018) Deep sign: enabling robust statistical continuous sign Language recognition via hybrid CNN-HMMs [J]. *Int J Comput Vision* 126:1311–1325
2. Hu H, Zhao W, Zhou W et al (2021) SignBERT: Pre-training of hand-model-aware representation for sign language recognition; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, [C]
3. Laines D, Gonzalez-Mendoza M, Ochoa-Ruiz G et al (2023) Isolated sign language recognition based on tree structure skeleton images; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, [C]
4. Lin K, Wang X (2024) SKIM: Skeleton-Based isolated sign Language recognition with part mixing [J]. *IEEE Trans Multimedia* 26:4271–4280
5. Aloysius N (2020) Understanding vision-based continuous sign Language recognition [J]. *Multimedia Tools Appl* 79(31):22177–22209
6. Hu L, Gao L, Liu Z et al (2024) Scalable frame resolution for efficient continuous sign Language recognition [J]. *Pattern Recogn* 145:109903
7. Rastgoo R, Kiani K (2024) Word separation in continuous sign Language using isolated signs and post-processing [J]. *Expert Syst Appl* 249:123695
8. Channayanamath M, Math A, Peddigari V et al (2021) [C] Dynamic hand gesture recognition using 3d-convolutional neural network; proceedings of the Communication Software and Networks: Proceedings of INDIA 2019, F, Springer
9. Cui R, Liu H, Zhang C (2017) Recurrent convolutional neural networks for continuous sign language recognition by staged optimization; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, [C]
10. Liang W, Xu X (2021) [C] Skeleton-based sign language recognition with attention-enhanced graph convolutional networks; proceedings of the Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10, F, Springer
11. Camgoz Nc, Koller O et al (2020) [C] HADFIELD S., Multi-channel transformers for multi-articulatory sign language translation; proceedings of the Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, F, Springer
12. Jinkai M, Jianjun P, Zhidong X et al (2024) Review of modular continuous sign Language recognition methods and techniques [J]. *J Chin Comput Syst* 45(10):2428–2441
13. Li D, Yu X, Xu C et al (2020) Transferring cross-domain knowledge for video sign language recognition; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, [C]
14. Pu J, Zhou W, Li H (2019) Iterative alignment network for continuous sign language recognition; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]

15. Rastgoo R, Kiani K, Escalera S (2021) Sign Language recognition: A deep survey [J]. *Expert Syst Appl* 164:113794
16. Koller O (2017) ZARGARAN S, NEY H. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, [C]
17. Carreira J (2017) ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset; proceedings of the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, F, [C]
18. Tran D, Wang H et al (2018) TORRESANI L. A closer look at spatiotemporal convolutions for action recognition; proceedings of the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, F, [C]
19. Liu Z, Luo D, Wang Y et al (2020) Teinet: Towards an efficient architecture for video recognition; proceedings of the Proceedings of the AAAI conference on artificial intelligence, F, [C]
20. Cui Z, Zhang W, Li Z et al (2023) Spatial-temporal transformer for end-to-end sign Language recognition [J]. *Complex Intell Syst* 9(4):4645–4656
21. Saunders B, Camgoz N C, Bowden R (2020) Progressive transformers for end-to-end sign language production; proceedings of the Computer Vision–ECCV. : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, F, 2020 [C]. Springer
22. Vaswani A (2017) Attention is all you need [J]. *Advances in Neural Information Processing Systems*
23. Lea C, Flynn M D, Vidal R et al (2017) Temporal convolutional networks for action segmentation and detection; proceedings of the proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, F, [C]
24. Wu C-Y, Li Y, Mangalam K et al (2022) Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, [C]
25. Koller O, Forster J (2015) Continuous sign Language recognition: towards large vocabulary statistical recognition systems handling multiple signers [J]. *Comput Vis Image Underst* 141:108–125
26. Koller O, Camgoz N C, Ney H et al (2019) Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign Language videos [J]. *IEEE Trans Pattern Anal Mach Intell* 42(9):2306–2320
27. PU J, ZHOU W, Hu H et al (2020) Boosting continuous sign language recognition via cross modality augmentation; proceedings of the Proceedings of the 28th ACM international conference on multimedia, F, [C]
28. Du Y, Chen Z, Xie H et al (2024) SVTRv2: CTC beats Encoder-Decoder models in scene text recognition [J]. *arXiv preprint arXiv:2411.15858*
29. Fan R, Chu W, Chang P et al (2023) A ctc alignment-based non-autoregressive transformer for end-to-end automatic speech recognition [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1436–1448
30. Min Y, Hao A, Chai X et al (2021) Visual alignment constraint for continuous sign language recognition; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, [C]
31. Hu L, Gao L, Liu Z et al (2023) Continuous sign language recognition with correlation network; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, [C]
32. Zheng J, Wang Y, Tan C et al (2023) Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
33. Gao L, Shi P (2024) Cross-modal knowledge distillation for continuous sign Language recognition [J]. *Neural Netw* 179:106587
34. Zhou H, Zhou W, Zhou Y et al (2020) Spatial-temporal multi-cue network for continuous sign language recognition; proceedings of the Proceedings of the AAAI conference on artificial intelligence, F, [C]
35. Zuo R, Mak B (2022) C2slr: Consistency-enhanced continuous sign language recognition; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, [C]
36. Chen Y, Zuo R, Wei F et al (2022) Two-stream network for sign Language recognition and translation [J]. *Adv Neural Inf Process Syst* 35:17043–17056
37. Brox T, Malik J (2010) Large displacement optical flow: descriptor matching in variational motion Estimation [J]. *IEEE Trans Pattern Anal Mach Intell* 33(3):500–513
38. Abdullahi Sb, Bolon-Canedo V, Chamnongthai K et al (2024) Spatial-temporal feature-based End-to-end fourier network for 3D sign Language recognition [J]. *Expert Syst Appl* 248:123258
39. Wang X, Girshick R, Gupta A et al (2018) Non-local neural networks; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, [C]
40. Wang H, Tran D, Torresani L et al (2020) Video modeling with correlation networks; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
41. Kwon H, Kim M, Kwak S et al (2020) Motionsqueeze: Neural motion feature learning for video understanding; proceedings of the Computer Vision–ECCV. : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, F, 2020 [C]. Springer
42. Jiang B, Wang M, Gan W et al (2019) Stm: Spatiotemporal and motion encoding for action recognition; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, [C]
43. Li Y, Ji B, Shi X et al (2020) Tea: Temporal excitation and aggregation for action recognition; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
44. Wang Z, She Q, Smolic A (2021) Action-net: Multipath excitation for action recognition; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
45. Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF models for sequence tagging [J]. *arXiv preprint arXiv:1508.01991*
46. Wang L, Tong Z, Ji B et al (2021) Tdn: Temporal difference networks for efficient action recognition; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
47. Liu Z, Ning J, Cao Y et al (2022) Video swin transformer; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
48. Liu Z, Lin Y, Cao Y et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, [C]
49. Zhang S, Guo S, Huang W et al (2020) V4d: 4d convolutional neural networks for video-level representation learning [J]. *arXiv preprint arXiv:2002.07442*
50. Bai S, Kolter J Z, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling [J]. *arXiv preprint arXiv:1803.01271*

51. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, [C]
52. Camgoz N C, Hadfield S, Koller O et al (2018) Neural sign language translation; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, [C]
53. Huang J, Zhou W, Zhang Q et al (2018) Video-based sign language recognition without temporal segmentation; proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, F, [C]
54. Loshchilov I (2017) Decoupled weight decay regularization [J]. arXiv preprint arXiv:1711.05101
55. Loshchilov I, Hutter F (2016) Sgdr Stochastic gradient descent with warm restarts [J]. arXiv preprint arXiv:1608.03983
56. Niu Z, Mak B (2020) Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition; proceedings of the Computer Vision–ECCV : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, F, 2020 [C]. Springer
57. Camgoz N C, Koller O, Hadfield S et al (2020) Sign language transformers: Joint end-to-end sign language recognition and translation; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
58. Cheng KL, Yang Z, Chen Q et al (2020) Fully convolutional networks for continuous sign language recognition; proceedings of the Computer Vision–ECCV. : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16, F, 2020 [C]. Springer
59. Zhou H, Zhou W, Qi W et al (2021) Improving sign language translation with monolingual data by sign back-translation; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, [C]
60. Chen Y, Wei F, Sun X et al (2022) A simple multi-modality transfer learning baseline for sign language translation; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, [C]
61. Zuo R (2022) Local Context-aware Self-attention for continuous sign Language Recognition}} [J]. Proc Interspeech 2022:4810–4814
62. Hao A, Min Y, Chen X (2021) Self-mutual distillation learning for continuous sign language recognition; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, [C]
63. Jang Y, Oh Y, Cho J W et al (2023) [C] Self-sufficient framework for continuous sign language recognition; proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), F, IEEE
64. Hu L, Gao L, Liu Z et al (2022) [C] Temporal lift pooling for continuous sign language recognition; proceedings of the European conference on computer vision, F, Springer
65. Hu L, Gao L, Liu Z et al (2023) Self-emphasizing network for continuous sign language recognition; proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, F, [C]
66. Senhua X, Liqing G, Liang W et al (2024) Multi-scale context-aware network for continuous sign Language recognition [J], vol 6. Virtual Reality & Intelligent Hardware, pp 323–337. 4
67. Wang Z, Li D, Jiang R et al (2025) Continuous sign Language recognition with Multi-Scale Spatial-Temporal feature enhancement [J]. IEEE Access
68. Aloysius N, Geetha M, Nedungadi P (2025) Optimized Multi-Modal Conformer-based framework for continuous sign Language recognition [J]. IEEE Open J Comput Soc
69. Aloysius N, Nedungadi P (2024) Continuous Sign Language Recognition with Adapted Conformer via Unsupervised Pretraining [J]. arXiv preprint arXiv:2405.12018
70. Chen H, Wang J, Guo Z et al (2024) SignVTCL: Multi-Modal continuous sign Language recognition enhanced by Visual-Textual contrastive learning [J]. arXiv preprint arXiv:2401.11847
71. Cihan Camgoz N, Hadfield S, Koller O et al (2017) Subnets: End-to-end hand shape and continuous sign language recognition; proceedings of the Proceedings of the IEEE international conference on computer vision, F, [C]
72. Yang Z, Shi Z, Shen X et al (2019) Sf-net: structured feature network for continuous sign Language recognition [J]. arXiv preprint arXiv:1908.01341
73. Zhu Q, Li J, Yuan F et al (2024) Continuous sign Language recognition based on cross-resolution knowledge distillation [J]. Arab J Sci Eng. : 1–13
74. Hu L, Shi T, GAO L et al (2024) Improving continuous sign Language recognition with adapted image models [J]. arXiv preprint arXiv:2404.08226
75. Zhu Q, Li J, Yuan F et al (2024) Multiscale Temporal network for continuous sign Language recognition [J]. J Electron Imaging 33(2):023059–023059
76. Zuo R, Mak B (2024) Improving continuous sign Language recognition with consistency constraints and signer removal [J]. ACM Trans Multimedia Comput Commun Appl 20(6):1–25

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.