Dual-View Sign Language Recognition via Front- View Guided Feature Fusion for Automatic Sign Language Training

Siyuan Jing * and Gaorong Yan

Posted Date: 20 January 2026

doi: 10.20944/preprints202601.1551.v1

Keywords: sign language recognition; Chinese national sign language training; human–machine interaction; front-view guided fusion; deep neural network

 doi:10.20944/preprints202601.1551.v1

Article

Dual-View Sign Language Recognition via Front-View Guided Feature Fusion for Automatic Sign Language Training

Siyuan Jing [1,2]* and Gaorong Yan [1]

[1] Sichuan Province Key Laboratory of Philosophy and Social Science for Language Intelligence in Special Education, Leshan Normal University, Leshan China, 614000

[2] Key Laboratory of Internet Natural Language Intelligent Processing of Sichuan Provincial Education Department, Leshan Normal University, Leshan China, 614000

* Correspondence: syjing628@126.com

Abstract

The foundation of an automatic sign language training (ASLT) system lies in word-level sign language recognition (WSLR), which refers to the translation of captured sign language signals into sign words. However, two key issues need to be addressed in this field: (1) the number of sign words in all public sign language datasets is too small, and the words do not match real-world scenarios, and (2) only single-view sign videos are typically provided, which makes solving the problem of hand occlusion difficult. In this work, we design an efficient algorithm for WSLR which is trained on our recently released NationalCSL-DP dataset. The algorithm first performs frame-level alignment of dual-view sign videos. A two-stage deep neural network is then employed to extract the spatiotemporal features of the signers, including hand motions and body gestures. Furthermore, a front-view guided early fusion (FvGEF) strategy is proposed for effective fusion of features from different views. Extensive experiments were carried out to evaluate the algorithm. The results show that the proposed algorithm significantly outperformed existing dual-view sign language recognition algorithms and that compared with the state-of-the-art algorithm, the recognition accuracy was improved by 10.29%.

Keywords: sign language recognition; Chinese national sign language training; human–machine interaction; front-view guided fusion; deep neural network

1. Introduction

Sign language teaching is an important task in modern special education because sign language serves not only as a communication tool but also as a key to unlocking educational and social opportunities for the estimated 466 million deaf individuals around the world. However, sign language teaching is confronted with three major challenges: (1) the mastery of sign language is hindered by its intricate hand gestures and dynamic movements; (2) the absence of communication partners and authentic training environments amplifies the challenges in the learning process; and (3) a notable gap exists in the availability of sign language instructors, coupled with a dearth of digital teaching resources. Consequently, the development of an automatic sign language training (ASLT) system is imperative.

The foundation of an ASLT system is rooted in sign language recognition (SLR) technology. SLR can be broadly categorized into word-level sign language recognition (WSLR) [1,10,11] and continuous sign language recognition (CSLR) [14,15,24] on the basis of the recognition objective. The key distinction is that the former focuses on translating sign language signals into individual sign

 doi:10.20944/preprints202601.1551.v1

words (glosses), whereas the latter aims to decode signals into sequential glosses. Given our objective of enabling learners to master sign language vocabulary proficiently, this study focuses exclusively on WSLR. In terms of technical approach, SLR can also be divided into wearable-based and vision-based SLR [2,42,45]. Wearable-based SLR employs data gloves (e.g., EMG sensors) to capture fine-grained hand motions and body gestures. This approach offers advantages such as high precision in motion feature extraction and robustness to background noise. However, it requires users to wear specialized devices, which compromises usability and limits real-world applicability. In contrast, vision-based SLR uses RGB/depth cameras to capture visual signals of signers' movements and leverages computer vision and deep learning for recognition [55,56]. Owing to its noninvasive nature and superior user experience, this type of SLR has become the dominant trend in the past decade.

Recently, vision-based SLR has undergone substantial advancements, driven by the proliferation of sign language datasets and the rapid development of deep learning technologies. In the field of WSLR, numerous influential datasets, such as WLASL [28], MS-ASL [22], DEVISIGN [5], AUTSL [43], and NMFs-CSL [16], have been publicly released. However, a critical limitation is that most of these datasets provide only single-view RGB sign videos, which poses significant challenges in addressing hand occlusion during WSLR. As illustrated in Figure 1, hand occlusion occurs frequently in Chinese sign language (CSL) scenarios. The figure displays both front-view and left-view perspectives of a signer. In the front-view videos, the signer's hands are often occluded, which makes extracting hand features (e.g., shape and motion) difficult. In contrast, the left-view videos clearly reveal the signer's two hands, thus highlighting the importance of multiview data for overcoming occlusion issues.

In this paper, we present an efficient WSLR algorithm that is trained on the NationalCSL-DP dataset, which comprises 134,140 sign videos across 6,707 sign words [20]. The algorithm begins with frame-level alignment of dual-view sign videos to ensure temporal consistency between two perspectives. A two-stage deep neural network is then employed to extract the spatiotemporal features of the signers, including hand motion dynamics and body postures. To address the challenge of multiview feature integration, we introduce a front-view guided early fusion (FvGEF) strategy that adaptively weights features from different views to increase recognition accuracy. The theoretical basis for this design is explained in detail in the subsequent sections. Extensive experiments conducted on NationalCSL-DP demonstrate that the proposed algorithm significantly outperforms existing dual-view WSLR methods. Specifically, it achieved a 10.29% improvement in recognition accuracy compared with the state-of-the-art algorithm.

 doi:10.20944/preprints202601.1551.v1

Figure 1. Examples of sign words {protection, air-conditioner, forced, and edit} in CSL with hand occlusion.

The contributions of this study include the following:

(1) We present a novel WSLR algorithm tailored for the NationalCSL-DP dataset, which serves as a foundational component for developing an ASLT system. The algorithm leverages a two-stage deep neural network architecture as its backbone.

(2) We introduce two key strategies: (a) a temporal frame-level alignment method for dual-view sign videos and (b) a front-view guided early fusion (FvGEF) strategy to enhance cross-view feature integration and improve recognition accuracy.

(3) Comprehensive experiments validate the algorithm's efficacy and efficiency. The results demonstrate a 10.29% improvement in recognition accuracy compared with state-of-the-art methods, thus underscoring the practical utility of our approach.

The remainder of the paper is organized as follows. Section 2 reviews related work on WSLR and multiview action recognition. In Section 3, we describe the dual-view Chinese sign language training system and the NationalCSL-DP dataset. Section 4 explains the proposed algorithm in detail. Section 5 presents the experiments and provides an in-depth analysis. Finally, Section 6 concludes the paper and outlines future directions.

