VSNet: Focusing on the Linguistic Characteristics of Sign Language

Yuhao Li[1], Xinyue Chen[1], Hongkai Li[1], Xiaorong Pu[1,3], Peng Jin[2], Yazhou Ren[1,3,*]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China; [2]Sichuan Provinical Key Lab of Philosophy and Social Science for Language Intelligence in Special Education, Leshan Normal University, Leshan, China; [3]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China

Abstract

Sign language is a visual language expressed through complex movements of the upper body. The human skeleton plays a critical role in sign language recognition due to its good separation from the video background. However, mainstream skeleton-based sign language recognition models often overly focus on the natural connections between joints, treating sign language as ordinary human movements, which neglects its linguistic characteristics. We believe that just as letters form words, each sign language gloss can also be decomposed into smaller visual symbols. To fully harness the potential of skeleton data, this paper proposes a novel joint fusion strategy and a visual symbol attention model. Specifically, we first input the complete set of skeletal joints, and after dynamically exchanging joint information, we discard the parts with the weakest connections to other joints, resulting in a fused, simplified skeleton. Then, we group the joints most likely to express the same visual symbol and discuss the joint movements within each group separately. To validate the superiority of our method, we conduct extensive experiments on multiple public benchmark datasets. The results show that, without complex pretraining, we still achieve new state-of-the-art performance. The code is available at https://github.com/atinyboy/VSNet.

1. Introduction

Sign language is a visual language expressed through body movements and serves as the primary communication tool for the deaf community [13]. Although sign language and spoken language have distinctly different grammatical structures [1, 7], they share common linguistic features [39, 44, 59]. Accurate sign language translation can facilitate communication between deaf and hearing people, fostering a more inclusive and accessible society. Among the key subtasks in sign language translation is isolated sign language recognition (ISLR), which focuses on translating individual glosses. ISLR has garnered significant attention as it aids people in accurately understanding numerous gestures and quickly learning sign language.

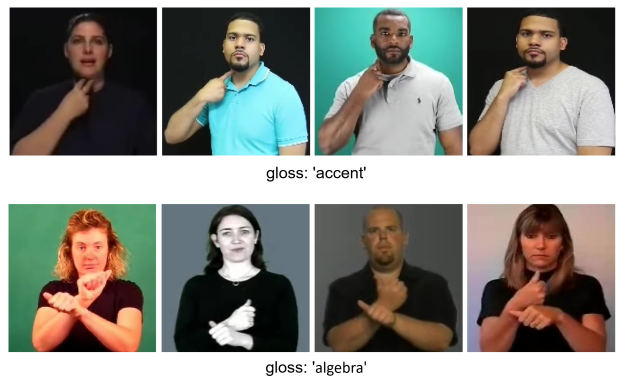

Figure 1. In the sign for "accent", the index and middle fingers (or only index finger) are extended toward the neck while other fingers are curled, forming the "pointing" VS. In the sign for "algebra", movements can be decomposed into multiple VS such as "crossed arms," "fist," and "thumb-up."

In recent years, human skeletons have been increasingly applied in the ISLR field due to their strong operability and clear separation from video backgrounds. Although most skeleton-based recognition methods have achieved promising results, they typically treat sign language as typical human motion [12, 21, 23, 30, 31, 41, 58], overlooking the intrinsic relationships among joints and the unique linguistic characteristics [25, 52] of sign language itself. We believe that: (1) The importance of joints in sign language is not equivalent; joints with stronger connections to others are more significant than isolated joints, and connections among joints with similar movement tendencies hold more meaning than those among dispersed joints. (2) Similar to how letters form words, each sign language gloss can be decomposed into smaller visual symbols (VS). While VS

*Corresponding author (yazhou.ren@uestc.edu.cn).

lack explicit meaning on their own, from a skeletal perspective, they capture distinct movement patterns of joints, effectively decomposing sign language actions (see Fig. 1). Therefore, our work focuses on these two core aspects:

Core 1: Focus on identifying essential joints. Including too many joints can impede training, while too few can lead to information loss. On the one hand, many prior models have used manually trimmed, simplified skeletons, but manually selecting the optimal joints is not trivial. On the other hand, enabling a model to automatically choose key joints can be computationally intensive. Fortunately, the dynamically learned skeletons from GCNs approximate real structures more closely, enabling us to reduce the dependency on an initial skeleton. By computing the weight of each joint, we can identify and retain the most important joints while discarding less critical local joints. Specifically, for the complete hand skeleton, we discard a small number of joints based on computed weights after each GCN layer, relearning new skeleton connections in the subsequent GCN layer. We control the number of joints discarded within each region to ensure that the final GCN output retains a recognizable structure. Notably, even joints with the lowest weights undergo at least one convolution, so their information is not entirely discarded.

Core 2: Focus on emphasizing visual symbols (VS) in sign language. If each joint is considered the smallest element of a sign language action, then the combination of joints across all frames becomes too vast to compute self-attention. In practice, sign language is not random motion; it is designed for human recognition. When a signer interprets sign language, they do not scrutinize the position of every joint but instead form a general impression of each hand. Thus, in this work, we concentrate on recognizing VS. Unfortunately, isolating each VS from movements is as challenging as the entire SLR task itself. Therefore, we adopt a simplified approach, segmenting the skeleton into a generalized VS in a fixed manner, as shown in Fig. 2. We then design two self-attention modules to capture spatial and temporal relationships between these VS. Empirically, we find that this approach significantly enhances the model's discriminative power.

