Dynamic Spatial-Temporal Aggregation for Skeleton-Aware Sign Language Recognition

Lianyu Hu, Liqing Gao, Zekang Liu, Wei Feng[*]

College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

{hly2021, lqgao, lzk100953, wfeng}@tju.edu.cn

Abstract

Skeleton-aware sign language recognition (SLR) has gained popularity due to its ability to remain unaffected by background information and its lower computational requirements. Current methods utilize spatial graph modules and temporal modules to capture spatial and temporal features, respectively. However, their spatial graph modules are typically built on fixed graph structures such as graph convolutional networks or a single learnable graph, which only partially explore joint relationships. Additionally, a simple temporal convolution kernel is used to capture temporal information, which may not fully capture the complex movement patterns of different signers. To overcome these limitations, we propose a new spatial architecture consisting of two concurrent branches, which build input-sensitive joint relationships and incorporates specific domain knowledge for recognition, respectively. These two branches are followed by an aggregation process to distinguishe important joint connections. We then propose a new temporal module to model multi-scale temporal information to capture complex human dynamics. Our method achieves state-of-the-art accuracy compared to previous skeleton-aware methods on four large-scale SLR benchmarks. Moreover, our method demonstrates superior accuracy compared to RGB-based methods in most cases while requiring much fewer computational resources, bringing better accuracy-computation trade-off. Code is available at `https://github.com/hulianyuyy/DSTA-SLR`

Keywords: Sign language recognition, skeleton data, dynamic feature graph aggregation, complex temporal movement patterns.

1. Introduction

Sign language is the primary means of communication for deaf individuals, conveyed through dynamic hand gestures, body posture, and facial expressions (Dreuw et al., 2007; Ong and Ranganath, 2005). However, understanding sign language can be challenging and time-consuming for the hearing people, requiring significant time and effort. Fortunately, recent advancements in machine learning and computer vision have made significant progress in the automatic interpretation of sign language. These technologies greatly facilitate communication for deaf individuals, enabling them to interact more smoothly with others in their daily lives.

Sign language recognition (SLR) can be broadly categorized into two types: vision-based SLR and skeleton-aware SLR. Vision-based SLR directly predicts signs from RGB image streams, which can be computationally expensive. Skeleton-aware SLR involves predicting a single sign label from a sequence of 2D or 3D skeleton representations. Compared to vision-based SLR, skeleton-aware SLR requires much fewer computations from 2D (Kay et al., 2017; Yan et al., 2018) or 3D (Liu et al., 2019; Shahroudy et al., 2016) skeleton inputs and is less affected by environmental factors like camera motion, lighting conditions, and changes in viewpoint (Liu et al., 2020),

which is thus more robust in extracting human action representations.

Current methods for skeleton-aware SLR (Tunga et al., 2021; Jiang et al., 2021b; Boháček and Hrúz, 2022) often use a combination of spatial graph modules and temporal modules to capture spatial and temporal features, respectively. However, there are two main limitations to these methods. Firstly, current methods rely on graph convolutional networks (GCNs) with a fixed graph structure or a single learnable graph, which can hinder the modeling of dynamic connections between human body joints that are specific to different input samples. Secondly, current methods typically use a simple temporal convolution to capture temporal information, which may not be able to effectively model the complex temporal dependencies present in sign language.

In this paper, we address the aforementioned limitations from two aspects. First, we propose to enhance the flexibility of spatial modules to comprehensively capture the correlations between different body joints. Specifically, we introduce a graph correlation module that dynamically builds spatial graphs for each input sample in a channel-wise manner, eliminating the need for a predefined graph based on the human body structure. Furthermore, we introduce virtual super nodes in addition to the physical human body joints to incorporate specific domain knowledge into the model. An

---

Corresponding author

adaptive graph is then constructed to distinguish the inherently important connections between human joints. Secondly, to capture the complex temporal movements of various input samples, we propose a parallel temporal convolutional module to aggregate multi-scale temporal information. This module consists of a series of temporal convolutions with varying receptive fields, which enables capturing human dynamics across a wide range of temporal durations.