2. Related Works

2.1. Word-Level Sign Language Recognition

The existing WSLR methods can be categorized into three paradigms according to the input modality used: vision-based, pose-based, and multimodal approaches.

A. Vision-Based Approaches

Inspired by breakthroughs in action recognition using convolutional neural networks (CNNs), vision-based methods have become a dominant direction in the field of WSLR. Various works have employed CNNs to extract frame-level visual features (e.g., hand shapes and body postures), followed by recurrent neural networks (RNNs) or temporal convolutional networks (TCNs) to model temporal dynamics in sign sequences [10,37,55]. With the rise of attention mechanisms, transformers have emerged as powerful alternatives for sequence modeling. The vision transformer (ViT) [9] has been increasingly adopted for spatial feature extraction, whereas video-specific architectures such as

 doi:10.20944/preprints202601.1551.v1

S3D [52], MViT [30], and ViViT [3] have demonstrated superior performance in capturing spatiotemporal dependencies. However, most of them cannot address the hand occlusion issue.

B. Pose-Based Approaches

In contrast with vision-based methods, pose-based WSLR models excel in terms of robustness to background clutter, lighting changes, and occlusions by explicitly encoding human kinematic data (e.g., hand joints and limb trajectories). These approaches typically leverage pretrained human pose estimation frameworks, including OpenPose [35], MediaPipe [33], and MMPose [40], to extract skeletal keypoints [2,13,47]. Subsequently, spatiotemporal graph neural networks (ST-GNNs) are employed to model the structural and temporal dependencies within the skeletal sequences to explicitly capture the kinematic relationships between hand joints and body limbs. This framework enables the encoding of dynamic sign language features, such as gesture trajectories and joint movements, while mitigating the effects of background noise and lighting variations [29]. A seminal approach, SL-GCN, uses spatiotemporal graph convolutional networks (ST-GCNs) to model sign poses as dynamic graphs, where nodes represent anatomical joints and edges encode spatial–temporal dependencies. Subsequent studies have expanded this framework by integrating transformers (e.g., pose transformers) to enhance long-range temporal modeling and have demonstrated improved performance in noisy environments [14,53]. Hu et al. proposed a self-supervised pretrained framework named SignBERT+ that incorporates a pose-based model-aware hand prior to enhance the performance in sign language understanding tasks [15]. However, current pose data are derived by general-domain pose estimation tools, which may inevitably introduce pose estimation errors and result in undetected keypoints due to motion blur and hand occlusion in sign videos. Besides that, some works exploit multimodal approaches to address this issue [7,19.56]. However, applying this paradigm to multiview SLR remains challenging because current pose estimation tools struggle to retrieve high-fidelity skeletal keypoint data from lateral-view sign videos, where self-occlusion and perspective distortion frequently degrade the detection quality.

2.2. Multiview Action Recognition

The development of effective and robust real-world action recognition (AR) systems necessitates the integration of multiview learning. Recent progress in AR has yielded diverse multiview approaches, including dictionary learning [27], view-adaptive neural networks [4,6], and attention-based architectures [18,21]. For example, ViT models have adopted cuboid-shaped spatial–temporal modeling to handle multiview data robustly. Contemporary methods prioritize learning view-invariant representations for downstream tasks, and view-specific and view-agnostic modules are fused to enable accurate predictions [29,34,36]. Moreover, techniques such as feature factorization and cascaded residual autoencoders address challenges in partial-view classification and incomplete-modality fusion, respectively [46,49,50]. In this paper, we focus on dual-view vision-based WSLR. Previous work has demonstrated that dual-view features, by capturing complementary spatial–temporal dynamics, significantly increase recognition accuracy compared with single-view baselines [12,21]. This highlights the potential of leveraging multiview data to construct more robust and precise WSLR systems, particularly in complex real-world scenarios with occlusions or viewpoint variations.

3. Method for Dual-View Sign Language Training

3.1. A System for Dual-View Sign Language Training

This section presents an ASLT system that leverages dual-view WSLR technology (see Figure 2). The system employs a dual-camera setup: an RGB camera and a teleprompter are positioned in front of the signer, whereas a second RGB camera captures lateral-view footage from the left side. When the signer initiates training in the ASLT studio, the teleprompter displays preprogrammed sign language tasks, thus prompting the signer to perform specified sign words. Concurrently, the dual cameras stream real-time RGB videos of the signer's actions from frontal and lateral views to a server,

 doi:10.20944/preprints202601.1551.v1

where inference via the WSLR model determines whether the performed signs match the prompted vocabulary.

The training and inference pipeline of the WSLR model is illustrated on the right side of Figure 2. Upon receiving dual-view sign videos, the system first performs data preprocessing, including frame sampling, resolution normalization, and temporal alignment, to synchronize cross-view sequences (e.g., to correct for camera latency or motion disparities). The preprocessed frames are then fed into a trained WSLR model to extract and fuse spatiotemporal features from both perspectives to capture hand dynamics and body kinematics from the frontal and lateral views. Finally, the model outputs the recognition results, thus providing real-time feedback on the accuracy of the signer's performance.

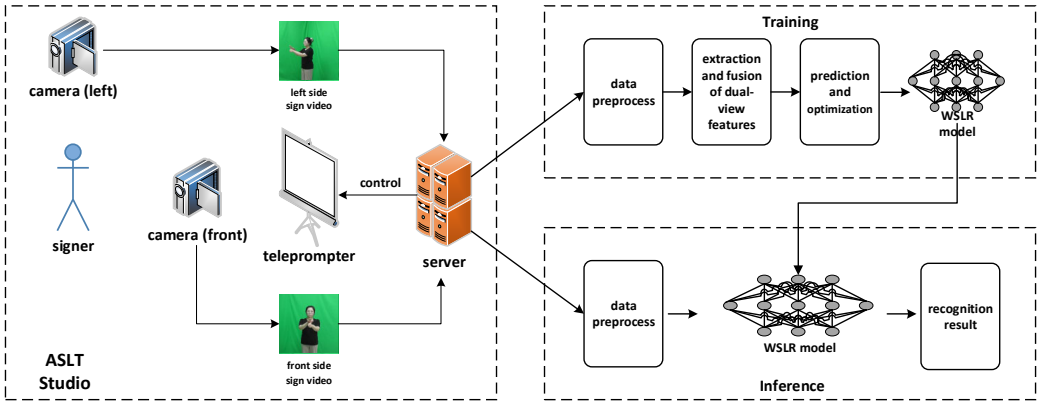

Figure 2. ASLT system based on dual-view WSLR technology.

