



DYNAMIC ADAPTIVE REPRESENTATION CALIBRATION FOR KEYPOINT-BASED SIGN LANGUAGE RECOGNITION

Ying Ma*

Nanchang Hang Kong University, marian0225@163.com
Nanchang, China

Nan Lin

Nanchang Hang Kong University, lindazhuang0815@outlook.com
Nanchang, China

* Corresponding author

ABSTRACT

Aim/Purpose	Individual differences in signers' speed and amplitude cause variations in keypoint sequences, leading to redundant features that degrade the classification performance of sign language recognition models.
Background	To reduce the impact of individual differences among signers, this paper proposes a Dynamic Adaptive Representation Calibration (DARC) module that helps the model focus on the most informative features.
Methodology	The study utilizes component-based modeling of high-dimensional semantic features and introduces sample-adaptive learnable weights. The method was validated using large-scale sign language datasets, specifically WLASL2000 and MSASL1000.
Contribution	The proposed DARC module dynamically adjusts the importance of different features for each input, allowing the model to focus more on informative patterns and reducing the influence of individual differences among signers.
Findings	The proposed method achieved a Top-1 accuracy of 56.86% on the WLASL2000 dataset (a 3.2% relative improvement over the state-of-the-art DSTA-SLR) and 65.88% on the MSASL1000 dataset. Visualizations confirm that the proposed module significantly improves feature discriminability and clustering clarity. Additionally, multi-stream training validations demonstrate consistent performance improvements across spatial and temporal dynamic streams, effectively raising the model's overall representation ceiling.

Accepting Editor Geoffrey Z. Liu | Received: May 13, 2026 | Revised: July 15, July 22, 2026 |

Accepted: July 23, 2026.

Cite as: Ma, Y., & Lin, N. (2026). Dynamic adaptive representation calibration for keypoint-based sign language recognition. *Interdisciplinary Journal of Information, Knowledge, and Management*, 21, Article 14.

<https://doi.org/10.28945/5859>

(CC BY-NC 4.0) This article is licensed to you under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/). When you copy and redistribute this paper in full or in part, you need to provide proper attribution to it to ensure that others can later locate this work (and to ensure that others do not accuse you of plagiarism). You may (and we encourage you to) adapt, remix, transform, and build upon the material for any non-commercial purposes. This license does not permit you to use this material for commercial purposes.

Recommendations for Practitioners	Developers of practical sign language recognition systems should consider integrating dynamic adaptive calibration to better handle natural variations among different signers in real-world environments.
Recommendations for Researchers	Researchers should further explore dynamic feature recalibration techniques to improve model robustness against individual variations in sequence-based recognition tasks.
Impact on Society	By improving the accuracy of keypoint-based computer vision models, this research helps develop more accessible and scalable tools to bridge the communication gap between the deaf/hard-of-hearing community and hearing individuals.
Future Research	Future studies should explore the application of the DARC module to other video-based action recognition tasks or investigate its performance in real-time continuous sign language translation.
Keywords	sign language recognition, keypoint sequences, feature calibration, dynamic adaptive representation

INTRODUCTION

Sign language serves as a vital communication medium for the deaf community, embodying rich linguistic structures and cultural connotations (Seaman et al., 1988). With advancements in computer vision and deep learning, sign language recognition (SLR) has emerged as a crucial research direction to enable barrier-free information exchange (Wu et al., 2025). In recent years, keypoint-based SLR methods have garnered significant attention. For instance, Deng, Leng, Chen, et al. (2024) developed a multi-stream network to extract multi-level features directly from skeleton keypoints, while Tran Anh et al. (2025) achieved efficient recognition by optimizing keypoint selection for Graph Convolutional Networks (GCNs). Ultimately, these approaches rely on keypoints extracted from human pose estimation to represent sign trajectories in a compact, structured format. This enables focused capture of the spatiotemporal characteristics of gestures and body movements. Compared with wearable devices that require complex sensors, keypoint-based SLR methods offer greater convenience and scalability (Renjith et al., 2025).

The two typical examples shown in Figure 1(a) and (b) illustrate the major challenges in sign language recognition (D. Li et al., 2020). Variations in gesture amplitude and signing speed among different signers often lead to diverse representations of the same sign vocabulary (Zuo et al., 2023). One signer may express a sign with slow, broad arm movements, while another may perform the same sign with rapid, compact hand gestures. These individual differences alter the spatial scale and temporal duration of the extracted keypoint sequences, thereby affecting the model’s discriminative ability.



(a) Execution of the vocabulary “Umbrella” by Signer 1 and Signer 2



(b) Execution of the vocabulary “Add” by Signer 3 and Signer 4

Figure 1. Variation in sign language execution

Existing keypoint-based sign language recognition models typically employ Graph Convolutional Networks (GCNs) or their variants (Vázquez-Enríquez et al., 2021). GCN is an architecture designed to process graph-structured data by treating human joints as nodes and bones as edges. Building on this structure, these models learn spatial dependencies among hand and upper-limb joints while incorporating temporal convolutions or attention mechanisms to capture dynamic changes in sign movements. Such approaches demonstrate strong performance across both spatial and temporal dimensions (Comivi Alowonou & Han, 2024).

However, analysis of existing research reveals that while most models can adaptively capture variations in joint structures and temporal patterns, their adaptability to different samples remains limited, presenting the following issues. First, as shown in Figure 1, the same vocabulary executed by different signers exhibits significant variations. Current sign language recognition models, which rely on static feature-weight networks, struggle to effectively adapt to substantial differences in motion amplitude and speed among signers (Khan et al., 2025). Second, high-dimensional semantic features contain redundant and weakly correlated components that interfere with the model’s discriminative decisions, thereby limiting overall classification capability. Finally, existing dynamic adjustment methods often introduce substantial additional parameters, resulting in high computational overhead that hinders practical deployment.

To address these challenges, this paper proposes a Dynamic Adaptive Representation Calibration (DARC) module for deep feature calibration. This module comprises two components: dynamic weight generation and residual gating. The dynamic weight generation component adaptively generates a set of feature weight coefficients based on the input sample’s feature distribution, applying dimension-wise weighting to the input vector. This approach enables the model to enhance key components, suppress redundant components, and weaken weakly correlated components for different action samples, thereby improving overall recognition performance and model adaptability. Simultaneously, the residual gating mechanism introduces an adjustable fusion ratio coefficient, λ , to blend the original features with dynamically weighted features. This ensures the model adapts to individual variations while preserving the original feature information, effectively preventing a degradation in discriminative power due to overreliance on dynamically generated weights.

The main contributions of this work are as follows: In the field of sign language recognition, we propose a DARC module for deep feature calibration that generates dynamic weights to enhance key components of sample features. The effectiveness of this module is validated on two large-scale sign language datasets, WLASL2000 and MSASL1000. Experimental results demonstrate that it significantly improves the model’s classification performance and generalization capability.

LITERATURE REVIEW

SIGN LANGUAGE RECOGNITION

Research on sign language recognition in the field of computer vision can be broadly divided into two major categories according to the data modality processed: RGB-image-based methods and human-keypoint-based methods (Al Abdullah et al., 2024). RGB-based approaches typically process continuous raw color video streams directly, aiming to jointly capture the global motion and fine-grained appearance features of the signer at the pixel level. In recent years, researchers have extensively explored RGB-based architectures. For example, Bertasius et al. (2021) proposed the TimeSformer architecture, which entirely eschews convolutional operations and is purely based on a spatiotemporal self-attention mechanism. It decomposes videos into frame-level sequences of image patches and independently computes temporal and spatial attention, thereby overcoming the limitations of the local receptive fields inherent in traditional 3D convolutions. Wong et al. (2025) proposed SignRep, a scalable self-supervised masked autoencoding framework that randomly masks portions of the input continuous stream and forces the network to reconstruct the missing representations using spatio-temporal context, thereby enabling the model to resist visual disturbances caused by local occlusions and background noise.