With the advantages of high spatial and temporal capacity, our proposed method achieves new state-of-the-art accuracy on four commonly used SLR benchmarks. Especially, compared to vision-based methods which typically demand much higher computations and memory usage, our model achieves superior results in most cases with a better accuracy-computation trade-off, thereby bringing us one step closer towards real-life applications. Several ablation experiments confirm the effectiveness of our proposed modules.

2. Related Work

2.1. Sign Language Recognition

Sign language recognition (SLR) aims to classify an input video into a single sign label, which can be roughly divided into vision-based SLR and skeleton-aware SLR. Vision-based SLR usually receives RGB videos or RGB+D videos as inputs. Hand pose is used as guidance to pool spatio-temporal feature maps from different layers of a 3D CNN in (Hosain et al., 2021) to capture multi-cue features. SignBERT (Hu et al., 2021a) performs self-supervised pre-training on available sign data and incorporates hand prior in a model-aware method to model hierarchical context. SAM-SLR-v2 (Jiang et al., 2021a) proposes to fuse the representations of multiple modalities to enhance the representations. Despite high accuracy, these vision-based methods usually require a lot of computations. Skeleton-aware SLR methods receive skeleton sequences as inputs to perform recognition, which are more invariant to background information and more lightweight in 2D or 3D format compared to RGB streams, thus requiring fewer computational resources. Previous skeleton-aware SLR methods mostly rely on graph networks to process non-Euclidean skeleton data by treating human body joints as graph nodes. SAM-SLR (Jiang et al., 2021b) proposes a sign language graph convolutional network and a separable spatial-temporal convolutional network to exploit skeleton features. SLGTformer (Song, 2022) leverages learnable graph relative positional encodings and temporal twin self-attention to guide skeleton feature extraction. SPOTER (Boháček and Hrúz, 2022) introduces new normalization and augmentation techniques as well as a transformer-
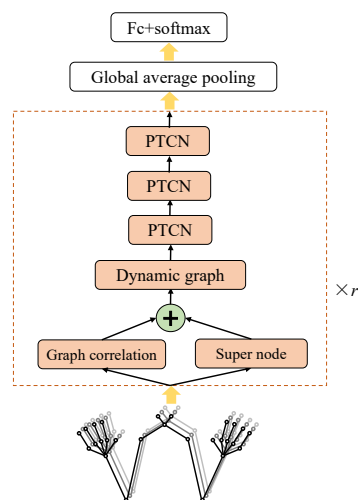

Figure 1: Overview for our proposed model.

based framework to handle skeleton data. We argue that the graph structure adopted in these methods may not well fit complex human dynamics and propose to dynamically build graph structures.

2.2. Skeleton-based Action Recognition

Skeleton-based action recognition aims to recognize human actions from a series of input skeleton sequences. Earlier methods always adopt convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to depict action representations by modeling human dynamics as a pseudo-image (Ke et al., 2017; Li et al., 2017, 2018a) or a series of coordinates along time (Du et al., 2015; Liu et al., 2016; Li et al., 2018b; Zhang et al., 2017; Si et al., 2019). Nevertheless, they overlook the internal relationships between joints which are better captured by graph networks due to their natural advantage over handling non-euclidean data. ST-GCN (Yan et al., 2018) has first proposed to model the human structure as a spatial graph by representing joints as nodes and bones as edges with a spatial-temporal graph convolutional network (GCN). However, the spatial receptive field in GCN is only limited to 1-hop neighbors and the edge weights are fixed in such conditions. Some late methods propose to leverage the higher polynomial order of the skeleton adjacency matrix to enlarge the spatial receptive field, where distant joints can exchange messages directly (Gao et al., 2019; Li et al., 2019). Some other methods propose to attach an attention-based adaptively computed graph along with the predefined graph to better exploit relationships between all joints and alleviate the predefined edge weights (Shi et al., 2020; Zhang et al., 2020; Xu et al., 2022; Ke et al., 2022; Su et al., 2021; Yang et al., 2021; Chen et al., 2021).