3.2. NationalCSL-DP Dataset

The NationalCSL-DP dataset is the first sign language dataset that covers the entire Chinese national sign language (CNSL) vocabulary [20]. It contains recordings of ten signers, namely, 2 males and 8 females, with a mean age of 19.82±0.28 years. Among them, 8 were deaf students, and 2 were hearing students, all of whom were highly proficient in CNSL. The dataset contains 6707 Chinese Sign Language words, thus surpassing the vocabulary sizes of existing datasets. In addition, it provides 134140 sign videos from two perspectives. Unlike the recently proposed multiview sign language datasets, i.e., MM-AUSLAN [41] and Multi-VSL [8], the two perspectives in the NationalCSL-DP dataset are vertical. More details can be found in [20] Specifically, we shuffled the sign words and selected the top-*K* words to build the subsets, where K = {200, 500, 1000, 2000, 6707}. Finally, five subsets, named NationalCSL200, NationalCSL500, NationalCSL1000, NationalCSL2000, and NationalCSL6707, are provided.

4. Proposed Algorithm for Dual-View WSLR

4.1. Dual-View Word-Level Sign Language Recognition

The NationalCSL-DP dataset, which contains a large collection of sign videos that comprehensively cover the CNSL vocabulary from two orthogonal views (frontal and lateral), presents significant challenges for recognition. In previous work [21], researchers proposed a CNN+Transformer architecture with a simple plus fusion strategy, which achieved a top-1 accuracy of 69.61% on the NationalCSL6707 dataset. This approach involves training two independent deep neural networks on frontal and lateral sign videos. During inference, features from the two models are fused via a plus operator before final classification. However, the study revealed that early fusion strategies failed to perform effectively on NationalCSL-DP. The analytical results suggested that camera latency-induced frame misalignment between the two views might have been the root cause.

 doi:10.20944/preprints202601.1551.v1

To address this issue, this paper first introduces frame-level alignment (see Figure 3) to synchronize cross-view frames before feature extraction, which is similar to [44]. Additionally, in consideration of the biomechanical characteristics of sign language, where the right hand typically serves as the dominant executor and the left hand as an auxiliary, we hypothesize that the frontal view plays a primary role in WSLR, whereas the lateral view acts as a complementary source during hand occlusion events. The proposed algorithm pipeline (see Figure 3) begins with frame-level alignment, which identifies dynamic onset frames by signer movement tracking to mitigate temporal misalignment. A novel front-view guided early fusion (FvGEF) strategy is then introduced. Spatial feature extractors first capture visual cues from frontal and lateral videos independently, after which cross-view features are fused under FvGEF. Finally, a transformer module models the temporal dynamics of sign actions to generate sequential feature representations. Detailed algorithmic descriptions are provided in the subsequent sections.

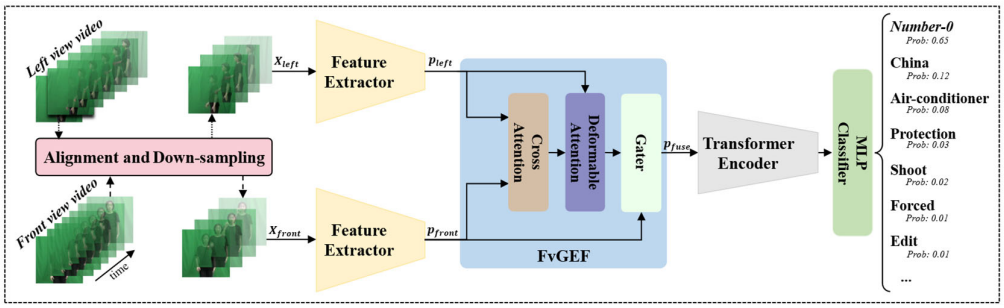

Figure 3. Proposed dual-view WSLR recognition framework. *Prob* represents the probability of the results.

4.2. An Efficient Algorithm for WSLR

A. Framework

As depicted in Fig. 3, our framework begins with frame-level alignment and downsampling of sign videos across dual views. The synchronized frames are then channeled into dedicated spatial feature extractors for each view. The outputs from these extractors are fused via a novel front-view guided early fusion (FvGEF) module to generate discriminative cross-view spatial features. These features are subsequently processed by a transformer encoder to extract temporal dynamics. Finally, an MLP classifier is used to produce the final prediction results.

B. Frame-level alignment

When constructing the NationalCSL-DP dataset, two assistants independently operated the frontal and lateral cameras for recording, which inherently introduced asynchrony between the cross-view sign videos. The authors directly fused unaligned spatial features using an early fusion strategy, which surprisingly yielded lower recognition accuracy than did single-view WSLR models on the NationalCSL-DP dataset [21]. Although the two orthogonal RGB cameras in the proposed ASLT system are synchronized via signal control, residual cross-view asynchrony (e.g., camera latency) may still persist. To address this issue, we introduce a frame-level alignment method. The core principle is that, regardless of the initial synchronization status between the two views, the precise detection of sign motion onset frames from both perspectives provides a robust solution for mitigating video asynchrony in sign language recognition [26].

As shown in Fig. 3, the input of the algorithm is a dual-view sign video, which is denoted as $V = \{f_t\}_{t=1}^{T} \in R^{T\times 3\times H\times W}$, where T, H, and W represent the number of frames, height and width, respectively, of each sign video. To obtain the precise location of the sign motion onset frame, the frames are first converted to grayscale and then subjected to Gaussian blurring to suppress high-frequency noise, which enhances the reliability of motion estimation. We subsequently compute the absolute pixelwise differences between consecutive frames to generate a motion intensity map for

 doi:10.20944/preprints202601.1551.v1

each frame [23,25]. The global motion intensity is then quantified by calculating the spatial average of each frame's intensity map, which yields a scalar motion intensity value m for each frame, and the values for all frames form a motion intensity sequence $M = \{m_1, m_2, \ldots, m_{T-1}\}$.

For determining the onset frame in a sign video, we introduce an adaptive threshold strategy that is grounded in robust statistics of the motion intensity distribution. Specifically, the threshold τ is defined as the sum of the median (*Med*) and median absolute deviation (*MAD*), where *Med* represents the median of the motion intensity sequence, and *MAD*, which is a robust measure of dispersion, is calculated as the difference between the median of the motion intensity *M* and *Med* [39]. This adaptive threshold effectively mitigates variations in background motion and illumination, thereby enabling accurate onset detection across diverse scenarios.