However, although these RGB-based methods can capture extremely rich visual information, their heavy reliance on raw pixels makes them highly susceptible to external environmental factors such as cluttered backgrounds, illumination changes, and variations in the signer’s clothing (Darmayunata et al., 2025). Moreover, processing high-dimensional continuous video streams often incurs substantial computational and storage costs. To overcome these drawbacks, keypoint-based methods that leverage pose estimation algorithms (e.g., OpenPose, MediaPipe) have gradually emerged (Guan et al., 2025). Keypoint-based sign language recognition methods typically first extract 2D or 3D spatial coordinates of hand, face, and body keypoints frame-by-frame from continuous video streams, and then organize these coordinate sequences into spatio-temporal graphs or structured sequence vectors, which are fed into deep networks to extract high-level semantic and motion trajectory features. This lightweight data representation not only effectively eliminates interference from external environmental factors but also naturally conforms to the physical topological laws of the human skeleton. With the development of deep learning, Graph Convolutional Networks (GCNs) have been widely adopted, and the research focus has gradually shifted from early heuristic physical topology modeling to adaptive spatial-topology learning and refined extraction of temporal dynamics.

In terms of spatial topology and feature aggregation, the Dynamic Spatial-Temporal Attention for Sign Language Recognition (DSTA-SLR) model proposed by L. Hu et al. (2024) breaks the rigid topological constraints of traditional GCNs by employing an input-adaptive graph structure. Instead of being restricted to the natural physical skeletal connections, this model dynamically computes the adjacency matrix based on the features of the current input frame, enabling direct information exchange between nodes that are not physically connected but are semantically highly coordinated (e.g., the left and right hands), thus more accurately capturing complex gesture evolutions. Deng, Leng, Hu, et al. (2024) proposed a Skeleton-based Multi-feature Learning (SML) method that introduces a Self-guided Adaptive Residual Decoupled Graph Convolution (SGAR-DGCN), which divides high-dimensional graph node features into multiple independent subspaces for separate convolution operations, thereby reducing redundant channel coupling. Meanwhile, they employ knowledge distillation to construct a supervised metric loss, guiding the model to spontaneously focus on hard, confusing samples, thereby significantly improving feature-extraction efficiency and classification boundary clarity on large-scale datasets.

At the same time, current models have also made progress in handling temporal dynamics. Sign language sequences contain action pauses and rapid transitions of varying lengths. To address this challenge, Z. Liu et al. (2020) proposed a decoupled and unified multi-scale temporal graph convolutional

network (MS-G3D), which uses parallel temporal convolution kernels with different receptive fields to finely extract short-term rapid actions while also capturing long-range dependencies spanning dozens of frames. Targeting the unique properties of sign language, Özdemir et al. (2023) proposed a multi-cue temporal modeling method that further optimizes the capture of complex temporal dynamics from skeleton sequences. Recently, Min et al. (2025) conducted a systematic review of skeleton-based continuous sign language recognition, discussing in detail the representational capacity and temporal boundary issues in continuous action scenarios. These cutting-edge explorations of temporal dynamics have effectively captured temporal dependencies in sign language and laid the foundation for designing more flexible temporal perception modules in the future.

Table 1 summarizes representative sign language recognition methods from recent years, including their core methodologies, performance on benchmark datasets, and potential areas for improvement.

Table 1. Summary of sign language recognition research

Study	Methodology	Performance achieved	Potential areas for improvement
Bertasius et al. (2021)	Proposed TimeSformer, adapting ViT to video understanding via pure self-attention.	Achieved 79.7% and 81.8% Top-1 accuracy on Kinetics-400/600 .	Requires large-scale training data and high memory; limited scalability to long sequences.
Wong et al. (2025)	Proposed the SignRep self-supervised framework with masked representation learning guided by sign priors.	Achieved 61.05% (WLASL2000), 84.1% (NMFs-CSL), and 81.37% (ASL-Citizen) Top-1 accuracy.	Limited multimodal fusion, random masking lacks semantic guidance, and feature importance is not explicitly modeled.
L. Hu et al. (2024)	Proposed the DSTA framework with graph correlation module, super-node module and Parallel Temporal Convolution Network (PTCN) for multi-scale temporal modeling.	Achieved 53.68% Top-1 accuracy on WLASL2000 , 65.74% on MSASL1000 , 98.1% on SLR500 , and 77.9% on NMFs-CSL .	Convolution-based temporal modeling limits long-range dependency learning; lacks dynamic feature calibration.
Deng, Leng, Hu, et al. (2024)	Proposed the SML framework, the Multi-feature Aggregation (MFA) module to fuse joint, bone, and motion features, employs SGAR-DGCN for skeleton feature learning.	Achieved 55.85% (WLASL2000) and 96.85% (AUTSL); further improved to 61.71% and 98.59% with SamSLRV2.	Relies on pose estimation and handcrafted feature fusion; limited long-range temporal modeling.
Z. Liu et al. (2020)	Proposed the MS-G3D skeleton action recognition network, which combines multi-path STGC modules with joint-bone two-stream fusion to achieve multi-scale spatio-temporal feature learning.	Achieved SOTA on NTU RGB+D 60/120 and Kinetics Skeleton 400 (91.5% CS, 96.2% CV on NTU60).	Fixed graph modeling and limited long-range temporal dependency learning.
Özdemir et al. (2023)	Proposed a multi-cue temporal modeling framework combining ST-GCN, DeepHand, DAN, and MC-LSTM.	Achieved 92.58% (BosporusSign22k) and 90.85% (AUTSL).	Shallow cross-modal interaction, high model complexity, and limited long-range temporal modeling.

As summarized in Table 1, recent studies have primarily focused on improving network architectures, multi-scale spatio-temporal modeling, and multimodal feature fusion, resulting in substantial performance gains. However, most existing methods focus on extracting richer features while paying limited attention to the unequal contributions of deep features in decision-making. Individual differences in signing style, speed, and motion amplitude often produce redundant or less informative feature components, which may degrade recognition performance. Moreover, explicit mechanisms for dynamically calibrating feature importance remain largely unexplored. To address this issue, this paper proposes the DARC module, which adaptively reweights deep feature representations to enhance discriminative information while suppressing redundant components, thereby improving the robustness and generalization of sign language recognition models.

DYNAMIC WEIGHT ADJUSTMENT MECHANISM

Unlike standard convolutional networks, where filter weights remain fixed after training, dynamic weight adjustment mechanisms introduce a data-dependent paradigm that shifts the model from “learning general parameters” to “generating specific parameters.” Dynamic convolution, proposed by Chen et al. (2020), generates attention weights based on the input, thereby significantly enhancing the model’s generalization capability without increasing network depth. In sequence modeling, the self-attention mechanism of the Transformer architecture is also inherently a form of adaptive dynamic weight adjustment, which has proven highly effective (Vaswani et al., 2017).

In the field of skeleton-based action recognition, dynamic mechanisms also play an important role in capturing critical components. Qiu et al. (2022) achieved prioritized attention to informative body parts by dynamically adjusting the importance weights of individual joints on a frame-by-frame basis. Lee et al. (2023) proposed a hierarchically decomposed graph convolutional network (HD-GCN), which employs hierarchical graph convolution and adaptive adjacency weight learning, enabling the model to dynamically adjust the association strengths between joints based on the input action, thereby more effectively modeling complex human movements. Chi et al. (2022) introduced Infogcn, which incorporates an information bottleneck and an adaptive feature reweighting mechanism into the graph convolution process, dynamically enhancing discriminative joint features while suppressing redundant information, thus improving skeleton-based action recognition performance.

RESIDUAL GATING

Residual connections and gating mechanisms are key techniques for optimizing feature representations in deep neural networks. Residual learning (Shafiq & Gu, 2022) effectively alleviates the vanishing gradient problem in deep networks, ensuring the stable transmission of basic spatio-temporal features. Gating mechanisms, such as the Squeeze-and-Excitation (SE) network (J. Hu et al., 2020), can dynamically evaluate and re-assign channel-wise feature weights, thereby suppressing background noise and redundant information. The combination of these two is typically achieved through a residual gating mechanism (Savarese & Figueiredo, 2017), which provides a simple yet effective means of improving network optimization, enabling the model to transmit high-information features and suppress irrelevant ones selectively.