Our main contributions are summarized as follows:

- We propose a novel module with an automated skeleton simplification function for sign language recognition. This module is designed to dynamically learn global skeleton connections while progressively discarding isolated local joints.
- Considering the linguistic characteristics of sign language, we introduce a new self-attention model called VSformer, which captures temporal and spatial patterns in sign language motions and extracts knowledge from the human skeleton based on the concept of VS.
- We conduct extensive experiments on ISLR to validate the effectiveness of our proposed approach. Compared to previous methods, our approach achieves SoTA performance on four public benchmarks, namely **WLASL** [32], **MSASL** [26], **NMFs-CSL** [16], and **Slovo** [27].

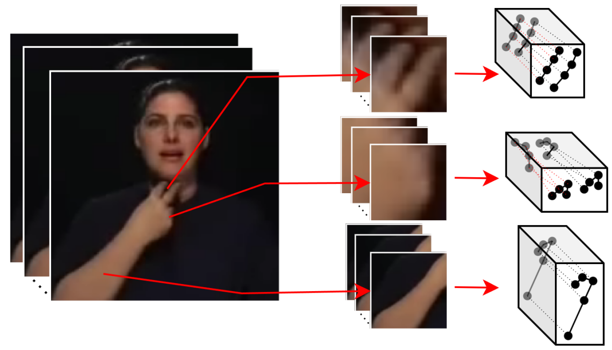

Figure 2. Decomposing generalized visual symbols from videos

2. Related Works

2.1. Sign Language Recognition

Sign language is the primary means of communication within the deaf community. With significant advances in machine learning and computer vision over the past decade, exploring sign language recognition (SLR) has become increasingly important [28, 43, 45], as it can automatically interpret sign language and help deaf and mute individuals communicate more easily in daily life. Unlike directly translating sign language sentences, isolated sign language recognition (ISLR) is a subtask in the field of sign language translation [45], aimed at translating each individual gloss [40, 53]. ISLR is not only useful for guiding beginners in learning sign language and assisting in the creation of new sign language gestures, but it also lays a solid foundation for SLR [1, 4, 53]. Research in this area can be divided into two categories based on the input modality: RGB-based methods and skeleton-based methods.

RGB-based Recognition. RGB-based methods, similar to other video recognition tasks, typically use CNNs to extract visual features from RGB videos [5, 15, 19, 20, 29, 55–57], and then aggregate temporal and spatial information to predict sign language. However, RGB-based models face two significant challenges: first, they are heavily affected by changes in the video background, and second, video data occupies a large amount of memory resources.

Skeleton-based Recognition. Compared to RGB images, human skeletons [8, 47, 50, 54] describe the motion information conveyed in sign language videos using simplified points and edges, reducing interference from complex backgrounds and the appearance of the signer. Moreover, operations on three-dimensional spatial points with actual physical meanings offer stronger interpretability and

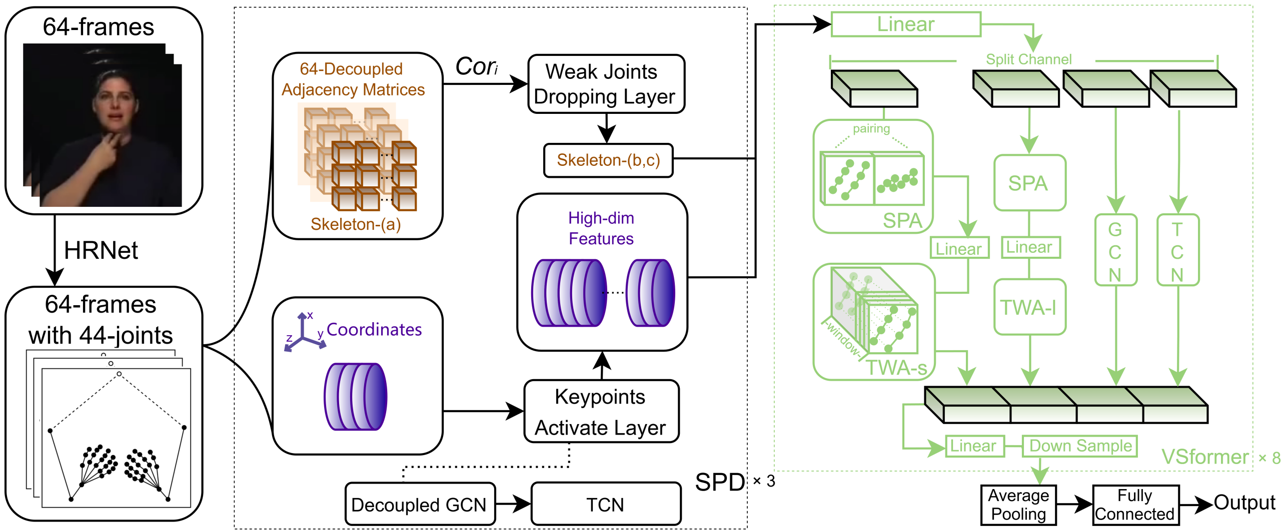

Figure 3. Overview of our VSNet. Here, *Down sample* consists of a convolution layer with a kernel size of 7 and a stride of 2 along the temporal dimension, followed by a batch norm layer, *Average Pooling* represent transforming the data into a single-dimensional feature by taking the mean across both temporal and spatial dimensions. For details on the *Skeleton-(a,b,c)*, please refer to Fig. 4.

lower memory consumption. In earlier works, identifying the correct human keypoints in a video was a major challenge. However, with the development of deep learning, human pose estimation models have become increasingly reliable [10, 24, 51].