3. Methods

Fig 1 illustrates the overview of our proposed method, which consists of r blocks followed by a global average pooling layer, a fully connected (fc) layer, and a softmax function for recognition. Each block is first composed of a concurrent graph correlation module and a super node transform module to dynamically capture spatial joint relationships, followed by a dynamic graph aggregation module to distinguish inherently important joints. Three consecutive parallel temporal convolutional modules are followed to perform temporal modeling to capture complex human dynamics. In this section, we will first describe the process of building skeleton graphs and then provide a detailed explanation of our proposed method.

3.1. Graph Construction

Wearable motion capture devices such as Kinect V2 (Pagliari and Pinto, 2015) have been widely used in popular skeleton benchmarks such as NTU RGB+D 60 (Shahroudy et al., 2016), NTU RGB+D 120 (Liu et al., 2019) and NW-UCLA (Wang et al., 2014). However, these devices cannot provide precise annotations for the hands, which are critical for SLR. To address this limitation, we use a pretrained whole-body pose estimation network (HRNet (Sun et al., 2019) from MMPose (Contributors, 2020)) to offer annotations for people in the videos, resulting in 133 keypoints following the COCO format. To reduce the noise contained in the large number of nodes, we reduce the number of nodes to 27, which only contain 10 nodes for each hand and 7 nodes for the upper body, which is based on the observation that sign language is mainly conveyed by joints located in the upper human body.

Given the input skeleton sequences, adjacent keypoints are connected in the spatial dimension according to the natural connections of the human body to form spatial graphs in GCNs. Formally, a graph is represented as G=(V, E), where V = $\{v_1,...,v_N\}$ is a set of N graph nodes representing joints and E is a series of graph edges representing connectivity (bones) between joints. An adjacent matrix A with size $N \times N$ is adopted to depict the connectivity where $A_{i,j}$ represents the connection strength between node i and j. Especially, A is constructed as :

$$\mathbf{A}_{i,j} = \begin{cases} 1 & \text{if } d\left(v_i, v_j\right) = 1 \\ 0 & \text{else} \end{cases} \quad (1)$$

where $d(v_i, v_j)$ calculates the minimum distance between skeleton node v_i and v_j. The action sequences are represented as a feature tensor $x \in \mathcal{R}^{C \times T \times N}$ where each node $v_n \in N$ has a C dimensional feature vector over total T frames.
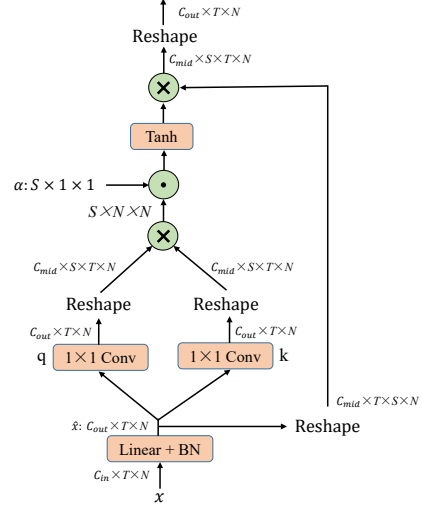

Figure 2: Overview for the proposed graph correlation module.

3.2. Graph Correlation Module

Recently, graph convolutional networks (GCNs) are broadly introduced to model human body dynamics in action recognition and SLR. Specifically, given the input tensor $x \in \mathcal{R}^{C \times T \times N}$, it performs layer-wise graph convolution with 1-hop normalized graph structure A at each time step as:

$$Y = \sigma(WXA) \quad (2)$$

where Y is the output and W is a trainable weight matrix. $\sigma(\cdot)$ acts as an activation function for output. In the graph convolutional network, the graph A is predefined according to human body structure in eq. 1 where each node can only aggregate information from 1-hop neighbors with fixed weights. However, distant nodes may still have significant correlations in sign language recognition, and these connections are not well captured by the fixed A. Moreover, the relationships between joints should be dynamically computed as different joints may play distinct roles in different sign language actions. We propose to dynamically compute joint relationships for each input sample to better capture important connections.