Additionally, to mitigate local fluctuations further, we apply 1D Gaussian smoothing to the motion intensity sequence M prior to detection. To ensure robust onset localization, the starting index is defined as the earliest time index where the motion intensity stays consistently above the adaptive threshold τ for a minimum of *B* consecutive frames:

$$\text{starting\_index} = \min\{i \in \mathcal{T} \mid \widehat{M}[j] > \tau,\ \forall\, j \in [i,\ i + B]\} \quad (1)$$

where $\mathcal{T} = \{1,2,\ldots,T\}$ refers to the set of all frame indices, $\widehat{M}$ denotes the Gaussian-smoothed motion sequence, and *B* denotes the minimum required duration of motion for down-sampling in the algorithm, which is set to 5 in this study.

Within our framework, we perform frame-level alignment on both the front-view and lateral-view videos of the NationalCSL-DP dataset, thus generating separate starting indices for the front and lateral views. This process effectively alleviates the effect of temporal misalignment and ensures consistent temporal synchronization across viewpoints. Following alignment, we downsample the original sign videos starting from the detected onset frames. This preprocessing step establishes a robust foundation for subsequent feature extraction, cross-view fusion, and recognition tasks.

C. Frontal view-guided early fusion

Understanding the structure of sign language requires going beyond surface-level visual patterns. Unlike spoken language, which relies on a linear auditory stream, sign languages encode meaning through spatially distributed and simultaneous articulations, including handshape, movement, location, palm orientation, and facial expressions. These articulatory parameters form complex morphological units that map directly to linguistic meaning, which is a phenomenon commonly referred to as form-meaning mapping. In particular, the spatial configuration and dynamic trajectory of both hands contribute to disambiguating semantically similar signs.

Despite advances in deep learning-based sign language recognition, many existing methods treat the task as a black-box video classification problem and often overlook the linguistic structure embedded in spatiotemporal expressions. Our proposed dual-view word-level recognition algorithm aims to bridge this gap by modeling spatial cues more faithfully by using synchronized views to recover occluded articulators and better capture spatial relationships.

We hypothesize that semantic consistency must be preserved across views; although the visual input differs between frontal and lateral perspectives, both reflect the same underlying sign semantics. Therefore, any fusion strategy should seek to (i) extract complementary cues from each view and (ii) ensure their semantic coherence in the fused representation. This motivates the use of a front view-guided early fusion (FvGEF) mechanism that adaptively emphasizes or suppresses view-specific features on the basis of their contribution to semantic clarity.

After preprocessing, which results in a front-view word-level sign video with T' frames $X_{front} = \{X_t\}_{t=1}^{T'} \in R^{T' \times 3 \times H \times W}$ and a left-view video $X_{left} = \{X_t\}_{t=1}^{T'} \in R^{T' \times 3 \times H \times W}$, the proposed framework extracts frame-level features using two spatial extractors, S_{front} and S_{left}. Therefore, the framewise representation of dual-view videos $p_{front} = \{p_t\}_{t=1}^{T'} \in R^{T' \times d_s}$ and $p_{left} = \{p_t\}_{t=1}^{T'} \in R^{T' \times d_s}$, where d_s is the spatial representation dimension, is calculated by Equation (2), where $view \in \{front, left\}$.

$$p_{view} = S_{view}(X_{view}) \quad (2)$$

 doi:10.20944/preprints202601.1551.v1

In previous work, we analyzed the mutual information (MI) between individual view features and gloss labels and reported that frontal-view features consistently resulted in higher MI values [21,38]. A comparison of the results of front-view recognition and left-view recognition suggests that frontal-view features carry stronger semantic signals and thus should be weighted more heavily during fusion. Our proposed FvGEF framework operationalizes this insight through a gated attention mechanism, which enables the model to dynamically regulate the contribution of each view at both the spatial and temporal levels. Therefore, $p_{fuse} = \{p_t\}_{t=1}^{T'} \in R^{T' \times d_s}$ is fused by p_{front} and p_{left} through the proposed FvGEF.

Furthermore, $\Phi_{CrossAttention}$ is used to interact p_{front} and p_{left} by applying a cross-attention mechanism in which the *query* originates from the front view and the *key* and *value* are generated from the left view as follows:

$$CA_{Out} = \Phi_{CrossAttention}(p_{front}, p_{left}) \tag{3}$$

By applying $\Phi_{CrossAttention}$, the cross-view feature is first computed to facilitate preliminary semantic fusion between different views. Moreover, deformable attention has been proposed for enhancing higher-level semantic coherence and alleviating misalignment [48]. Therefore, we introduce $\Phi_{DeformableAttention}$ to improve semantic coherence across views while simultaneously mitigating occlusion influence, which is validated to be beneficial in our experiments, thus enabling finer-grained synchronization as follows:

$$DA_{Out} = \Phi_{DeformableAttention}(CA_{Out}, p_{left}) \tag{4}$$

Finally, Φ_{Gater} employs a gating network to learn fusion weights on the basis of the concatenation of p_{front} and the output of $\Phi_{DeformableAttention}$ by implementing a weighting strategy that prioritizes from the front view as follows:

$$\begin{aligned} p_{fuse} &= \Phi_{Gater}(p_{front}, DA_{out}) \\ &= \sigma\left(Linear\left(Cancat(p_{front}, DA_{out})\right)\right) \odot p_{front} \\ &\quad + \left(1 - \sigma\left(Linear\left(Cancat(p_{front}, DA_{Out})\right)\right)\right) \odot DA_{Out} \end{aligned} \tag{5}$$

where σ is the sigmoid function, $Cancat$ is utilized for concatenation, $Linear$ represents the linear transformation layer and $\odot$ represents the Hadamard product. The choice of the front view as the dominant source is grounded in both empirical observations and statistical analysis. From a linguistic standpoint, most signers are trained to produce signs facing forward, thus ensuring that key articulators (handshapes and facial expressions) are visible to an audience or camera positioned in front. Our data indicate that frontal videos generally exhibit higher information density in terms of visual discriminative features (e.g., greater variance in pixel movement and handshape visibility). Moreover, the left-hand view serves as a critical supplement in situations that involve self-occlusion, such as crossing hands, overlapping gestures, or partially obscured facial cues.

Taken together, the theoretical underpinnings of our fusion strategy are grounded in both sign language linguistics and information-theoretic principles; thus, our fusion strategy provides an interpretable and robust approach to multiview WSLR. Through FvGEF, front view features are strengthened, and left view noise is weakened, which are the goals of using the left view from NationalCSL as a supplementary perspective for efficient ISLR.

5. Experiments

In this section, we detail the experimental setup and procedures. First, we evaluate how different deep neural network modules affect the algorithm's performance. Second, we compare the proposed algorithm with state-of-the-art (SOTA) methods to validate its effectiveness. Finally, we perform ablation studies and provide a thorough analysis of the results.