2.2. Human Skeleton

Due to the natural physical connections between skeleton joints, Graph Convolutional Network (GCN) has been widely applied to model spatio-temporal relationships [12, 23, 31, 34, 41, 46]. GCN require predefined connectivity of the key points to construct the graph network. However, the physically natural skeleton connections may not always be the most suitable edges for recognition. Relying solely on the skeleton itself to construct the graph can limit the information flow between nodes. To address this issue, SL-GCN [21] employs a dynamic graph with decoupling between different frames, while TD-GCN [37] adds supernodes to control the connections between nodes. A common drawback of these approaches is that they require complex networks to model the key points. Furthermore, while dynamic graphs alleviate the limitations of the original skeleton, blindly allowing all key points to connect introduces redundant information, leading to suboptimal models.

Therefore, we argue that relying solely on GCN does not fully exploit the true potential of skeleton data.

In recent years, attention-based methods [3, 6, 9, 38, 42, 48] have also made progress in human skeleton recognition. Unlike GCN, attention mechanisms do not require the researcher to provide information on joint connections, instead equally attending to all nodes. However, capturing the correlations between all joints across all frames requires substantial memory resources. To address this, SLGTformer [49] decomposes the spatio-temporal pose sequence into spatial graphs and temporal windows, while SkateFormer [11] partitions joints and frames and performs self-attention on each partition. Based on these theories, we design a model where GCNs and self-attention networks complement each other. GCNs are responsible for exploring the spatial correlations between human joints, while self-attention networks focus on capturing the motion of the skeleton across multiple time sequences and spatial combinations.

3. Methods

3.1. Data Pre-Processing

Key Point Extraction. Sign language is a visual language primarily conveyed through hand and body movements, with facial expressions also playing a crucial role. However, the human skeleton mainly describes bones and joints, excluding muscles, making it unable to accurately capture facial expressions in sign language videos. Our investigation revealed that a significant number of signers, especially beginners, do not exhibit extensive facial expressions during sign language communication. Therefore, our main experiments exclude facial expressions, while the appendix provides additional details on how the concept of VS can be extended to the face.

In previous GCN models, finger joints were often simplified to avoid making the graph too complex, leading to a loss of information. In this study, we selected a complete

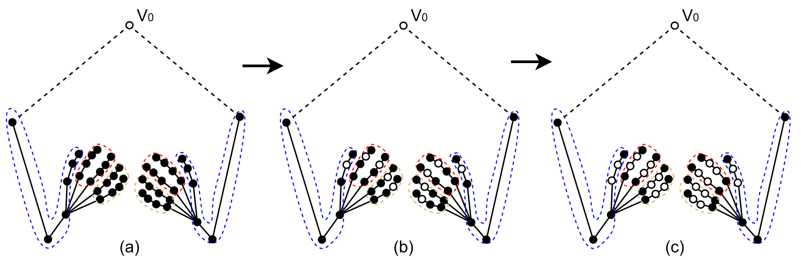

Figure 4. Illustration of the skeleton and grouping type-1. V_0 represents a localization node, the dashed box indicates nodes that are grouped together and hollow dots indicate dropped nodes. Please note that (a) is the input real skeleton diagram, while (b) and (c) are simplified representations drawn after dropping for ease of analysis.

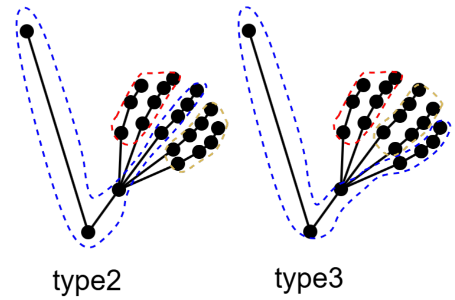

Figure 5. Skeleton groupings type-2 and type-3. The other hand is grouped in a mirrored fashion.

set of 44 key points from both hands out of the 133 human body joints extracted using HRNet [24, 51], as shown in Fig. 4(a).

Data Enhancing. Given a video $X \in \mathbb{R}^{T\times V\times C}$, where T, V, and C represent the number of frames, joints, and coordinates respectively. We replace abnormal points—such as those out of range (less than 0, more than 512) or clearly distorted—caused by model extraction errors, with the average value from adjacent frames. Additionally, to standardize body positions and movement amplitudes across different signers in the sign language videos, we first use the hollow node V_0 in Fig. 4 as the reference node for the entire video. Then, all key points in each frame are shifted based on this reference coordinate, as follows:

$$X(t, v, c) = X(t, v, c) - X(t, v_0, c) + [x, y, z] \quad (1)$$

where $[x, y, z]$ is the average value of all the coordinates of V_0 in the dataset.

In the time dimension, we randomly select 64 frames to standardize the video length and perform random mirroring and shifting of the video to generate augmented data.

3.2. Self-Pacing Dropping Block (SPD)

As shown in Fig. 3, our SPD consists of two components: “Keypoint Activation” and “Weak Joints Dropping”. The input is a complete hand skeleton where the node information is independent, and the output is a high-dimensional feature skeleton after activation. The entire block is stacked three times, where the third block only outputs features and does not build a new skeleton.

Keypoint Activation Layer. In this part, we define the interaction of each joint with others as an activation. We employ a decoupled graph convolution [21] to perform this activation, represented as:

$$F^t_{out} = \frac{1}{p}\sum_{k=0}^{p} D^{-\frac{1}{2}} A_k D^{-\frac{1}{2}} \phi(F^t_{in}) \quad (2)$$

where A is the learnable adjacency matrix initialized as $I + In + Out$, D presents the diagonal degree of A. Here, I denotes self-connections, while In and Out represent inward and outward connections, respectively. The symbol ϕ represents a linear layer, and p represents the decoupling quantity, with a different A applied each time, followed by averaging the results. Following the graph convolution, we apply the simplest Temporal Convolutional Network (TCN) to capture features along the temporal dimension.