The overview of the graph correlation module is shown in fig. 2. In each layer, the input features $x \in \mathcal{R}^{C_{in} \times T \times N}$ pass through a linear layer and a BatchNorm layer to obtain $\widehat{x}$, transforming the channels into C_{out}. Afterwards, the features from the query and key go through two 1×1 convolutions to act as query and key, respectively. Then the features are reshaped into $\mathcal{R}^{C_{mid} \times S \times T \times N}$ to compute the adjacent matrices. In this process, S independent adjacent matrices are generated to encode various joint relationships to expand the network capacity, where $C_{mid}=C_{out}/S$. Finally, the features from both branches are multiplied to produce the expected adjacent matrices

$\widehat{A} \in \mathcal{R}^{S \times N \times N}$. To better represent the significance of each subset, we introduce a learnable parameter α that is randomly initialized and multiplied with the computed adjacent matrices. The resulting ajacent matrices $\widehat{A}$ are passed through a tanh activation function to transform their values into the range of [-1, 1]. Negative values in $\widehat{A}$ are expected to suppress corresponding messages, while positive values are expected to aggregate useful information from corresponding nodes. The reshaped $\widehat{x}$ is finallt multiplied with the adjacent matrices $\widehat{A}$ to aggregate information from other body joints. It's worth noting that in this process, the adjacent matrices are dynamically and independently computed for each input sample, getting rid of the predefined pattern of previous methods. Thus, it could dynamically build relationships with distant nodes, avoiding fixed weights within only 1-hop neighbors.

3.3. Super Node Transform Module

Currently, only human body joints are assigned as graph nodes, with specific physical meanings in practice. However, we argue that the network could also contain task-specific knowledge inherently to aid recognition. For instance, in SLR, we expect the network to contain the rules to perform signs and execute body movements to help recognition. To achieve this, we propose adding super (virtual) nodes in each layer and allowing input features to exchange information with them to absorb useful knowledge. The overview of our proposed module is illustrated in fig. 3.

Specifically, given the input features $x \in \mathcal{R}^{C_{in} \times T \times N}$ in each layer, they first pass through a LayerNorm layer and a linear layer to enhance their representations. These features are then multiplied with the super node features $u \in \mathcal{R}^{C_{in} \times E}$ to obtain the similarity matrix $\tilde{A} \in \mathcal{R}^{E \times T \times N}$. Here, E denotes the number of super nodes. To encode the importance of each super node, we introduce a randomly initialized learnable parameter $\beta \in \mathcal{R}^{E \times 1 \times 1}$, and multiply it with the similarity matrix. The similarity matrix $\tilde{A}$ then passes through a tanh activation function to transform its values into [-1, 1], whose negative values would suppress corresponding connections and positive values would enhance the received messages. Finally, the super node features $u \in \mathcal{R}^{C_{in} \times E}$ are multiplied with the computed similarity matrix $\tilde{A} \in \mathcal{R}^{E \times T \times N}$ to aggregate beneficial information from super nodes for each body joint, to absorb specific domain knowledge to help recognition.

3.4. Dynamic Graph Aggregation Module

In addition to the dynamic adjacent matrix that is computed independently for each input, we
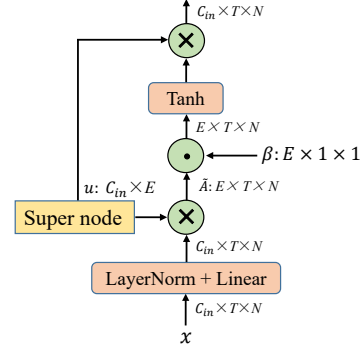

Figure 3: Overview for the proposed super node transform module.

also integrate static graphs into our model design, which are invariant over inputs to encode inherently important connections between body points. Practically, we introduce a fully learnable graph $P \in \mathcal{R}^{C_{in} \times N \times N}$ to capture joint relationships. P is randomly initialized and updated via backward-gradient-based network propagation, which is channel-wisely set to learn unique connection characteristics for each channel to encode body movements. To exchange messages between different graph nodes, we multiple input features $x \in \mathcal{R}^{C_{in} \times T \times N}$ with $P \in \mathcal{R}^{C_{in} \times N \times N}$ to let each node aggregate information from all other joints in a per-channel manner, resulting in output features $y \in \mathcal{R}^{T \times C_{in} \times N}$.