Recent research continues to validate the direct application of residual gating mechanisms in temporal filtering and spatial feature calibration. Gai et al. (2025) proposed a dual-path residual gating system that ingeniously combines residual connections with gating mechanisms to retain critical information selectively. J. Liu et al. (2026) introduced a residual gated fusion module that adaptively injects fine-grained skeleton motion into feature representations via dynamic conditional gating. In sign language recognition, high-dimensional abstract features often exhibit substantial redundancy; introducing residual gating strategies can effectively filter out weakly correlated components without losing core semantic trajectories. This synergistic effect provides a solid theoretical foundation for the residual gating design in the DARC module proposed in this work.

These successful cross-domain applications offer strong theoretical support for the design of the DARC module. By computing input-dependent calibration masks, DARC can dynamically scale features and precisely respond to each signer’s unique speed and spatial amplitude.

METHODOLOGY

OVERVIEW

This research aims to achieve adaptive individual variation through the DARC. The module’s architecture comprises two core components: Dynamic Weight Generation and Residual Gating. Dynamic Weight Generation addresses the challenge faced by sign language recognition models in effectively handling variations in gesture amplitude and speed across different signers. By generating dynamic weight coefficients, it enhances key components while suppressing redundant and weakly correlated components. Residual gating integrates original features with dynamically weighted features. By introducing an adjustable fusion ratio coefficient λ , it effectively prevents model degradation caused by over-reliance on dynamically generated weights.

As shown in Figure 2, original features ($x_{original}$) from the feature extractor (architecture as shown in the figure) undergo Dropout to obtain basic classification features (x_{static}). The model retains the basic classification features (x_{static}) while introducing the DARC module. The original feature ($x_{original}$) is passed through a dynamic weight generation module to produce a set of dynamic weights, denoted as Dyna. Then, $x_{original}$ is multiplied element-wise with Dyna to obtain the dynamic feature ($x_{dynamic}$). Subsequently, the base classification feature (x_{static}) and the dynamic feature ($x_{dynamic}$) are fed into a feature fusion module, where they are fused according to a certain ratio. Finally, the fused feature (x_{dl}) is obtained. Then, x_{static} and x_{dl} are fused via element-wise addition to obtain the enhanced feature ($x_{enhance}$). Simultaneously, x_{static} and $x_{enhance}$ are trained using a joint cross-entropy loss function. This approach maintains the baseline model’s computational complexity while enhancing the model’s adaptability to individual differences among signers, thereby improving recognition performance.

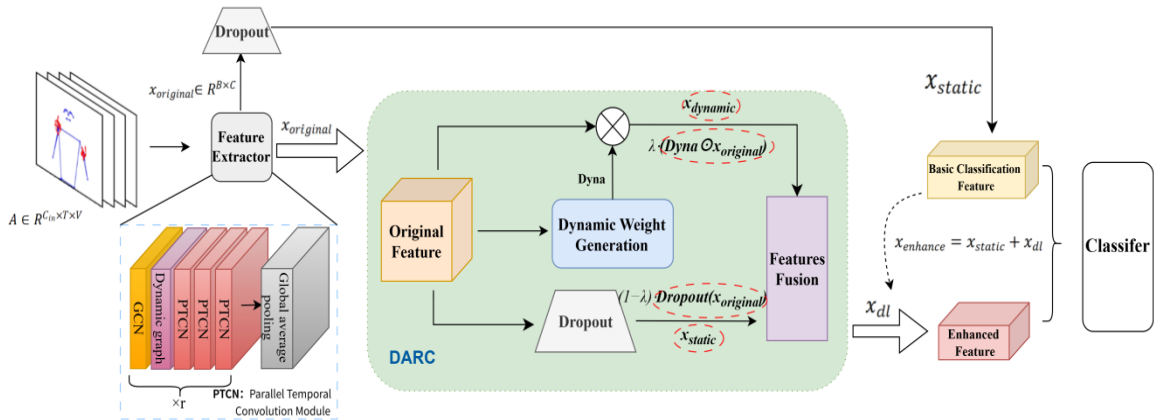


Figure 2. Overall model workflow overview

DYNAMIC ADAPTIVE REPRESENTATION CALIBRATION MODULE

As shown in Figure 3, given a sequence of keypoints $A \in R^{C_{in} \times T \times V}$, where C_{in} denotes the number of coordinate channels, T denotes the number of frames, and V denotes the number of joints. The base model processes this sequence of keypoints through multiple layers of spatiotemporal graph

convolutional modules. Its core functional role is to simultaneously capture the spatial-structural dependencies of the physical joints and the temporal dynamics of the gestures, yielding spatio-temporally fused original features. $x_{original} \in R^{B \times C}$ are obtained, where B denotes the batch size, and C denotes the feature dimension (512 in our implementation).

$$x_{original} = BaseModel(A) \quad (1)$$

Generate dynamic weight coefficients through linear mapping and the Softmax function:

$$Coef = Softmax(Linear(x_{original})) \quad (2)$$

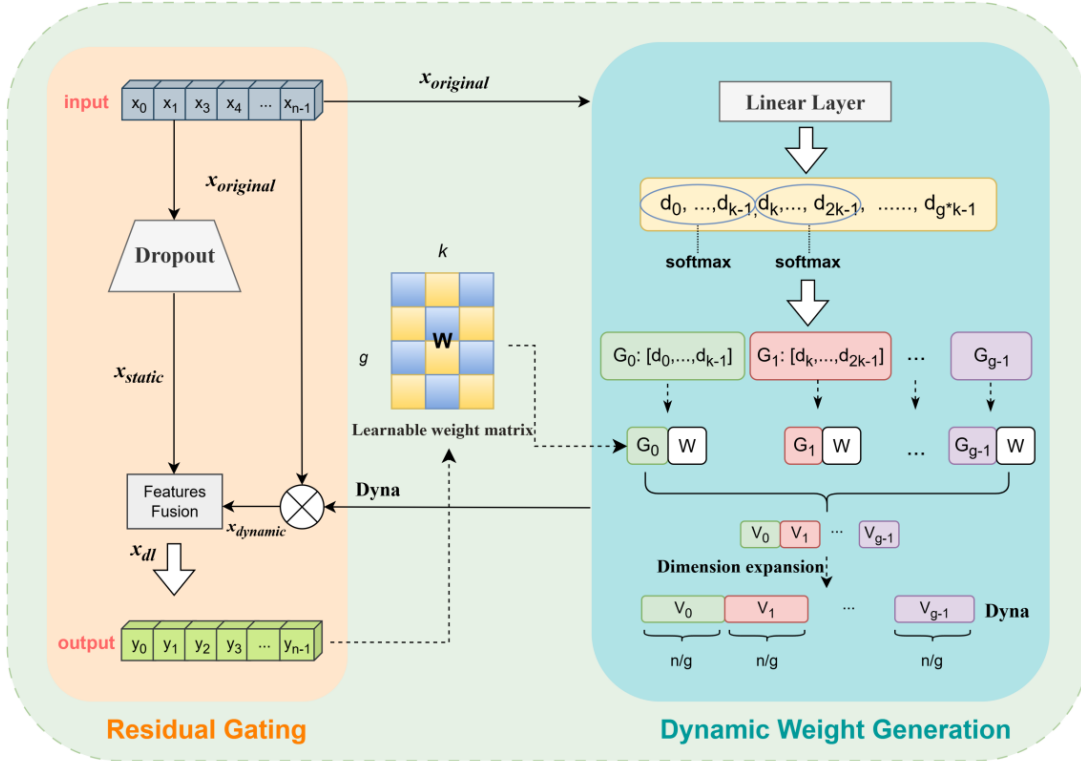


Figure 3. Process overview of the proposed DARC module

Coef is then reshaped into the form $G \times K$, where G denotes the number of groups and K denotes the number of kernels per group, enabling subsequent intra-group fusion. To obtain the final dynamic weights, the model acquires a learnable base weight matrix W during training. An initial base weight matrix W is randomly sampled from a standard normal distribution. Throughout training, all samples collectively contribute gradients to optimize the base weight matrix. During inference, all samples continue using the same trained weight matrix.