Weak Joints Dropping Layer. Compared to the simplified skeletons used in [21, 49], our approach utilizes a complete set of skeleton keypoints that contains richer information. However, this comes at the cost of significantly increasing the model’s burden and raising the risk of introducing outlier points. To leverage these additional points with minimal overhead, we have designed an adaptive skeleton simplification strategy.

Before formally introducing this layer, it is necessary to present our approach to skeleton decomposition. Based on the definition of VS described in *Introduction*, the decomposition concept involves grouping joints that exhibit strong overall correlation or frequently form VS together in sign language. These groupings can be viewed as a generalized or macro-level VS. A simple approach is grouping each finger and arm separately, but this creates too many groups, potentially affecting later interactions. Based on this, we conducted various experiments on VS, including treating individual fingers and distant fingers as VS. However, these experiments did not yield optimal results (the average performance in WLASL dataset decreased by 1.20% and 1.64%, respectively). We believe this is because the information from a single finger is insufficient to form a meaningful VS, while the information disparity between distant fingers is even greater. This also serves as the skeleton grouping basis for Figs. 4 and 5.

After graph convolution in Eq. 2, $A(i, j)$ has learned the weighted edge between node i and node j. These edges reflect the connection strength between i and j. Since these connections are not normalized via softmax, we can use A as a guide to calculate the total correlation of all keypoints, Cor, as follows:

$$Cor_i = \sum_j \max(A_0(i,j), A_1, \ldots, A_p) \tag{3}$$

A lower Cor to other keypoints indicates that the point is more independent. We refer to the two most independent points in each group as weak joints, which are dropped from the skeleton. To ensure that the number of keypoints in each group remains consistent, the groups involving both arms will have two fewer joints dropped.

3.3. VSformer

After the joint fusion block, we obtained six different joint groupings, as shown in Fig. 4(c). Based on the concept of skeletal disassembly, we treat each grouping as a generalized VS. We now need to focus on two main questions: Q1, what are the relationships between different VS? And Q2, do these VS remain consistent throughout the entire video, or do they change after a brief period?

Inspired by [11], which proposes an efficient model for partitioning human limbs in action recognition, we employ two types of self-attention to constrain the model's focus on the aforementioned questions. Specifically, we calculate spatial pairing attention among different combinations within a frame to address Q1, and we compute temporal window attention for both global and local motion of the combinations across different frames to address Q2.

Spatial Pairing Attention (SPA). Given a frame $F_t \in \mathbb{R}^{G \times C}$, where $G = 6$ represents VS groupings, the ideal approach would involve calculating attention among all VS pairwise. However, this is inefficient and often purposeless, as VS that are far apart in the skeletal structure tend to have weak correlations. For instance, in the sign language gesture for "accent", it is unnecessary to compare the fingers of the left hand with those of the unused right hand. Based on this observation, we ultimately partition F_t into pairs as follows:

$$F_t = [P_{\text{arm}}, P_{\text{left}}, P_{\text{right}}] \tag{4}$$

where $P_{\text{arm}} \in \mathbb{R}^{2 \times C}$ represents the grouping of both arms, P_{left} denotes the four fingers of the left hand, and P_{right} represents the remaining four fingers of the right hand. Represented as:

$$SPA_{\text{out}} = \text{concat}\left(\text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{h}}\right) V_i\right) \tag{5}$$

where Q_i, K_i, V_i are obtained from P_i after mapping through linear layers. The attention is actually calculated among the 8 joints within each pairing and is ultimately concatenated to form the output.

Temporal Window Attention (TWA). Depending on different sign language videos, the VS can be classified into

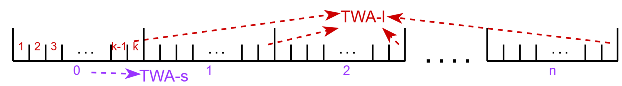

Figure 6. TWA windows. The small ticks represent frames, and the large ticks represent windows.

two types: short duration and long duration, representing their respective duration periods. Accordingly, our TWA is also divided into two types: TWA-s and TWA-l. Given a video $X \in \mathbb{R}^{T \times G \times C}$, where T represents the total number of frames, the timeline is divided into n windows, each containing k frames, such that $T = n \times k$, as shown in Fig. 6. For TWA-s, attention is computed within each window. For TWA-l, frames with the same index are selected sequentially from each window. Finally, the results from both types of attention are concatenated to produce the output, represented as:

$$TWA_{out} = \text{softmax}(\phi_1(W)\phi_2^T(W) + b)\phi_3(W) \tag{6}$$

where W represents the selected windows, ϕ_n denotes dividing the C dimension into three equal parts and selecting the n-th part, and b represents the added bias.

It is worth noting that the discussions on Q1 and Q2 are not fully addressed by these two attention layers. The SPA only addresses the relationships between paired VS, while the windows defined in TWA do not entirely correspond to the duration of the VS. Therefore, we incorporate a simple GCN and TCN to enhance the model's ability to address the remaining scenarios. The connection method of SPA and TWA is shown in the right side of Fig. 3. After passing the input data $X \in \mathbb{R}^{T \times G \times C}$ through a linear layer, we divide its channel dimension C into four equal parts, sending each part into one of the four modules and later concatenated. Additionally, unlike the Decoupled GCN in Sect. 3.2, the GCN here utilizes an adjacency matrix shared across certain frames, which significantly reduces the computational load.