3.5. Parallel Temporal Convolution Module

Current techniques typically utilize a single temporal convolutional module with a large kernel to extract temporal information. Nonetheless, we argue that this approach may not be adequate to represent the complex movements when performing signs. For instance, various signers may perform signs with distinct speeds and body movements, and the action durations may vary significantly. Using a single temporal receptive field to capture such diverse movement patterns may not be optimal. Therefore, we propose a parallel temporal convolution module (PTCN) that consists of multiple temporal convolutions in parallel with different kernel sizes to capture these dynamics.

Fig. 4 depicts the architecture of our proposed module. It comprises L parallel branches with various kernel sizes, i.e., K_1, K_2, ..., K_L, to capture temporal information of various scales. Given the input features $x \in \mathcal{R}^{C_{in} \times T \times N}$, we split x along the channel dimension into L branches of $x_1, \cdots, x_L \in \mathcal{R}^{C_{in}/L \times T \times N}$ independently. In each of the L branches, such as the l_{th} branch, x_l undergoes a temporal convolution with a kernel size of K_l to capture temporal information. Then, a BatchNorm layer calibrates the representations. Finally, the
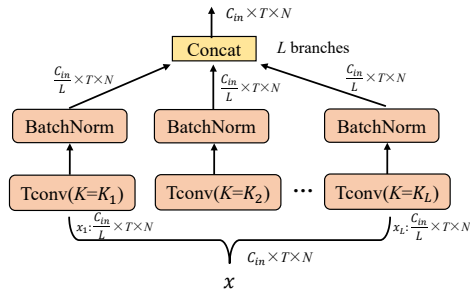
Figure 4: Overview for the proposed parallel temporal convolution module.

features from different branches are concatenated along the channel dimension as outputs to mix information from different scales.

As shown in fig 1, we serially stack three PTCNs to increase the temporal receptive field and enlarge the model capacity to capture complex human dynamics.

3.6. Multiple-Stream Fusion.

Following previous methods (Jiang et al., 2021a,b), we fuse beneficial information from multiple streams to perform recognition, including joint (original input data), bone, joint-motion and bone motion streams. For the bone stream, we obtain information by subtracting the source joint (which is close to the center of gravity of the skeleton) from the target joint (which is far away from the center of gravity) for each connected joint pair of the human body. Similarly, we obtain motion information by subtracting the difference between adjacent frames for either joint or bone streams. We calculate softmax scores for all four streams and then average them to give the final prediction.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

SLR500 (Zhang et al., 2016) is a Chinese sign language dataset recorded in the controlled lab environment with a solid-color background. It contains 500 words performed by 50 signers through 5 times. Totally, there are 125,000 videos in SLR500.

WLASL (Li et al., 2020a) is a challenging American Sign Language dataset collected from web videos with a vocabulary size of 2000 words. It contains 21,083 samples performed by 119 signers with unconstrained recording conditions. We follow previous methods to use four subsets of WLASL, i.e., WLASL100, WLASL300, WLASL1000 and WLASL2000 to evaluate our method.

Model configurations	Top-1(%)
Our model	**51.44**
w/o Graph correlation	49.25 (-2.19)
w/o Super node transform	50.82 (-0.62)
w/o Dynamic aggregation	50.23 (-1.21)
w/o PTCN	49.69 (-1.75)

Table 1: Ablations on the effectiveness of each proposed module.

MSASL (Joze and Koller, 2018) is an American sign language dataset, consisting of 16,054, 5,287, and 4,172 samples in the training, testing, and validation set, respectively. It's recorded in unconstrained real-life scenarios with a vocabulary size of 1,000. We follow previous methods to use four subsets of MSASL, i.e., MSASL100, MSASL200, MSASL500 and MSASL1000 to evaluate our method.