By multiplying the dynamic weight coefficients with the base weights, the dynamic weights for each group can be computed, where g and k denote possible values:

$$Dyna^{(G)} = \sum_{k=1}^K Coef^{(g,k)} \cdot W^{(g,k)}, \quad G = 1, 2, 4, \dots, g, \quad K = 1, 2, 4, \dots, k \quad (3)$$

After concatenating all group results, the global dynamic weight vector is obtained:

$$Dyna = [Dyna^{(1)}, Dyna^{(2)}, \dots, Dyna^{(g)}] \quad (4)$$

To match the dimension of the original feature $x_{original}$, a repeated expansion operation is applied to map **Dyna** to the same dimension as $x_{original}$. Finally, the basic classification features and dynamic features are fused through a residual gating mechanism:

$$x_{dl} = (1 - \lambda) \cdot \text{Dropout}(x_{original}) + \lambda \cdot (\text{Dyna} \odot x_{original}) \quad (5)$$

Here, $\odot$ denotes element-wise multiplication. The adjustable dynamic fusion ratio coefficient $\lambda \in [0, 1]$ is introduced to control the degree of dynamic fusion.

This module maintains model compactness while introducing a dynamic weight-generation mechanism, thereby achieving adaptive enhancement of key components across different sample features. Consequently, the model demonstrates significant performance improvements on both the WLASL2000 (D. Li et al., 2020) and MSASL1000 (Joze & Koller, 2018) datasets.

SPECIFIC ARCHITECTURE

This research builds upon the DSTA-SLR model as the foundation for feature extraction, embedding the proposed DARC module into its global feature aggregation stage. Input samples undergo feature extraction through a combination of multi-layer dynamic graph convolution and parallel temporal convolution, yielding rich spatio-temporal features $X \in \mathbb{R}^{B \times C \times T \times V}$, where T denotes the number of frames, and V denotes the number of joints.

To obtain original features, average pooling is performed on both temporal and joint direction information:

$$x_{original} = \text{AvgPool}_{T,V}(X) \quad (6)$$

The resulting global feature representations for each sample are fed into the DARC module to obtain fused features.

$$x_{dl} = \text{DARC}(x_{original}) \quad (7)$$

Based on this, the same input undergoes Dropout processing to obtain the basic classification features x_{static} . These are then combined with the fused features to form the final enhanced features:

$$x_{enhance} = x_{static} + x_{dl} \quad (8)$$

Finally, independent fully connected layers (FC) output the enhanced prediction and basic classification prediction:

$$y_{enhance} = \text{Softmax}(\text{FC}(x_{enhance})) \quad (9)$$

$$y_{static} = \text{Softmax}(\text{FC}(x_{static})) \quad (10)$$

During training, the model simultaneously outputs two sets of predictions: y_{static} and $y_{enhance}$. Learning occurs by minimizing cross-entropy loss given the class label. Two loss functions are defined:

- a. Cross-entropy loss function for enhanced features:

$$L_{enhance} = L_{\text{cross-entropy}}(y_{enhance}, \text{label}) \quad (11)$$

- b. Cross-entropy loss function for basic classification features:

$$L_{static} = L_{\text{cross-entropy}}(y_{static}, \text{label}) \quad (12)$$

To prevent the model from losing the discriminative power of basic classification features while incorporating enhanced features, this method employs a joint optimization strategy. If only the enhanced features are supervised, the model is prone to over-relying on the DARC module, leading to feature degradation. Therefore, the basic classification loss L_{static} is combined with the enhanced

loss $L_{enhance}$. The effect of this joint optimization is to provide a shortcut for gradient back-propagation through the model, alleviating gradient vanishing and accelerating convergence, while enabling complementary advantages between the basic classification features and the enhanced features. A tunable balancing coefficient α is introduced to adjust the trade-off between $L_{enhance}$ and L_{static} . This is achieved by introducing an adjustable balance coefficient α to balance $L_{enhance}$ and L_{static} .

$$L = L_{static} + \alpha L_{enhance} \quad (13)$$

The primary purpose of the balance coefficient α is to regulate the gradient back-propagation strength of DARC during training, coordinating the joint learning of basic classification features and enhanced features. Its value directly influences the training dynamics and final classification performance of the model. When α is set to a small value (e.g., approaching 0), the optimization process is predominantly driven by the basic classification feature loss L_{static} . In this case, the model struggles to learn effectively from the enhanced features, failing to exploit the advantages of deep feature calibration. Conversely, when α is excessively large (e.g., >2), the enhanced feature loss $L_{enhance}$ dominates the total loss. This imbalance causes the network to overfit the feature calibration introduced by DARC during back-propagation, thereby overshadowing the representation of basic classification features. Moreover, an overly large α may induce gradient instability in the early training stages, reducing the model’s generalization ability on unseen data. When α is set to a moderate value, the model can effectively leverage DARC for deep feature calibration while preserving robust basic classification feature representations. An appropriate α value fosters a positive synergistic effect between the two types of features, achieving an optimal trade-off between base and enhanced features in the optimization space.

MODEL TRAINING

For data partitioning, the dataset’s official partitioning is used, dividing the data into a training set and a test set. During training, the model uses a stochastic gradient descent (SGD) optimizer with a momentum coefficient of 0.9, a weight decay coefficient of 0.0001, and an initial learning rate of 0.1. The learning rate scheduling strategy combines warm-up and segmented decay, using linear warm-up over the first 20 epochs to gradually increase the learning rate from a small value to the initial learning rate. When the epoch reaches 150 and 200, the learning rate is decayed to 0.1 times its original value, respectively, to promote model convergence. The training batch size is set to 24, with a total of 250 training epochs. Model checkpoints are saved every 5 epochs for subsequent performance evaluation and optimal model selection.

To enhance model robustness, a dynamic DropGraph regularization technique is introduced (Cheng et al., 2020). This mechanism controls the probability that nodes are retained in the graph structure via the `keep_prob` parameter. During the first 100 epochs, `keep_prob` linearly decays from 1.0 to 0.9, and then remains fixed at 0.9. This is equivalent to randomly “dropping” some skeleton connections during network training, forcing the model to learn more robust feature representations.

In sign language recognition tasks, it is difficult to fully depict complex spatiotemporal dynamic information with a single feature representation. Therefore, this paper adopts a multi-stream training strategy, constructing four input forms and training models independently:

- (1) *Joint Stream*: Taking the original keypoint coordinates as input, it models the spatial position information and static pose features of sign language;
- (2) *Bone Stream*: Constructing a skeletal representation by calculating the vector difference between adjacent joints, which is used to model the structural relationships between joints;
- (3) *Joint Motion Stream*: Differencing the time dimension to extract motion information of key points, which is used to describe the dynamic changes of sign language;

- (4) *Bone Motion Stream*: Further calculate temporal differences based on skeletal representation to model the process of skeleton structure changes over time.

Due to the heterogeneity of the four aforementioned data streams in the feature space, fusing them prematurely can easily lead to interference among features, thereby blurring critical information and introducing noise. The late fusion strategy, by decoupling the feature-learning stages, allows the four data streams to independently perform deep feature extraction of spatial structure and temporal dynamics without mutual interference. This delayed interaction mechanism not only preserves the uniqueness of each spatiotemporal representation to the greatest extent but also effectively aggregates information at the final decision stage, thereby enhancing the model’s generalization ability and classification reliability. Therefore, in the inference stage, this paper adopts a late fusion strategy to perform weighted fusion of the predictions from four models to obtain the final classification result (Pereira et al., 2023). This strategy can effectively integrate information from different input forms, thereby enhancing the overall recognition performance and robustness of the model.

PERFORMANCE METRICS

To comprehensively evaluate the model’s performance on large-scale sign language recognition tasks, this paper uses Top-1 and Top-5 accuracy as the primary evaluation metrics.

(1) Top-1 accuracy

Top-1 accuracy is the core performance metric in sign language recognition, representing the proportion of samples for which the category with the highest predicted probability by the model is exactly consistent with the true label. Its calculation formula is:

$$Top - 1 Accuracy = \frac{N_{correct}}{N_{total}} \quad (14)$$

where $N_{correct}$ represents the number of correctly predicted samples, and N_{total} represents the total number of samples in the test set.

(2) Top-5 accuracy

The Top-5 accuracy measures the proportion of samples with true labels among the top five categories predicted by the model based on probability ranking. The calculation formula is:

$$Top - 5 Accuracy = \frac{N_{top5_correct}}{N_{total}} \quad (15)$$

Among them, $N_{top5_correct}$ represents the number of samples in which the true label appears among the top 5 categories with the highest predicted probabilities. Top-1 accuracy is the most intuitive and widely used evaluation metric in sign language recognition tasks, as it directly reflects the model’s overall recognition performance. In large-scale category scenarios (such as WLASL2000), this metric can effectively measure the discriminative ability of the model in complex classification tasks.