3.4. Multi-Grouping Ensemble

In the previous sections, we introduced the concept of skeletal disassembly based on VS. However, since we cannot precisely determine the true VS composition for each sign language, using a fixed disassembly method appears to be a compromise, potentially weakening interpretability. Therefore, we adopt an ensemble strategy to reduce such discrepancies. Specifically, we designed two additional skeleton grouping schemes, as shown in the Fig. 5, and trained the model separately to obtain the output q from the final fully connected layer for each scheme. Following [21], we aggregate these results using a weighted sum, represented as: $q_{\text{total}} = \alpha_1 q_{\text{type-1}} + \alpha_2 q_{\text{type-2}} + \alpha_3 q_{\text{type-3}} + \alpha_4 q_{\text{bone}}$, where α represents the weights. For a detailed discussion on α,

Methods	MSASL1000				MSASL200				MSASL100			
	P-I		P-C		P-I		P-C		P-I		P-C	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
RGB-based												
HMA [15]	49.16	69.75	46.27	68.60	66.30	84.03	67.47	84.03	73.45	89.70	74.59	89.70
I3D [5]*	-	-	57.69	81.05	-	-	81.97	93.79	-	-	81.76	95.16
BSL [2]	64.71	85.59	61.55	84.43	-	-	-	-	-	-	-	-
TCK [33]†	-	-	-	-	80.31	91.82	81.14	92.24	83.04	93.46	83.91	93.52
skeleton-based												
ST-GCN [58]	36.03	59.92	32.32	57.15	52.91	76.67	54.20	77.62	59.84	82.03	60.79	82.96
BEST [60]	58.82	81.18	54.87	80.05	76.60	91.54	76.75	91.95	80.98	95.11	81.24	95.44
SignBERT [14]†	59.80	81.86	57.06	80.94	77.34	91.10	78.02	91.48	81.37	93.66	82.31	93.76
SignBERT+ [17]†	62.42	83.49	60.15	82.44	78.51	-	79.35	-	84.94	-	85.23	-
MASA [61]†	63.47	83.89	60.79	83.29	79.25	92.86	79.70	93.33	83.22	95.24	83.19	95.46
SSRL [62]†	65.22	85.09	62.68	84.38	82.93	94.78	83.43	95.03	86.26	96.96	86.63	96.79
Ours	70.71	86.36	68.38	85.71	84.62	94.04	85.19	94.08	86.26	95.24	86.68	95.39
Ours(4-crops)	**73.18**	**88.64**	**70.88**	**87.59**	**86.17**	**95.51**	**86.64**	**95.61**	**88.11**	**97.09**	**88.32**	**97.18**

Table 1. Comparison with previous work on **MSASL**. "†" indicates that the model was pre-trained, and "*" indicates methods that applied result fusion with other modalities in post-processing. Bold indicates the best performance, while underlined denotes the second-best.

Methods	WLASL2000				WLASL300				WLASL100			
	P-I		P-C		P-I		P-C		P-I		P-C	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
RGB-based												
I3D [5]*	32.48	57.31	-	-	56.14	79.94	56.24	78.38	65.89	84.11	67.01	84.58
HMA [15]	37.91	71.26	35.90	70.00	-	-	-	-	-	-	-	-
BSL [2]	46.82	79.36	44.72	78.47	-	-	-	-	-	-	-	-
TCK [33]†	-	-	-	-	68.56	89.52	68.75	89.41	77.52	91.08	77.55	91.42
NLA-SLR V64 [63]*	51.15	83.43	48.14	82.20	-	-	-	-	-	-	-	-
skeleton-based												
P3D [30]	44.47	79.69	42.18	78.52	67.18	89.01	67.62	89.24	76.71	91.97	78.27	92.97
BEST [60]	46.25	79.33	43.52	77.65	67.66	89.22	68.31	89.57	77.91	91.47	77.83	92.50
SignBERT [14]†	47.46	83.32	45.17	82.32	70.36	88.92	71.17	89.36	79.07	93.80	80.05	94.17
MASA [61]†	49.06	82.90	46.91	81.80	73.65	91.77	74.33	92.13	83.72	93.80	84.47	94.30
NLA-SLR K64 [63]*	49.10	82.00	46.18	80.71	-	-	-	-	-	-	-	-
SL-GCN [22]*	51.50	84.94	48.87	84.02	-	-	-	-	-	-	-	-
DSTA [18]	53.68	-	51.17	-	79.97	-	80.56	-	82.38	-	83.09	-
Ours	53.54	83.18	51.18	82.03	79.04	94.01	79.74	94.25	84.50	94.57	85.25	94.33
Ours(4-crops)	**55.98**	**87.07**	**53.54**	**86.04**	**80.09**	**94.76**	**80.85**	**94.92**	**85.66**	**94.96**	**86.25**	**95.08**

Table 2. Comparison with previous work on **WLASL**. "†" indicates that the model was pre-trained, and "*" indicates methods that applied result fusion with other modalities in post-processing. Bold indicates the best performance, while underlined denotes the second-best.

please refer to Sect. 4.4. The term *bone* refers to the joint data of type-1 represented in vector form, generated by directing from the source joint to its target joint according to the natural connections of the human body.