5.1. Implementation Details

 doi:10.20944/preprints202601.1551.v1

Following the protocols of prior studies, each video in our experiments was sampled into 16 frames with a temporal stride of 5 frames. Specifically, the starting frames for each view were detected via the proposed frame-level alignment method, which served as the origins for downsampling. Two separate Swin transformers were then employed to generate 512-dimensional feature vectors for each frame. The proposed FvGEF module was subsequently utilized to perform efficient fusion guided by the front view and supplemented with the left view. To process the sequence in the latent space, we applied a transformer encoder that comprised 4 layers of 8-head attention with a self-attention dimension of 512. Our framework was implemented using PyTorch on CentOS with an NVIDIA 4090D GPU. Training was conducted using the RGD optimizer at a learning rate of 0.01 and terminated when the validation loss plateaued for 5 consecutive epochs.

5.2. Evaluation Metric

The top-*k* accuracy measures the proportion of test samples for which the ground-truth label appears within the model's top-*k* predictions ranked by confidence scores. This metric is particularly well suited for tasks such as WSLR that involve large output spaces with numerous possible classes. Mathematically, top-*k* accuracy is computed as follows:

$$\text{Top} - \text{k accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(y_i \in \hat{Y}_i^k\right) \tag{6}$$

Here, N represents the total number of samples (i.e., videos), and the indicator function I returns 1 if the true label y_i for the i_{th} sample falls into the subset of the top-k predicted labels $\hat{Y}_i^k$ that are output by the inference model.

5.3. Experiments on Different Feature Extractors

To systematically assess how different 2D feature extractors affect our framework's performance, we performed comprehensive experiments on the NationalCSL6707 dataset using diverse architectural designs. We did not conduct experiments on small-scale datasets, as our primary objective was to ensure that the proposed algorithm achieves strong performance on the full dataset. The models evaluated included classic CNNs (e.g., ResNet), vision transformers (ViT-base and ViT-large) [9], and Swin transformers at varying scales (Swin-tiny, Swin-small, and Swin-base) [31,32]. All the models were trained under identical conditions to ensure a fair comparison, with pretraining conducted on ImageNet. As shown in Table 1, significant performance disparities emerged across different feature extraction backbones. Among the CNNs, ResNet established a robust baseline with a top-1 accuracy of 75.09%. Notably, substituting ResNet with vision transformer architectures yielded consistent improvements across all the evaluated metrics. Specifically, ViT-base and ViT-large achieved top-1 accuracies of 76.61% and 77.05%, respectively, thus demonstrating that global attention mechanisms excel at capturing long-range dependencies in sign language sequences.

Among all the evaluated architectures, Swin-small delivered the highest performance, with 80.99% top-1 accuracy, which exceeded that of the second-best network (ViT-large) by more than 3%. This result highlights the effectiveness of hierarchical feature learning combined with local window attention in the Swin transformer, which excels at modeling fine-grained motion patterns and semantic structures in dual-view Chinese WSLR tasks. Notably, increasing the model size does not universally improve the performance. For example, ViT-large slightly outperformed ViT-base, which was likely because of insufficient training data to support the model's capacity. In contrast, Swin-tiny and Swin-small demonstrated a superior balance between model complexity and generalization capability and achieved strong performance with relatively few parameters. Therefore, we ultimately adopted Swin-small as the spatial feature extractor for the proposed algorithm.

Table 1. Experiments on various 2D feature extractors on the NationalCSL6707 dataset. Performance is measured in terms of the top-1, top-5, and top-10 accuracies (%). The best results are highlighted in bold.

 doi:10.20944/preprints202601.1551.v1

Extractor / Metric	ResNet	ViT-base	ViT-large	Swin-base	Swin-tiny	Swin-small
Top-1	75.09	76.61	77.05	75.80	75.98	**80.99**
Top-5	90.80	90.95	90.85	91.20	91.81	**93.22**
Top-10	93.50	93.47	93.19	93.87	94.58	**95.14**

5.4. Comparison with State-of-the-Art Algorithms

In this section, we benchmark our proposed algorithm against several SOTA methods on the NationalCSL dataset to validate its efficacy in the dual-view Chinese WSLR task. The compared approaches include widely used architectures such as S3D, SL-GCN, CNN-Transformer, and MViT, which encompass diverse modeling paradigms that span 3D convolutional neural networks (3DCNNs), graph convolutional networks (GCNs), and video transformers. The experiments were carried out on five datasets, and three metrics were employed, namely, Top-1, Top-5 and Top-10. We performed a thorough evaluation to assess the proposed algorithm's performance against SOTA methods across multiple metrics on datasets with diverse scales. Notably, the CNN-Transformer algorithm was proposed by Jing (Jing, 2025). It employs a plus fusion strategy, where two independent deep neural networks are trained on sign videos from two perspectives. During the recognition phase, the recognition results of the two deep neural networks are integrated, and the final recognition result is output. In the experiments, S3D, SL-GCN, and MViT, which adopt the same strategy as the CNN-Transformer algorithm, were also evaluated.

As shown in Table 2, the proposed algorithm consistently outperformed prior SOTA methods across all evaluation metrics and NationalCSL-DP subsets. Notably, compared with the most competitive baselines (i.e., CNN-Transformer or MViT), our approach achieved significant top-1 accuracy improvements of +3.50% (NationalCSL200), +4.20% (NationalCSL500), +7.80% (NationalCSL1000), +8.80% (NationalCSL2000), and +10.29% (NationalCSL6707). Moreover, the improvements were not limited to the top-1 accuracy; our approach also outperformed the existing models in terms of both the top-5 and top-10 accuracies across most datasets. For example, on NationalCSL6707, our model achieved a top-1 accuracy of 80.99% (a margin of over 10% improvement), alongside a top-5 accuracy of 93.22% and a top-10 accuracy of 95.14%. These results indicate that our framework both identified the correct class more accurately and maintained strong discrimination capabilities among similar categories. These advantages were most pronounced on larger subsets, where the exponential increase in sign language variation complexity amplified the benefits of our algorithm's cross-view fusion and feature learning.

Table 2. Top-K accuracy (%) comparison of the proposed approach and state-of-the-art methods on the NationalCSL-DP dataset. The best results are highlighted in bold.