In large-scale sign language recognition tasks, the number of vocabulary categories is vast (WLASL2000 contains approximately 2000 categories), and some words are visually highly similar. Relying solely on a single best guess may not fully reflect the model’s ability to capture semantic associations. Top-5 accuracy provides a more lenient yet practical evaluation perspective, effectively measuring the model’s ability to constrain correct labels within a smaller candidate set.

EVALUATION METHODS

DATASET USED

WLASL2000 (Word-Level American Sign Language 2000) is a large-scale American Sign Language word-level dataset comprising approximately 2000 vocabulary categories, each containing multiple sign language video samples. The dataset features diverse video sources, including different signers,

camera angles, and background environments, exhibiting significant variations in motion, speed, and individual differences. This diversity poses challenges for model robustness and generalization capabilities. The dataset has been widely used in sign language recognition research and is well-suited for evaluating model recognition performance across diverse semantic contexts and complex gestural scenarios. The WLASL100 dataset is a word-level subset of American Sign Language (ASL), comprising approximately 2,000 video samples covering 100 distinct ASL vocabulary items. Recorded across multiple contexts, each ASL vocabulary item is demonstrated by different individuals, with varying sample sizes per item.

MSASL1000 (Multi-Signer American Sign Language 1000) is another large-scale American Sign Language dataset encompassing approximately 1000 vocabulary categories. Unlike WLASL2000, MSASL1000 emphasizes the variability among multiple signers. Each sign category is demonstrated by signers with different identities, increasing the challenge of cross-individual recognition. This dataset exhibits significant variations in video resolution, shooting conditions, and sign styles, effectively evaluating a model’s adaptability to signer variations and real-world application scenarios.

Both datasets provide 27 keypoint data extracted using the MMpose model, including 10 hand keypoints and 7 upper-body keypoints, forming a complete body pose representation (Sengupta et al., 2020). A consistent data preprocessing workflow was applied, encompassing 120-frame temporal sampling, coordinate normalization, and standardized data augmentation strategies, ensuring the reliability and comparability of experimental results.

PROTOTYPE IMPLEMENTATION & PLATFORM

The experiment was implemented based on the PyTorch deep learning framework, with both training and inference completed on a computing platform equipped with an NVIDIA RTX 4090 GPU. The experimental environment utilized Python 3.9, CUDA 11.0, and cuDNN.

For data processing, a custom data feeder (Feeder) is used to uniformly preprocess keypoint sequences. The input data consists of a sequence of 3D skeleton keypoints with dimensions of $[C=3, T=120, V=27]$, representing the number of coordinate channels, time frames, and key-points, respectively. During the training phase, various data augmentation operations are applied to the input sequence, including random selection, random shifting, and random mirroring (mirror probability = 0.5), and the keypoint coordinates are normalized to enhance the model’s robustness to different signers and action variations. In the testing phase, only normalization is retained without introducing random perturbations to ensure the stability and reproducibility of the evaluation results.

RESULT

COMPARISON WITH OTHER MODELS

When comparing against state-of-the-art methods on the two reference datasets, approaches are broadly categorized into RGB-based and keypoint-based methods by their input modalities. As shown in Tables 2 and 3, the proposed DARC achieves Top-1 accuracy exceeding current state-of-the-art methods on both WLASL2000 and MSASL1000, reaching 56.86% and 65.88%, respectively. This demonstrates the effectiveness of DARC in leveraging keypoint features for SLR, even exhibiting advantages on large-scale datasets.

ABLATION STUDY

The value of λ in formula (5) controls the contribution of the dynamic feature to enhanced feature representations. To investigate the specific impact of this coefficient on model performance and determine the optimal fusion ratio, experiments were conducted by setting λ within the range $\{0, 0.1, 0.3, 0.5, 0.7, 1.0\}$. The effects on classification accuracy were observed through ablation experiments on the WLASL100 dataset, with results shown in Table 4.

Table 2. Performance comparison of the proposed method against other state-of-the-art methods on the WLASL2000 dataset

Dataset	Type	Method	P-I (%)	
			Top-1	Top-5
WLASL2000	RGB-based	I3D (Carreira & Zisserman, 2017)	32.48	57.31
		BSL (Albanie et al., 2020)	46.82	79.36
		TimeSformer HR* (Bertasius et al., 2021)	43.66	80.14
	Keypoint-based	SLGTformer (Song & Xiang, 2022)	47.42	79.58
		BEST (Zhao et al., 2023)	46.25	79.33
		ST-GCN (Wang et al., 2022)	34.40	66.57
		LGF-SLR (Gao et al., 2025)	52.74	86.06
		Baseline (L. Hu et al., 2024)	53.68	85.75
		Ours	56.86	88.80

Table 3. Performance comparison of the proposed method against other state-of-the-art methods on the MSASL1000 dataset

Dataset	Type	Method	P-I (%)	
			Top-1	Top-5
MSASL1000	RGB-based	I3D+BLSTM (D. Li et al., 2020)	40.99	-
		BSL (Albanie et al., 2020)	64.71	85.59
	Keypoint-based	ST-GCN (Wang et al., 2022)	36.03	59.92
		HMA (H. Hu et al., 2021)	56.02	76.51
		BEST (Zhao et al., 2023)	58.82	81.18
		Baseline (L. Hu et al., 2024)	65.74	83.20
		Ours	65.88	83.36

Table 4. Performance analysis with different λ on the WLASL100 dataset

λ value	Top-1 accuracy (%)	Top-5 accuracy (%)
0	82.38	96.64
0.1	82.38	96.64
0.3	82.55	95.81
0.5	82.89	95.64
0.7	80.37	96.14
1.0	80.37	95.47

Experimental results indicate that when $\lambda = 0$, the model degenerates into using only basic classification features, thereby losing the advantage of dynamic weight adjustment and achieving relatively low accuracy. As λ increases from 0 to 0.3, the model’s recognition performance shows an upward trend. This indicates that introducing the DARC effectively strengthens the model’s ability to capture key components, thereby improving feature discriminability. When $\lambda = 0.5$, the model achieves its highest Top-1 accuracy of 82.89%. This result validates that during feature fusion, it is necessary to retain some basic classification features while supplementing them with moderate dynamic features. As λ further increases, model performance begins to gradually decline. Notably, when λ reaches 1.0, accuracy shows a significant drop, indicating that excessively high λ values cause the model to over-rely on dynamically generated weights, thereby weakening the expression of basic classification features extracted through multi-layer graph convolutions.

Based on determining the optimal model parameter λ ($\lambda = 0.5$), this paper further investigates the impact of setting the number of groups (groups) and the number of weights per group (kernel) in the DARC module on model performance. Experiments were conducted with different combinations of groups $\in \{1, 2, 4, 8\}$ and kernel $\in \{1, 2, 4, 8\}$, training models, and evaluating performance on the validation set to observe the effects of groups and kernel on dynamic weighting.

As shown in Table 5, first fixing kernel=1 reveals that when groups=1, model performance is relatively low. As the number of groups increases to moderate values (e.g., 4), model performance improves significantly, indicating that moderate grouping effectively suppresses redundant and weakly correlated components in input features. The accuracy reaches its optimum when groups=8.

Table 5. Performance analysis with different groups on the WLASL100 dataset

Groups	Kernel	WLASL100	
		Top-1 Accuracy (%)	Top-5 Accuracy (%)
1	1	80.54	96.14
2	1	80.37	95.47
4	1	82.05	96.98
8	1	82.89	95.64

As shown in Table 6, with groups=8 fixed, model performance deteriorates noticeably as kernel increases from 1 to 4, with Top-1 accuracy on the WLASL100 dataset dropping from 82.89% to 80.03%. While accuracy improves at kernel=8, kernel=1 still yields the best performance. These results demonstrate the critical importance of the grouping mechanism and the number of weights per group within the DARC module. Reasonably tuned groups and kernel parameters can effectively enhance model performance.