4. Experiments

4.1. Parameter Settings

We employ the AdamW optimizer with a plateau learning rate schedule. The initial learning rate is set to 1×10^{-3}, with a minimum learning rate of 1×10^{-5} and a decay rate of 0.5. The batch size is 32 and the loss function used is cross-entropy with label smoothing. Training spans a maximum of 200 epochs, with the first 15 epochs serving as a warm-up period starting from a learning rate of 1×10^{-7}. A random sample of 64 frames is selected as input, and with a 50% probability, image mirroring is applied. A random jitter in the range $(-10, 10)$ is added to each joint point. In the last two layers of SPD, joints are randomly masked with a probability of $(1 - p_{kr})/100 \times epoch$, where p_{kr} is set to 0.9. All experiments were performed with NVIDIA A100.

Methods	NMFs-CSL	
	Top-1	Top-5
RGB-based		
I3D [5]*	64.4	88.0
HMA [15]	64.7	91.0
GLE-Net [36]	69.0	88.1
skeleton-based		
ST-GCN [58]	59.9	86.8
BEST [60]	68.5	94.4
MASA [61]†	71.7	<u>97.0</u>
SignBERT [14]†	74.9	93.2
Ours	<u>75.3</u>	95.4
Ours(4-crops)	**76.7**	**97.2**

Table 3. Comparison on **NMFs-CSL**. Bold indicates the best performance, while underlined denotes the second-best.

Methods	Slovo	
	Top-1	Top-5
Swin-large-48 [38]	55.66	-
MViTv2-small-48 [35]	62.18	-
MViTv2-small-32 [35]	64.09	-
Ours	<u>76.53</u>	<u>93.08</u>
Ours(4-crops)	**77.83**	**95.17**

Table 4. Comparison with the baselines on **Slovo**. The final number indicates the number of frames input to the model.

Metrics. Following [22, 61, 63], we report top-1 and top-5 classification accuracy, as well as per-instance (P-I) and per-class (P-C) precision metrics. Each table includes results for single-crop and 4-crop inference, with all ablation studies conducted in the single-crop setting. For 4-crops inference, we sum the predictions from the three types of skeleton groupings along with the bone data to obtain the final prediction.

4.2. Dataset

We conduct experiments on four publicly available benchmarks: WLASL, MSASL, NMFs-CSL, and Slovo. All training occurs on the training set, with hyperparameter tuning performed on the validation set.

MSASL [26] is a widely used dataset for American Sign Language (ASL). It contains a total of 16,054, 5,287, and 4,172 samples in the training, validation, and test sets, respectively, with a vocabulary of 1,000 signs. Two subsets, consisting of the top 100 and 300 words, are defined as MSASL100 and MSASL300, respectively.

WLASL [32] has a larger vocabulary but fewer total samples than MSASL. It contains 2,000 signs performed by over 100 signers, with 14,289, 3,916, and 2,878 samples in the training, validation, and test sets. Similarly, two subsets, WLASL100 and WLASL300, are created. Both ASL datasets are collected from real-world scenarios, offering greater authenticity.

NMFs-CSL [16] is a challenging Chinese Sign Language (CSL) dataset with a vocabulary of 1,067 and consists of 25,608 and 6,402 samples in the training and test sets, respectively.

Slovo [27] is the latest dataset for Russian Sign Language (RSL). It contains 20,000 videos across 1,000 isolated RSL gestures from 194 signers. Notably, the authors provide 42 hand landmarks based on MediaPipe, so we do not re-extract joint points.

4.3. Comparison with State-of-the-art Methods

In this section, we compare our model with several state-of-the-art approaches. For single-modality methods, we categorize them as either RGB-based or skeleton-based. For multi-modality methods, we report the performance of their single-modality models before ensemble.

MSASL [26]. As shown in the Tab. 1, RGB-based methods have developed slowly in recent years, and our approach shows a significant improvement when compared to them. Among skeleton-based methods, SignBERT [14] and BEST [60] designed self-supervised learning strategies to pre-train the model in order to alleviate issues related to limited interpretability and overfitting due to the small size of the SLR datasets. MASA [61] and SSRL [62] address the problem of neglecting explicit motion information and lexical signs during pre-training. However, none of these methods take into account the linguistic characteristics of sign language itself. Compared to them, our method achieves a substantial top-1 improvement of 7.96% in P-I and 8.2% in P-C on the full dataset. However, SSRL performs slightly better than our single-crop method on the 100-class subset, which we speculate is due to SPD requiring a certain amount of data to effectively learn the droppable nodes.

WLASL [32]. SAM-SLR [22] and NLA-SLR [63] are currently the two best-performing heavy multi-modality ensemble models. Among them, SL-GCN is the skeleton model used in SAM-SLR for later ensemble, employing a multi-stream approach to further improve performance. K64 and V64 are skeleton and RGB models in NLA-SLR that use 64-frame input, where the skeleton is represented as a heatmap sequence, allowing their keypoint encoder to share the same architecture with the video encoder. DSTA [18] was the previously most accurate model. Compared to their results, our top-1 accuracy achieved a 2.30% improvement on WLASL2000.