Dataset	*Metric*	*S3D*	*MViT*	*SL-GCN*	*CNN-Transformer*	*Ours*
NationalCSL200	Top-*1*	47.50	76.50	81.50	85.50	**89.00**
	Top-*5*	84.00	93.50	96.00	**98.00**	97.50
	Top-*10*	90.50	95.00	97.50	98.00	**98.00**
NationalCSL500	Top-*1*	49.00	77.00	80.20	84.00	**88.20**
	Top-*5*	82.40	93.20	91.20	94.20	**95.80**

 doi:10.20944/preprints202601.1551.v1

	Top-*10*	89.40	96.00	94.80	96.60	**97.20**
NationalCSL1000	Top-*1*	73.70	74.60	75.80	80.70	**88.50**
	Top-*5*	92.40	91.20	90.50	93.80	**96.20**
	Top-*10*	95.10	94.70	94.70	96.80	**97.20**
NationalCSL2000	Top-*1*	74.30	73.59	77.10	79.30	**88.10**
	Top-*5*	90.65	91.65	89.45	92.85	**95.50**
	Top-*10*	94.50	94.80	94.40	95.45	**96.55**
NationalCSL6707	Top-*1*	67.62	70.70	66.17	69.61	**80.99**
	Top-*5*	88.30	85.51	84.93	88.92	**93.22**
	Top-*10*	92.50	93.08	88.72	92.93	**95.14**

The success of our framework can be attributed to the effective frame-level alignment and the FvGEF strategy, which enable robust cross-view semantic understanding and better generalization to complex linguistic patterns. In contrast, existing methods often struggle with misalignments between different viewpoints or fail to capture fine-grained motion details due to limitations in their architectural design or modality fusion strategies. In summary, these comparisons clearly demonstrate that our framework establishes a new SOTA on the NationalCSL-DP benchmark, especially for large-scale WSLR tasks, thus highlighting its strong potential for developing an ASLT system.

5.5. Ablation Study

To better understand the contribution of each component in our framework, we conducted a comprehensive ablation study on two key modules, i.e., frame-level alignment and the FvGEF module. As summarized in Table 3, we evaluated the performance of four variants across five subsets of the NationalCSL-DP dataset. The baseline algorithm employs the Swin-small+transformer as the backbone and incorporates an early fusion strategy. The other two variants were the proposed algorithm without frame-level alignment and the proposed algorithm without the FvGEF strategy. The experimental results are shown in Table 3.

***A*. Effectiveness of frame-level alignment**

Table 3 clearly demonstrates the critical role of the frame-level alignment module in boosting recognition accuracy. Across all dataset subsets, the integration of frame-level alignment yielded substantial top-1 accuracy improvements over the baseline. Notably, on the largest subset, NationalCSL6707, the top-1 accuracy increased from 57.49% to 77.04% (+19.55%), thus underscoring the necessity of temporal and semantic synchronization between dual-view frame sequences. Consistent gains in the top-5 and top-10 accuracies further indicate that frame-level alignment enhances both the precision of top-ranked predictions and generalizability across the entire label space. Significantly, the improvement margins increased with the dataset scale, which suggests that frame-level alignment becomes progressively more effective as the complexity and variability of sign language instances escalate. This highlights the module's ability to mitigate viewpoint discrepancies and enhance robustness in high-complexity scenarios.

 doi:10.20944/preprints202601.1551.v1

Table 3. Results of the ablation study on the frame-level alignment and FvGEF modules across different subsets of NationalCSL-DP. The performance gains (↑) and decreases (↓) relative to the baseline are shown in parentheses. The best results are highlighted in bold.

Dataset	Metric	Baseline	w/alignment	w/FvGEF	Ours
NationalCSL200	Top-*1*	78.50	**89.00 (↑10.50)**	81.50 (↑3.00)	**89.00 (↑10.50)**
	Top-*5*	92.00	97.00 (↑5.00)	92.00	**97.50 (↑5.50)**
	Top-*10*	95.00	**98.00 (↑3.00)**	96.50 (↑1.50)	**98.00 (↑3.00)**
NationalCSL500	Top-*1*	77.40	87.80 (↑10.40)	78.20 (↑0.80)	**88.20 (↑10.80)**
	Top-*5*	92.60	**96.20 (↑3.60)**	93.40 (↑0.80)	95.80 (↑3.20)
	Top-*10*	95.80	**97.40 (↑1.60)**	95.60 (↓0.20)	97.20 (↑1.40)
NationalCSL1000	Top-*1*	76.20	88.10 (↑11.90)	77.10 (↑0.90)	**88.50 (↑12.30)**
	Top-*5*	91.10	**96.50 (↑2.40)**	91.60 (↑0.50)	96.20 (↑2.10)
	Top-*10*	94.60	97.10 (↑2.50)	94.40 (↓0.20)	**97.20 (↑2.60)**
NationalCSL2000	Top-*1*	75.10	86.90 (↑11.80)	75.80 (↑0.70)	**88.10 (↑13.00)**
	Top-*5*	90.35	95.45 (↑5.10)	89.90 (↓0.45)	**95.50 (↑5.15)**
	Top-*10*	93.05	**96.65 (↑3.60)**	93.20 (↑0.15)	96.55 (↑3.50)
NationalCSL6707	Top-*1*	57.49	77.04 (↑19.55)	61.37 (↑3.88)	**80.99 (↑23.50)**
	Top-*5*	72.34	88.37 (↑14.02)	77.32 (↑4.54)	**93.22 (↑15.81)**
	Top-*10*	77.41	91.43 (↑10.80)	81.95 (↑3.91)	**95.14 (↑11.98)**

***B.* Effect of the FvGEF module**

In contrast, the impact of the FvGEF module is context sensitive, with its effectiveness highly dependent on the dataset scale and complexity. On large-scale datasets such as NationalCSL6707, FvGEF demonstrated significant improvements: the top-1 accuracy increased from 57.49% to 61.37% (+3.88%), whereas the top-5 and top-10 accuracies increased by 4.54% and 3.91%, respectively. These

 doi:10.20944/preprints202601.1551.v1

results validate the effectiveness of the FvGEF module. Conversely, on small-scale datasets, FvGEF did not improve the performance uniformly across all the metrics. For example, on NationalCSL500, FvGEF decreased the top-5 and top-10 accuracies by 0.8% and 0.2%, respectively, whereas on NationalCSL2000, it decreased the top-5 accuracy from 90.35% to 89.90% (-0.45%). Notably, however, FvGEF consistently increased the top-1 accuracy across all datasets tested. Given that the top-1 accuracy is the primary performance indicator in the ASLT system, these findings collectively confirm the effectiveness of the FvGEF strategy in boosting the proposed system's overall performance.