Table 6. Performance analysis with different kernel on the WLASL100 dataset

Groups	Kernel	WLASL100	
		Top-1 accuracy (%)	Top-5 accuracy (%)
8	1	82.89	95.64
8	2	81.71	96.31
8	4	80.03	95.13
8	8	81.71	95.81

VISUALIZATIONS

To verify the effectiveness of the proposed DARC module in alleviating the imbalance in feature contributions, t-SNE was employed to visualize the dimensionality reduction of the high-dimensional features extracted by the model. A comparative analysis was conducted on the distribution changes before and after feature enhancement, as illustrated in Figure 4.

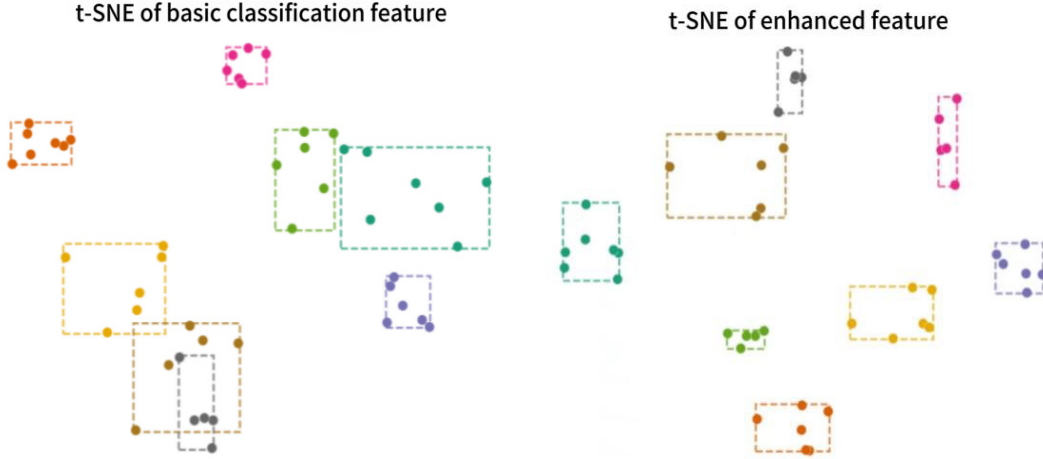


Figure 4. t-SNE plot of WLASL100 dataset, comparing basic classification feature with enhanced feature

Before introducing DARC, for the basic classification feature, due to differences in motion amplitude and speed among different sign language users, there was a certain degree of overlap between different categories, resulting in a blurred decision boundary. Some category samples exhibited significant discretization in the feature space. This discrete distribution reflected that the features extracted by the model were mixed with a large number of redundant and weakly correlated components, making it difficult for samples of the same category to form stable semantic representations.

For the enhanced feature after introducing DARC, various categories form clearer cluster structures in the feature space, with significantly increased inter-class separation and significantly reduced overlapping regions. The distribution of samples within the same category is more concentrated. This indicates that DARC can strengthen key components, suppress feature components with low discriminative contribution, reduce interference from redundant and weakly correlated components, and enhance the consistency of feature representation.

To further verify the effectiveness of DARC under different input modalities, the model was independently trained on joint stream (Joint), bone stream (Bone), joint motion stream (JM), and bone motion stream (BM), and compared with the baseline method. The changes in Top-1 accuracy during the training process are shown in Figure 5.

Overall, DARC outperforms the baseline model on most input streams, achieving the best accuracy of 77.68%, 44.97%, and 43.62% on Joint, JM, and BM streams, respectively, with improvements of 3.02%, 3.53%, and 6.71% compared to the baseline. The most significant improvements are observed on the dynamically related JM and BM streams. This indicates that the proposed DARC can effectively act on temporal dynamic features by adjusting the dynamic weights of different feature components, suppressing noise interference from differences in action amplitude and execution speed, thereby enhancing key components and improving the discriminative ability of the features.

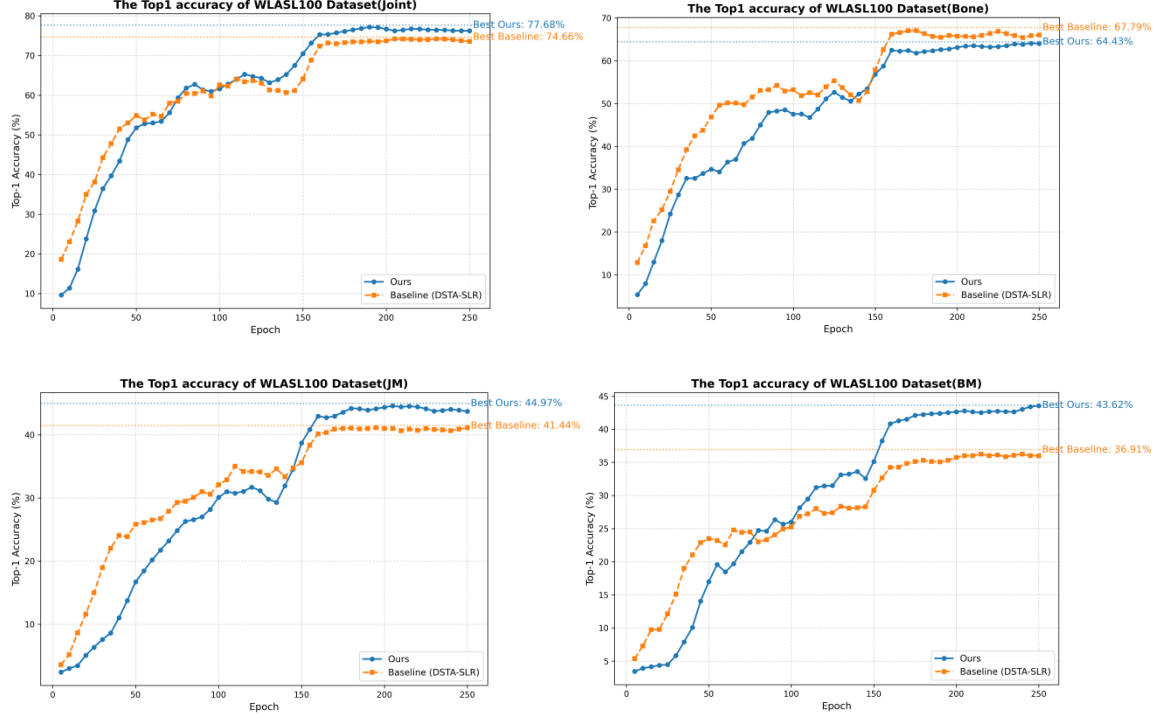


Figure 5. Comparison of Top-1 accuracy over training epochs between the proposed method and the baseline across four input streams (Joint, Bone, JM, and BM) on the WLASL100 dataset

From the trend of the curve, the convergence speed of the proposed method is generally consistent with the baseline in the early stages of training but gradually surpasses it in the later stages and reaches a higher performance ceiling. This indicates that the advantage of the proposed method is primarily reflected in the continuous optimization of feature representation capabilities, rather than in simple convergence acceleration. In contrast, the baseline method tends to encounter performance bottlenecks in later stages, with improvement saturating. Furthermore, DARC performs slightly lower than the baseline on the Bone stream, due to the limited redundant information contained in the skeletal representation, which restricts the action space of DARC. However, in input streams with richer information, this module’s mitigating effect on the imbalance in feature contributions becomes more pronounced.

DISCUSSION

The starting point of this paper is that the deep features extracted from a sequence of keypoints comprise multiple components with varying discriminative contributions, among which redundant and weakly correlated components can interfere with model decision-making (Z. Li et al., 2025). The role of DARC lies in dynamically adjusting the weights of these feature components, enhancing the contribution of key components, and thereby alleviating the imbalance in feature contribution.

Compared with existing methods, DARC achieves superior performance on both WLASL2000 and MSASL1000 datasets. Based on this result, it is not difficult to see that modeling the imbalance in feature contributions can also improve performance, compared to relying solely on improvements in structural or temporal modeling capabilities. The method proposed in this paper does not rely on ad-

ditional complex structures; instead, it adjusts the dynamic weights of deep feature components, indicating that the effectiveness of these components has a significant impact on the final recognition results (Cao et al., 2026).