NMFs-CSL (Hu et al., 2021c) contains 25,608 and 6,402 samples for training and testing, respectively, with a vocabulary of 1067 words.

4.1.2. Training details

We use r=4 basic blocks, and set the channels for each block in [64, 128, 256, 512]. We randomly/centrally sample 120 frames out of 150 input frames as input during training/testing. Input data is mean normalized, added with random noise of maxmium value of 20 and augmented with 50% horizontally flipping (only during training). The temporal stride of 3rd PTCN in each block is set as 2 to decrease the sequence length. We adopt the SGD optimizer to train our model by total 250 epochs and learning rate 0.1, which is decreased by a factor of 10 after 150 and 200 epochs. Our model is trained with batch size of 24 with weight decay 0.0001 on a 3090 GPU in PyTorch framework. We follow SAM-SLR-v2 (Jiang et al., 2021a) to add a STC Module after each block to calibrate the input features.

4.2. Ablation Study

We conduct experiments on the WLASL2000 dataset with joint stream only to verify the effectiveness of our proposed modules.

Study on the effectiveness of proposed modules. Tab. 1 verifies the effectiveness of our proposed modules with an overall 51.44% accuracy on the WLASL2000 dataset. It has been observed that eliminating any of the suggested modules leads to a decrease in accuracy. Especially, it's noticed that the graph correlation module and PTCN have the greatest impact on accuracy, resulting in a boost of +2.19% and +1.75%, respectively, by distinguishing beneficial spatial features

Model configurations	Top-1(%)
GCN	47.51
Learnable graph	49.86
Graph correlation module	**51.44**
$S = 1$	50.64
$S = 4$	50.91
$S = 8$	**51.44**
$S = 16$	50.93

Table 2: Ablations on the configuration of graph correlation module.

Model configurations	Top-1(%)
$E = 1$	50.98
$E = 6$	**51.44**
$E = 12$	51.26
$E = 24$	50.84
w/o β	50.96
Tanh $\rightarrow$ softmax	50.34

Table 3: Ablations on the configuration of super node transform module.

from other nodes or capturing complex human dynamics.

Study on the architecture of graph correlation module. In the upper part of tab. 2, we compare our proposed graph correlation module against other spatial aggregation architectures, e.g., GCN and a learnable graph. GCN aggregates spatial information from 1-hop neighbors with fixed weights. The learnable graph uses an adaptive adjacent matrix for all inputs. It's noticed that our proposed graph correlation module outperforms the other two architectures by a large margin in accuracy. We then examine the effects of the number of generated adjacent matrices S, and find that larger values of S lead to better performance, reaching a peak at 8. We thus set S=8 by default.

Study on the configurations of super node transform module. We first demonstrate the impact of varying the number of super nodes E. It's observed that larger E achieves better performance, which reaches a peak after equalling 6. Therefore, we set E=6 by default. Next, we evaluate the effectiveness of several proposed components in the super node transform module. Removing β leads to a decrease in accuracy, as it dynamically adjusts the importance weights of different super nodes. Additionally, replacing the tanh activation function with softmax results in decreased accuracy. The tanh activation function is crucial for highlighting important features with positive outputs and suppressing irrelevant features with negative outputs.

Model configurations	Top-1(%)
Kernels = [3,5,7,9]	50.05
Kernels = [3,5]	50.22
Kernels = [3,9]	51.05
Kernels = [5,7]	**51.44**
Kernels = [7,9]	51.03
1*PTCN	49.64
2*PTCN	50.52
3*PTCN	**51.44**
4*PTCN	50.66

Table 4: Ablations on the configuration of parallel temporal convolution module.

Study on the configurations of parallel temporal convolution module (PTCN). The upper part of tab. 4 investigates the architecture of PTCN. We find that a two-branch architecture using kernel sizes of [5,7] achieves better performance than other options, such as two branches or four branches. We believe that larger temporal kernels (e.g., 9) may introduce excessive noise, while smaller kernels may not adequately capture complex temporal movements. We then consider the number of stacked PTCNs. The bottom part of tab. 4 reveals that increasing the number of PTCNs leads to consistent improvements in performance, which reaches a peak after using 3 PTCNs. Thus, we use 3 PTCNs as the default setting.