Furthermore, we conducted ablation studies to explore the effectiveness of the deformable attention module in FvGEF. To ensure comparability and maintain methodological consistency, we adhered strictly to the previously described configurations for all other modules, strategies, and parameter settings. Subsequently, we systematically investigated the effect of incorporating the deformable attention module, contrasting it against scenarios where this module was omitted, in terms of recognition accuracy across five distinct datasets. Experimental results are shown in Table 4. The data show that when the deformable attention module was integrated, the algorithm achieved a slight improvement in Top-1 recognition accuracy across the five datasets compared to the scenario without this module. This improvement may be attributed to the fact that deformable attention can enhance semantic coherence and mitigate occlusion influence in WSLR.

Table 4. Experiments on deformable attention in the FvGEF. Performance is measured in Top-1 accuracy (%).

	NationalCSL-200	***NationalCSL-500***	***NationalCSL-1000***	***NationalCSL-2000***	***NationalCSL-6707***
baseline	78.50	77.40	76.20	75.10	57.49
***wo*/ DA**	81.00	78.00	76.95	75.65	61.24
***w*/ DA**	81.50	78.20	77.10	75.80	61.37

When integrated with frame-level alignment in the full model, FvGEF further improved the performance, particularly in terms of higher-order metrics such as Top-5 and Top-10. This suggests that FvGEF refines feature representations to enhance discrimination among semantically similar signs, but its benefits are contingent on the availability of sufficient data to support its intricate fusion mechanism. Overall, the ablation study confirms that frame-level alignment is a robust performance enhancer across all dataset sizes, whereas FvGEF serves as a complementary component that delivers incremental gains, especially on large-scale datasets. These findings emphasize the criticality of designing architectural components with both model compatibility and data characteristics in mind.

6. Conclusions

In this paper, we present a novel design for an ASLT system accompanied by an efficient dual-view WSLR algorithm tailored for Chinese sign language. The proposed algorithm employs a Swin transformer module and a transformer encoder for spatio-temporal feature modeling, which enables robust representation learning from dual-view word-level sign videos. To enhance recognition performance, we introduce a frame-level alignment module to synchronize temporal dynamics across dual views and an FvGEF strategy for cross-view feature integration. These components mitigate temporal misalignment and enhance inter-view semantic coherence, thereby yielding more accurate and stable recognition outcomes. The results of comprehensive experiments conducted on the NationalCSL-DP benchmark validate the effectiveness and efficiency of our framework. The proposed approach showed significant improvements in recognition accuracy across multiple metrics (e.g., Top-1, Top-5 and Top-10) and dataset subsets (ranging from NationalCSL200 to NationalCSL6707), thus confirming its superiority. Overall, this work provides a practical and scalable solution for dual-view Chinese WSLR, with substantial potential for real-world applications in education, assistive technologies, and communication systems.

 doi:10.20944/preprints202601.1551.v1

Contributions: Siyuan Jing is the director of the work and he designed the proposed algorithm and carried out the experiments and participated in manuscript writing; Gaorong Yan participated in manuscript writing.

Funding Declaration: This work is supported by the Humanities and Social Sciences Project of the Ministry of Education (Grant No. 25YJA740011), the General Project of the Philosophy and Social Sciences Foundation of Sichuan Province (Grant No. SCJJ24ND127), the Research Cultivation Project of Leshan Normal University (Grant No. KYPY2024-0002), and the Sichuan Provincial Key Laboratory of Philosophy and Social Sciences for Language Intelligence in Special Education (Grant No. YYZN-2025-6).

Data availability statement: The NationalCSL-DP dataset can be downloaded from https://figshare.com/articles/media/NationalCSL-DP/27261843.

Conflicts of interest: The authors declare that they have no conflicts of interest.:.

References

1. Alyami S, Luqman H, Hammoudeh M. (2024). Isolated Arabic sign language recognition using a transformer-based model and landmark keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1): 1-19.
2. Alyami S, Luqman H. (2025). Swin-MSTP: Swin transformer with multi-scale temporal perception for continuous sign language recognition. *Neurocomputing*, 617, 129015.
3. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, pp 6836-6846.
4. Bruce X, Liu Y, Zhang X, Zhong S, Chan K. (2022). MMnet: A model-based multimodal network for human action recognition in RGB-D videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3522-3538.
5. Wang H, Chai X, Hong X, Zhao G, Chen X (2016) Isolated sign language recognition with Grassmann covariance matrices. *ACM Transactions on Accessible Computing*, 8(4): 1-21.
6. Das S, Ryoo M. (2023). ViewCLR: Learning self-supervised video representation for unseen viewpoints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'23)*, pp 5573-5583.
7. De Coster M, Van Herreweghe M, Dambre J. (2021). Isolated sign recognition from RGB video using pose flow and self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, pp 3441-3450.
8. Dinh N., Nguyen T, Tran D, Pham N, Tran T, Tong N, Le Nguyen P. (2025). Sign language recognition: A large-scale multi-view dataset and comprehensive evaluation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'25)*, pp 7887-7897.
9. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Houlsby N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*:2010.11929.
10. Du Y, Xie P, Wang M, Hu X, Zhao Z, Liu J. (2022). Full transformer network with masking future for word-level sign language recognition. *Neurocomputing*, 500: 115-123.
11. Fink J, Poitier P, André M, Meurice L, Frénay B, Cleve A, Meurant L. (2023). Sign language-to-text dictionary with lightweight transformer models. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'23)*, pp 5968-5976.
12. Guan Z, Hu Y, Jiang H, Sun Y, Yin B. (2025). Multi-view isolated sign language recognition based on cross-view and multi-level transformer. *Multimedia Systems*, 31(3), 1-15.
13. Hasan K, Adnan M. (2025). EMPATH: MediaPipe-aided ensemble learning with attention-based transformers for accurate recognition of Bangla word-level sign language. In *International Conference on Pattern Recognition (ICPR'25)*, pp 355-371.
14. Hosain A, Santhalingam P, Pathak P, Rangwala H, Kosecka J. (2021). Hand pose guided 3D pooling for word-level sign language recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'21)*, pp 3429-3439.