NMFs-CSL [16]. As shown in Tab. 3, NMFs-CSL is a balanced dataset in terms of class distribution, so we only report the P-I accuracy. The pretraining strategy of SignBERT [14] is highly effective on this dataset; however, its performance on top-5 is suboptimal. Our model does not use any additional training data, yet it shows a significant

Skeleton	MSASL1000		MSASL200		MSASL100		WLASL2000		WLASL300		WLASL100		NMFs-CLR	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
type-1	**70.71**	**86.36**	84.62	94.04	86.26	95.24	**53.54**	83.18	**79.04**	**94.01**	84.50	**94.96**	**75.26**	95.47
bone	68.41	85.88	82.41	93.38	84.68	94.06	50.76	82.45	74.70	91.17	84.11	93.80	73.63	94.31
type-2	70.35	86.10	**85.28**	**94.11**	**87.45**	**95.64**	53.02	**83.32**	78.44	93.41	83.72	93.80	74.70	94.86
type-3	70.35	86.10	84.92	**94.11**	85.60	95.11	52.26	82.49	78.14	93.41	**84.88**	93.41	74.87	**95.55**

Table 5. Performance of multi-grouping VSformer on three datasets.

weighting	WLASL2000		MSASL1000		NMFs-CSL	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
sum	55.98	87.07	73.18	88.64	76.65	97.17
acc	56.01	87.18	73.13	88.64	76.65	97.17

Table 6. Performance of the two ensemble methods

Methods	WLASL300		WLASL2000	
	Top-1	Top-5	Top-1	Top-5
Linear	74.70	92.81	46.32	77.24
GCN	76.35	91.77	51.18	80.82
SPD(ours)	**79.04**	**94.01**	**53.54**	**83.18**

Table 7. Comparing our SPD with GCN and linear layers.

SPA	TMA	WLASL300		WLASL2000	
		Top-1	Top-5	Top-1	Top-5
✓		77.69	93.41	52.43	82.24
	✓	77.99	93.26	52.36	81.83
✓	✓	**79.04**	**94.01**	**53.54**	**83.18**

Table 8. Ablation study on the impact of our two types of attention mechanisms.

improvement over SignBERT in overall performance, with a 1.8% improvement in top-1 accuracy and a 4% improvement in top-5 accuracy.

Slovo [27]. Slovo is a recently released RSL dataset, with only limited experimental data available. We provide an up-to-date benchmark for better evaluating this dataset. As shown in Tab. 4, the results for *Swin* [38] and *MViTv2* [35] were provided as baselines by the dataset authors. Notably, the skeleton landmarks provided by the authors includes only 42 hand joints, without arm data, which partially limits our model's performance. Despite this, we still achieve results significantly above the baselines.

4.4. Ablation Study

In this section, we conduct several key ablation experiments on the WLASL dataset [32] and report P-I accuracy to validate the effectiveness of our approach.

SPD. We first replace SPD with two alternative methods. As shown in Tab. 7, *"Linear"* denotes a three-layer *Conv-2D* module with the same output dimension as SPD, while *"GCN"* refers to three decoupled graph convolution layer without the weak joints dropping strategy. To ensure consistent output data shape, these two methods directly input the simplified 24 joint points. The complete SFB implementation achieves the highest accuracy, indicating that our dropping strategy better facilitates the extraction of sign language action features.

Multi-Grouping. As shown in Tab. 5, we present the final performance of four types of crops. The overall performance of each grouping type varies only slightly, mainly depending on the sample distribution of the dataset. However, integrating all four crops significantly enhances action classification performance. The performance of the bone-based grouping is slightly lower than that of joint-based results, primarily because our model is designed for points and is less sensitive to vector-based data.

SAM-SLR [21] ultimately using manually set weights to control the ensemble. However, we argue that such weights lack interpretability. As shown in Tab. 6, we employed two methods to integrate the results: *"acc"* refers to using the individual top-1 accuracy of each crop as the weight, while *"sum"* represents directly summing the outputs of the final layer of the model. Both methods resulted in an about 2% improvement in top-1 accuracy, demonstrating that our ensemble approach is both general and robust.

VSformer. For our two types of self-attention, we replace one with a linear layer to test the performance of the other individually. As shown in Tab. 8, the full VSformer achieves significantly better performance, indicating that features in both temporal and spatial dimensions are crucial for distinguishing sign language actions.

5. Conclusion

In this work, we design a skeleton recognition model tailored specifically for ISLR, termed **VSNet**. The core idea of our approach is to focus on the truly essential elements of sign language. To this end, we first introduce a weak joints dropping strategy to prioritize more critical points in the skeleton. We then propose the VSformer to capture the linguistic characteristics of sign language, where Visual Symbols (VS) represent smaller components of sign language actions. We conduct extensive experiments to validate the effectiveness of the proposed VSNet, achieving state-of-the-art performance across four benchmarks.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 62476052), Sichuan Science and Technology Program (No. 2024NSFSC1473), the open project of Sichuan Provincial Key Laboratory of Philosophy and Social Science for Language Intelligence in Special Education (No. YYZN-2023-3), the Humanities and Social Sciences Project of the Ministry of Education (No. 23YJA740013), and Central Guidance for Local Science and Technology Development Fund Projects (No. 2024ZYD0268).

References

[1] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE TMM*, 24:1750–1762, 2021. 1, 2

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*. Springer, 2020. 6

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 3

[4] Amanda Cardoso Duarte, Samuel Albanie, Xavier Giró Nieto, and Gül Varol. Sign language video retrieval with free-form textual queries. In *CVPR*, pages 14074–14084, 2022. 2

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2, 6, 7

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[7] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*, pages 5120–5130, 2022. 1

[8] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *ECCV*, pages 536–553. Springer, 2020. 2

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *NeurIPS*, 34:9355–9366, 2021. 3

[10] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 3

[11] Jeonghyeok Do and Munchurl Kim. Skateformer: Skeletal-temporal transformer for human action recognition. *arXiv preprint arXiv:2403.09508*, 2024. 3, 5

[12] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, pages 1110–1118, 2015. 1, 3

[13] Karen Emmorey. *Language, cognition, and the brain: Insights from sign language research*. Psychology Press, 2001. 1

[14] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *ICCV*, pages 11087–11096, 2021. 6, 7