In sign language recognition, different performers exhibit differences in motion amplitude, speed, and other aspects, leading to a more dispersed distribution of features within the same category. Combined with the t-SNE visualization results, the introduction of DARC reveals a clearer inter-class boundary in feature distribution compared to the case without DARC, while intra-class samples are more concentrated. This indicates that by adjusting the weights of feature components, the model can suppress the interference from redundant and weakly correlated components, paying more attention to consistent discriminative information. Consequently, samples within the same category are distributed more densely in high-dimensional space, and the discriminative boundary between different categories becomes clearer, thereby enhancing overall classification performance. This also provides a good explanation for the improvement in accuracy from the perspective of feature distribution optimization.

In the multi-stream training framework, DARC achieves overall performance improvement under different input streams (joint, bone, joint-motion, bone-motion). These four input streams depict sign language information from both spatial-structure and temporal-dynamics perspectives, and their fusion process is essentially a superposition of multiple feature distributions. According to relevant research, multi-stream integration is prone to introducing redundant or even conflicting information when lacking constraints (Bai et al., 2025). However, after introducing DARC, the performance improvement from multi-stream integration becomes more pronounced, effectively alleviating information redundancy in the fusion process.

Based on current research trends, related studies are increasingly focusing on reducing feature redundancy or enhancing feature effectiveness. For instance, some research has highlighted the significant issue of redundant features in skeleton sequence modeling and proposed corresponding optimization strategies (Gunasekara et al., 2024), while other studies have sought to improve feature representation through channel recalibration (Zang et al., 2023). Compared to these methods, the proposed method in this paper does not rely on additional supervisory information or complex training strategies. Instead, it calibrates deep features within the model, achieving stable performance improvement while maintaining a concise overall structure.

For the development of practical sign language recognition systems, this study also has significant practical value. DARC can be integrated as a lightweight feature-calibration module into existing skeleton-based sign language recognition models, allowing developers to incorporate it without redesigning the overall network architecture; they should consider a reasonable integration scheme. This modular design offers good portability and can be conveniently applied to various skeleton encoders, multi-stream fusion frameworks, and actual deployment systems. Meanwhile, in application scenarios where computational resources are limited yet high recognition accuracy must be maintained, deep feature calibration is a practical approach to model optimization.

Nevertheless, the method proposed in this paper still has certain limitations. The DARC module introduces additional computational overhead, which may require further optimization in resource-constrained scenarios. Secondly, current methods primarily focus on deep feature calibration arising from differences in the contributions of various feature components, without explicitly modeling temporal dependencies over long time spans. For longer sign language sequences, the model needs to continuously retain early-action information and establish associations between temporal segments. However, as sequence length increases, such long-term dependencies often become more difficult to model effectively, which may adversely affect the recognition performance for complex continuous actions. Long-sequence modeling methods, such as Transformers, can leverage self-attention mechanisms to establish connections between distant time steps and thus hold promise for further improving the model's ability to capture long-term temporal information (Kilic et al., 2025). Future work

can explore integrating temporal modeling methods while maintaining the method’s simplicity. Additionally, the model has extremely high requirements for keypoint sequences, and its performance is greatly limited in complex backgrounds or severe occlusions. Future work could consider incorporating other modalities, such as RGB images.

In summary, the effectiveness of the method proposed in this paper is not only reflected in the overall performance improvement but also in the stable contribution of each component to the final result. By dynamically adjusting the weights of deep feature components, the imbalance in feature contributions can be effectively alleviated, thereby enhancing the model’s discriminative ability. This further demonstrates that dynamically adjusting deep feature components is an effective and interpretable optimization direction for sign language recognition tasks.

CONCLUSION

This study addresses the limitation of existing sign language recognition methods in adapting to inter-signer variability by proposing a DARC module for adaptive calibration of deep feature representations. The proposed module effectively enhances the informative components of high-dimensional semantic features while suppressing interference from weakly correlated components. Experimental results on two large-scale American Sign Language datasets, WLASL2000 and MSASL1000, demonstrate that the proposed method consistently outperforms existing state-of-the-art RGB-based and keypoint-based approaches, validating its effectiveness in improving sign language recognition performance.

The findings of this study indicate that adaptive feature modeling for inter-signer variability is a promising direction for improving large-scale sign language recognition. The proposed method enhances the generalization capability of recognition models in complex application scenarios and provides a new perspective for developing more accurate, robust, and practical sign language recognition systems. As related technologies continue to advance, approaches of this kind have the potential to reduce performance disparities across different signers further, thereby facilitating the deployment of sign language recognition in real-world applications such as assistive communication, accessible information services, education, and public services. Ultimately, these advances can contribute to more convenient and efficient communication support for deaf and hard-of-hearing communities.

Nevertheless, this study still has several limitations. The performance of the proposed method remains highly dependent on the accuracy of pose estimation, and its robustness to severe occlusion and complex backgrounds requires further improvement. Future work will focus on exploring multi-modal feature fusion among keypoint sequences, RGB images, and other complementary modalities to enhance further recognition performance and model generalization in challenging real-world environments.

REFERENCES

- Al Abdullah, A., Amoudi, G. A., & Alghamdi, H. S. (2024). Advancements in sign language recognition: A comprehensive review and future prospects. *IEEE Access*, 12, 128871–128895. <https://doi.org/10.1109/access.2024.3457692>
- Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., & Zisserman, A. (2020). BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In A. Vedaldi, H. Bischof, T. Brox & J. M. Frahm (Eds.), *Computer Vision – ECCV* (pp. 35–53). Springer. https://doi.org/10.1007/978-3-030-58621-8_3
- Bai, Y., Yang, D., Xu, J., Xu, L., & Wang, H. (2025). EHC-GCN: Efficient hierarchical Co-occurrence graph convolution network for skeleton-based action recognition. *Applied Sciences*, 15(4), 2109. <https://doi.org/10.3390/app15042109>

- Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *Proceedings of the 38th International Conference on Machine Learning*, 139, 813–824. <https://proceedings.mlr.press/v139/bertasius21a.html>
- Cao, H., Wang, Y., & Wang, T. (2026). Dynamic-static partitioning mask and multi-dimensional attention mechanism for skeleton-based action recognition. *Neural Processing Letters*, 58, Article 12. <https://doi.org/10.1007/s11063-025-11790-6>
- Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the Kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, 6299–6308. <https://doi.org/10.1109/CVPR.2017.502>
- Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: Attention over convolution kernels. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA*, 11030–11039. <https://doi.org/10.1109/CVPR42600.2020.01104>
- Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., & Lu, H. (2020). Decoupling GCN with DropGraph module for skeleton-based action recognition. In A. Vedaldi, H. Bischof, T. Brox & J. M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 536–553). Springer. https://doi.org/10.1007/978-3-030-58586-0_32
- Chi, H.-G., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., & Ramani, K. (2022). InfoGCN: Representation learning for human skeleton-based action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 20154–20164. <https://doi.org/10.1109/cvpr52688.2022.01955>
- Comivi Alowonou, K., & Han, J.-H. (2024). MSA-GCN: Exploiting multi-scale temporal dynamics with adaptive graph convolution for skeleton-based action recognition. *IEEE Access*, 12, 193552–193563. <https://doi.org/10.1109/access.2024.3520172>
- Darmayunata, Y., Setyanto, A., Utami, E., & Kusriani. (2025). Advances and challenges in gesture recognition and sign language recognition: A comprehensive review. *Proceedings of the 31st International Conference on Mechatronics and Machine Vision in Practice, Taguig, Philippines*, 191–196. <https://doi.org/10.1109/M2VIP67511.2025.11165688>
- Deng, Z., Leng, Y., Chen, J., Yu, X., Zhang, Y., & Gao, Q. (2024). TMS-Net: A multi-feature multi-stream multi-level information sharing network for skeleton-based sign language recognition. *Neurocomputing*, 572, 127194. <https://doi.org/10.1016/j.neucom.2023.127194>
- Deng, Z., Leng, Y., Hu, J., Lin, Z., Li, X., & Gao, Q. (2024). SML: A Skeleton-based multi-feature learning method for sign language recognition. *Knowledge-Based Systems*, 301, Article 112288. <https://doi.org/10.1016/j.knosys.2024.112288>
- Gai, Z., Das, T., & Choi, K. (2025). DRGNet: Enhanced VVC reconstructed frames using dual-path residual gating for high-resolution video. *Sensors*, 25(12), 3744. <https://doi.org/10.3390/s25123744>
- Gao, Q., Zhang, M., & Ju, Z. (2025). LGF-SLR: Hand local-global fusion network for skeleton-based sign language recognition. *IEEE Sensors Journal*, 25, 8586–8597. <https://doi.org/10.1109/jsen.2025.3527198>
- Guan, M., Wang, Y., Ma, G., Liu, J., & Sun, M. (2025). MSKA: Multi-stream keypoint attention network for sign language recognition and translation. *Pattern Recognition*, 165, Article 111602. <https://doi.org/10.1016/j.patcog.2025.111602>
- Gunasekara, S. R., Li, W., Yang, J., & Ogunbona, P. O. (2024). Asynchronous joint-based temporal pooling for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1), 357–366. <https://doi.org/10.1109/TCSVT.2024.3465845>
- Hu, H., Zhou, W., & Li, H. (2021). Hand-model-aware sign language recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2), 1558–1566. <https://doi.org/10.1609/aaai.v35i2.16247>
- Hu, J., Shen, L., Albanie, G., Sun, G., & Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), 2011–2023. <https://doi.org/10.1109/tpami.2019.2913372>
- Hu, L., Gao, L., Liu, Z., & Feng, W. (2024). Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (pp. 5450–5460). ELRA and ICCL. <https://doi.org/10.63317/2sjtkug9vfy5>