 doi:10.20944/preprints202601.1551.v1

15. Hu H, Zhao W, Zhou W, Li H. (2023). SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 11221-11239.
16. Hu H, Zhou W, Pu J, Li H. (2021). Global-local enhancement network for NMF-aware sign language recognition. *ACM transactions on Multimedia Computing, Communications, and Applications*, 17(3): 1-19.
17. Huang J, Zhou W, Li H, Li W. (2018). Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9): 2822-2832.
18. Ji Y, Yang Y, Shen H, Harada T. (2021). View-invariant action recognition via unsupervised attention transfer (UANT). *Pattern Recognition*, 113: 107807.
19. Jiang S, Sun B, Wang L, Bai Y, Li K, Fu Y. (2021). Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, pp 3413-3423.
20. Jin P, Li H, Yang J, Ren Y, Li Y, Zhou L, Liu J, Zhang M, Pu X, Jing S. (2025). A large dataset covering the Chinese national sign language for dual-view isolated sign language recognition. *Scientific Data*, 12:1-10.
21. Jing S, Wang G, Zhai H, Tao Q, Yang J, Wang B, Jin P. (2025). Dual-view spatio-temporal feature fusion with CNN-Transformer hybrid network for Chinese isolated sign language recognition. *arXiv:*2506.06966
22. Joze H, Koller O. (2019). MS-ASL: MS-ASL: A large-scale data set and benchmark for understanding American sign language. In *Proceedings of the 30th British Machine Vision Conference (BMVC'19)*, 100.
23. Kang X, Yao D, Jiang M, Huang Y, Li F. (2022). Semantic network model for sign language comprehension. *International Journal of Cognitive Informatics and Natural Intelligence,* 16(1): 1-19.
24. Koller O, Camgoz N, Ney H, Bowden R. (2019). Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9): 2306-2320.
25. Kusnadi A. (2015). Motion detection using frame differences algorithm with the implementation of density. In *Proceedings of the International Conference on New Media (ICNM'15)*, pp 57-61.
26. Kwak I, Guo J, Hantman A, Kriegman D, Branson K. (2020). Detecting the starting frame of actions in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'20)*, pp. 489-497.
27. Li B, Yuan C, Xiong W, Hu W, Peng H, Ding X, Maybank, S. (2017). Multi-view multi-instance learning based on joint sparse representation and multi-view dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2554-2560.
28. Li D, Rodriguez C, Yu X, Li H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'20)*, pp 1459-1469.
29. Li J, Wong Y, Zhao Q, Kankanhalli M. (2018). Unsupervised learning of view-invariant action representations. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'18)*, pp 1262-1272.
30. Li Y, Wu C, Fan H, Mangalam K, Malik J, Feichtenhofer C. (2022). Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR'22)*, pp 4804-4814.
31. Liu N, Li X, Wu B, Yu, Q, Wan L, Fang T, Zhang J, Li Q, Yuan Y. (2025). A lightweight network-based sign language robot with facial mirroring and speech system. *Expert Systems with Applications*, 262, 125492.
32. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR'21)*, pp 10012-10022.
33. Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, Zhang F, Chang C L, Yong M G, Lee J, Chang W T, Hua W, Georg M, Grundmann M. (2019). MediaPipe: A framework for building perception pipelines. *arXiv*: 1906.08172.
34. Ma Y, Yuan L, Abdelraouf A, Han K, Gupta R, Li Z, Wang Z. (2023). M2DAR: Multi-view multi-scale driver action recognition with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pp 5287-5294.
35. Cao Z, Hidalgo G, Simon T, Wei S E, Sheikh Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 43(1): 172-186

 doi:10.20944/preprints202601.1551.v1

36. Nguyen H, Nguyen T. (2021). Attention-based network for effective action recognition from multi-view video. *Procedia Computer Science*, 192: 971-980.
37. Rajalakshmi, E, Elakkiya R, Prikhodko A, Grif M, Bakaev M, Saini J, Subramaniyaswamy V. (2022). Static and dynamic isolated Indian and Russian sign language recognition with spatial and temporal feature detection using hybrid neural network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1): 1-23.
38. Ren T, Yao D, Yang C, Kang X. (2024) . The influence of Chinese characters on Chinese sign language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1): 6:1-6:31.
39. Rousseeuw P, Croux C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424): 1273-1283.
40. 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.
41. Shen X, Du H, Sheng H, Wang S, Chen H, Chen H, Yu X. (2024). MM-WLAuslan: multi-view multi-modal word-level Australian sign language recognition dataset. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'24)*.
42. Shi L, Zhang Y, Cheng J, Lu H. (2021). AdaSGN: Adapting joint number and model size for efficient skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, pp 13413-13422.
43. Sincan O, Keles H. (2020). AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods. *IEEE Access*, 8: 181340-181355.
44. Singla, N. (2014). Motion detection based on frame difference method. *International Journal of Information and Computation Technology*, 4(15): 1559-1565.
45. Wang H, Chai X, Hong X, Zhao G, Chen X. (2016). Isolated sign language recognition with Grassmann covariance matrices. *ACM Transactions on Accessible Computing*, 8(4): 1-21.
46. Wang L, Ding Z, Tao Z, Liu Y, Fu Y. (2019). Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV'19)*, pp 6212-6221.
47. Wu Z, Ma N, Wang C, Xu C, Xu G, Li M. (2024). Spatial–temporal hypergraph based on dual-stage attention network for multi-view data lightweight action recognition. *Pattern Recognition*, 151: 110427.
48. Xia Z, Pan X, Song S, Li L, Huang G. (2022). Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR'22)*, pp 4794-4803.
49. Xu Y, Jiang S, Cui Z, Su F. (2024). Multi-view action recognition for distracted driver behavior localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pp 7172-7179.
50. Yamane T, Suzuki S, Masumura R, Tora S. (2024). MVAFormer: RGB-based multi-view spatio-temporal action recognition with transformer. *In 2024 IEEE International Conference on Image Processing (ICIP'24)*, pp 332-338.
51. Yang Y, Liang G, Wu X, Liu B, Wang C, Liang J, Sun J. (2024). Cross-view fused network for multi-view rgb-based action recognition. In *2024 IEEE 8th International Conference on Vision, Image and Signal Processing (ICVISP'24)*, pp. 1-7.
52. Zhang D, Dai X, Wang X, Wang Y F. (2018). S3D: Single shot multi-span detector via fully 3D convolutional networks. In *Proceedings of the 2018 British Machine Vision Conference (BMVC'18)*, 293.
53. Zhang R, Hu C, Yu P, Chen Y. (2025). Improving multilingual sign language translation with automatically clustered language family information. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING'25)*, pp 3579–3588.
54. Zhang R, Zhao R, Wu Z, Zhang L, Zhang H, Chen Y. (2025). Dynamic feature fusion for sign language translation using hypernetworks. In *Findings of the Association for Computational Linguistics (NAACL'25)*, pp 6227–6239.
55. Zhou H, Zhou W, Zhou Y, Li H. (2021). Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24: 768-779.
56. Zuo R, Wei F, Mak B. (2023). Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pp 14890-14900.

 doi:10.20944/preprints202601.1551.v1