[15] Hezhen Hu, Wengang Zhou, and Houqiang Li. Hand-model-aware sign language recognition. In *AAAI*, pages 1558–1566, 2021. 2, 6, 7

[16] Hezhen Hu, Wengang Zhou, Junfu Pu, and Houqiang Li. Global-local enhancement network for nmf-aware sign language recognition. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(3): 1–19, 2021. 2, 7

[17] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE TPAMI*, 45 (9):11221–11239, 2023. 6

[18] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Dynamic spatial-temporal aggregation for skeleton-aware sign language recognition. In *LREC/COLING*, 2024. 6, 7

[19] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE TCSVT*, 29(9):2822–2832, 2018. 2

[20] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI*, 2018. 2

[21] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *CVPR*, pages 3413–3423, 2021. 1, 3, 4, 5, 8

[22] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Sign language recognition via skeleton-aware multi-model ensemble. *arXiv preprint arXiv:2110.06161*, 2021. 6, 7

[23] Tao Jiang, Necati Cihan Camgoz, and Richard Bowden. Skeletor: Skeletal transformers for robust body-pose estimation. In *CVPR*, pages 3394–3402, 2021. 1, 3

[24] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, pages 196–214. Springer, 2020. 3, 4

[25] Trevor Johnston and Adam Schembri. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007. 1

[26] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018. 2, 7

[27] Alexander Kapitanov, Kvanchiani Karina, Alexander Nagaev, and Petrova Elizaveta. Slovo: Russian sign language dataset. In *International Conference on Computer Vision Systems*, pages 63–73. Springer, 2023. 2, 7, 8

[28] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020. 2

[29] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms. *IJCV*, 126: 1311–1325, 2018. 2

[30] Taeryung Lee, Yeonguk Oh, and Kyoung Mu Lee. Human part-wise 3d motion context learning for sign language recognition. In *ICCV*, pages 20740–20750, 2023. 1, 6

[31] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018. 1, 3

[32] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 2, 7, 8

[33] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *CVPR*, pages 6205–6214, 2020. 6

[34] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 3595–3603, 2019. 3

[35] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4804–4814, 2022. 7, 8

[36] Jiajia Liao, Yujun Liu, Yingchao Piao, Jinhe Su, Guorong Cai, and Yundong Wu. Gle-net: A global and local ensemble network for aerial object detection. *International Journal of Computational Intelligence Systems*, 15(1):2, 2022. 7

[37] Jinfu Liu, Xinshun Wang, Can Wang, Yuan Gao, and Mengyuan Liu. Temporal decoupling graph convolutional network for skeleton-based gesture recognition. *IEEE TMM*, 26:811–823, 2023. 3

[38] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 3, 7, 8

[39] Anna Mindess. *Reading between the signs: Intercultural communication for sign language interpreters*. Nicholas Brealey, 2014. 1

[40] Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. Automatic dense annotation of large-vocabulary sign language videos. In *ECCV*, pages 671–690. Springer, 2022. 2

[41] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *CVPR*, pages 11865–11874, 2021. 1, 3

[42] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *NeurIPS*, 35:14501–14515, 2022. 3

[43] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021. 2

[44] Wendy Sandler. Sign language and linguistic universals. *Cambridge University*, 2006. 1

[45] Noha Sarhan and Simone Frintrop. Unraveling a decade: a comprehensive survey on isolated sign language recognition. In *ICCV*, pages 3210–3219, 2023. 2

[46] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 3

[47] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE TIP*, 29:9532–9545, 2020. 2

[48] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *CVPR*, pages 1227–1236, 2019. 3

[49] Neil Song and Yu Xiang. Slgtformer: An attention-based approach to sign language recognition. *arXiv preprint arXiv:2212.10746*, 2022. 3, 4

[50] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *proceedings of the 28th ACM international conference on multimedia*, pages 1625–1633, 2020. 2

[51] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 3, 4

[52] Clayton Valli. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, 2000. 1

[53] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Scaling up sign spotting through sign language dictionaries. *IJCV*, 130(6):1416–1439, 2022. 2

[54] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *CVPR*, pages 499–508, 2017. 2

[55] Hanjie Wang, Xiujuan Chai, and Xilin Chen. A novel sign language recognition framework using hierarchical grassmann covariance matrix. *IEEE TMM*, 21(11):2806–2814, 2019. 2

[56] Chengcheng Wei, Jian Zhao, Wengang Zhou, and Houqiang Li. Semantic boundary detection with reinforcement learning for continuous sign language recognition. *IEEE TCSVT*, 31(3):1138–1149, 2020.

[57] Pan Xie, Mengyi Zhao, and Xiaohui Hu. Pisltrc: Position-informed sign language transformer with content-aware convolution. *IEEE TMM*, 24:3908–3919, 2021. 2

[58] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 1, 6, 7

[59] Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlslt: Towards multilingual sign language translation. In *CVPR*, pages 5109–5119, 2022. 1

[60] Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. Best: Bert pre-training for sign language recognition with coupling tokenization. In *AAAI*, pages 3597–3605, 2023. 6, 7

[61] Weichao Zhao, Hezhen Hu, Wengang Zhou, Yunyao Mao, Min Wang, and Houqiang Li. Masa: Motion-aware masked autoencoder with semantic alignment for sign language recognition. *IEEE TCSVT*, 2024. 6, 7

[62] Weichao Zhao, Wengang Zhou, Hezhen Hu, Min Wang, and Houqiang Li. Self-supervised representation learning with spatial-temporal consistency for sign language recognition. *arXiv preprint arXiv:2406.10501*, 2024. 6, 7

[63] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *CVPR*, pages 14890–14900, 2023. 6, 7