- Joze, H. R. V., & Koller, O. (2018). *MS-ASL: A large-scale data set and benchmark for understanding American sign language*. PsyArXiv. <http://arxiv.org/abs/1812.01053>
- Khan, A., Jin, S., Lee, G. H., Arzu, G. E., Dang, L. M., Nguyen, T. N., Choi, W., & Moon, H. (2025). Deep learning approaches for continuous sign language recognition: A comprehensive review. *IEEE Access*, 13, 55524-55544. <https://doi.org/10.1109/ACCESS.2025.3554046>
- Kilic, U., Karadag, O. O., & Ozyer, G. T. (2025). AGMS-GCN: Attention-guided multi-scale graph convolutional networks for skeleton-based action recognition. *Knowledge-Based Systems*, 311, 113045. <https://doi.org/10.1016/j.knosys.2025.113045>
- Lee, J., Lee, M., Lee, D., & Lee, S. (2023). Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France*, 10410-10419. <https://doi.org/10.1109/ICCV51070.2023.00958>
- Li, D., Opazo, C. R., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass, CO, USA*, 1448-1458. <https://doi.org/10.1109/WACV45572.2020.9093512>
- Li, Z., Li, F., & Hua, G. (2025). Dynamic graph attention network for skeleton-based action recognition. *Applied Sciences*, 15(9), 4929. <https://doi.org/10.3390/app15094929>
- Liu, J., He, X., Liu, P., Shi, J., Zang, H., & Hao, Y. (2026). *Motion reinforces appearance: RGB-skeleton gated residual fusion for micro-gesture online recognition*. PsyArXiv. <https://doi.org/10.48550/arXiv.2606.11645>
- Liu, Z., Zhang, H., Chen, Z., Wang, Z., & Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA*, 140-149. <https://doi.org/10.1109/CVPR42600.2020.00022>
- Min, Y., Yang, Y., Jiao, P., Nan, Z., & Chen, X. (2025). A closer look at skeleton-based continuous sign language recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Honolulu, HI, USA*, 4968-4974. <https://doi.org/10.1109/ICCVW69036.2025.00515>
- Özdemir, O., Baytaş, İ. M., & Akarun, L. (2023). Multi-cue temporal modeling for skeleton-based sign language recognition. *Frontiers in Neuroscience*, 17, 1148191. <https://doi.org/10.3389/fnins.2023.1148191>
- Pereira, L. M., Salazar, A., & Vergara, L. (2023). On comparing early and late fusion methods. In I. Rojas, G. Joya, & A. Catala (Eds.), *Advances in Computational Intelligence* (pp. 365-378). Springer. https://doi.org/10.1007/978-3-031-43085-5_29
- Qiu, Z.-X., Zhang, H.-B., Deng, W.-M., Du, J.-X., Lei, Q., & Zhang, G.-L. (2022). Effective skeleton topology and semantics-guided adaptive graph convolution network for action recognition. *The Visual Computer*, 39(5), 2191-2203. <https://doi.org/10.1007/s00371-022-02473-7>
- Renjith, S., Sumi Suresh, M. S., & Rashmi, M. (2025). An effective skeleton-based approach for multilingual sign language recognition. *Engineering Applications of Artificial Intelligence*, 143, 109995. <https://doi.org/10.1016/j.engappai.2024.109995>
- Savarese, P., & Figueiredo, D. (2017). Residual gates: A simple mechanism for improved network optimization. *In Proc. Int. Conf. Learn. Representations*. <https://api.semanticscholar.org/CorpusID:211093189>
- Seaman, P. D., Kyle, J. G., & Woll, B. (1988). Sign language: The study of deaf people and their language. *Language*, 65(4), 898. <https://doi.org/10.2307/414977>
- Sengupta, A., Jin, F., Zhang, R., & Cao, S. (2020). mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sensors Journal*, 20(17), 10032-10044. <https://doi.org/10.1109/jsen.2020.2991741>
- Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied sciences*, 12(18), 8972. <https://doi.org/10.3390/app12188972>
- Song, N., & Xiang, Y. (2022). *SLGTformer: An attention-based approach to sign language recognition*. PsyArXiv. <https://doi.org/10.48550/arXiv.2212.10746>
- Tran Anh, V. H., Le Duc, Q., & Nguyen, T.-O. (2025). Efficient sign language recognition with skeleton data: A study of keypoint selection, pose estimators, and GCN models. *Proceedings of the International Conference*

on *Multimedia Analysis and Pattern Recognition*, Khanh Hoa, Vietnam, 1-6.

<https://doi.org/10.1109/MAPR67746.2025.11133720>

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 5998-6008.
- Vázquez-Enríquez, M., Alba-Castro, J. L., Docío-Fernández, L., & Rodríguez-Banga, E. (2021). Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3462–3471). IEEE.
<https://doi.org/10.1109/CVPRW53098.2021.00385>
- Wang, Q., Zhang, K., & Asghar, M. A. (2022). Skeleton-based ST-GCN for human action recognition with extended skeleton graph and partitioning strategy. *IEEE Access*, 10, 41403–41410.
<https://doi.org/10.1109/access.2022.3164711>
- Wong, R., Camgoz, N. C., & Bowden, R. (2025). Signrep: Enhancing self-supervised sign representations. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22804-22814). IEEE.
<https://doi.org/10.1109/ICCV51701.2025.02117>
- Wu, K., Li, Z., Zhao, W., Hu, H., Zhou, W., & Li, H. (2025). Cross-modal consistency learning for sign language recognition. *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 4117–4126). IEEE.
<https://doi.org/10.1109/CVPRW67362.2025.00392>
- Zang, T., Tu, J., Duan, M., & Liu, L. (2023). Skeleton-based action recognition with multi-stream, temporal-channel enhanced graph convolution network. *Proceedings of the 9th International Conference on Computer and Communications, Chengdu, China*, 2037–2041. <https://doi.org/10.1109/ICCC59590.2023.10507583>
- Zhao, W., Hu, H., Zhou, W., Shi, J., & Li, H. (2023). BEST: BERT pre-training for sign language recognition with coupling tokenization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3), 3597–3605.
<https://doi.org/10.1609/aaai.v37i3.25470>
- Zuo, R., Wei, F., & Mak, B. (2023). Natural language-assisted sign language recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14890–14900). IEEE.
<https://doi.org/10.1109/CVPR52729.2023.01430>

AUTHORS



Ying Ma received her PhD degree in Computer Science and Engineering from Chonnam National University, Korea, in 2023. She is currently a lecturer with the School of Software, Nanchang Hang Kong University. Her research interests include artificial intelligence, data science, deep learning, and computer vision.



Nan Lin is currently pursuing her Master's degree in Software Engineering at Nanchang Hang Kong University, Nanchang, China. Her current research interests include deep learning, computer vision, and artificial intelligence.