4.3. Comparison with the state-of-the-art

WLASL. Tab. 5 presents a comparison of our method with state-of-the-art approaches on the WLASL dataset. For fair comparison, we follow the previous methods to divide this dataset into four subsets, including WLASL100, WLASL300, WLASL1000, and WLASL2000, and evaluate using two accuracy metrics: per-instance accuracy (P-I) and per-class accuracy (P-C). The upper part of Tab. 5 compares our method with other skeleton-aware methods, while the bottom part compares it with other RGB-based methods. Our method achieves new state-of-the-art performance across all four subsets, surpassing other skeleton-aware approaches by a large margin, thanks to its superior ability to dynamically aggregate joint features and capture complex human dynamics. We especially find that our method outperforms other RGB-based approaches in most cases, despite consuming fewer computation. These results demonstrate the effectiveness and efficiency of our proposed method.

MSASL. Tab. 6 compares our method with state-of-the-art approaches on the MSASL dataset, which is divided into four subsets in-

Methods	WLASL100 P-I(%)	WLASL100 P-C(%)	WLASL300 P-I(%)	WLASL300 P-C(%)	WLASL1000 P-I(%)	WLASL1000 P-C(%)	WLASL2000 P-I(%)	WLASL2000 P-C(%)
Skeleton-aware								
SignBERT (Hu et al., 2021a)	79.07	80.05	70.36	71.17	-	-	47.46	45.17
HMA (Hu et al., 2021b)	-	-	-	-	-	-	46.32	44.09
SAM-CLR-v2 (Jiang et al., 2021a)	-	-	-	-	-	-	51.50	-
SLGTformer (Song, 2022)	-	-	-	-	-	-	47.42	-
BEST (Zhao et al., 2023)	77.91	77.83	67.66	68.31	-	-	46.25	43.52
RGB-based								
I3D (Li et al., 2020a)	65.89	-	56.14	-	47.33	-	32.48	-
Fusion-3 (Hosain et al., 2021)	75.67	-	68.30	-	56.68	-	38.84	-
TCK (Li et al., 2020b)	77.52	-	68.56	-	-	-	-	-
BEST (Zhao et al., 2023) (R+P)	81.63	81.01	75.60	76.12	-	-	**54.59**	**52.12**
Ours	**82.38**	**83.09**	**79.97**	**80.56**	**67.76**	**67.54**	53.68	51.17

Table 5: Comparison with other methods on the WLASL dataset. 'R+P' denotes the fused results of RGB and pose modalities.

Methods	MSASL100 P-I(%)	MSASL100 P-C(%)	MSASL200 P-I(%)	MSASL200 P-C(%)	MSASL500 P-I(%)	MSASL500 P-C(%)	MSASL1000 P-I(%)	MSASL1000 P-C(%)
Skeleton-aware								
SignBERT (Hu et al., 2021a)	81.37	82.31	77.34	78.02	-	-	59.80	57.06
HMA (Hu et al., 2021b)	78.57	79.48	72.19	73.52	-	-	56.02	52.98
BEST (Zhao et al., 2023)	80.98	81.24	76.60	76.75	-	-	58.82	54.87
RGB-based								
I3D+BLSTM (Li et al., 2020a)	72.07	-	-	-	-	-	40.99	-
TCK (Li et al., 2020b)	83.04	83.91	80.31	81.14	-	-	-	-
BSL (Albanie et al., 2020)	-	-	-	-	-	-	64.71	61.55
Ours	**84.16**	**84.54**	**81.58**	**81.89**	**75.06**	**74.68**	**65.74**	**62.31**

Table 6: Comparison with other methods on the MSASL dataset. 'R+P' denotes the fused results of RGB and pose modalities.

Methods	Top-1(%)
Skeleton-aware	
ST-GCN (Yan et al., 2018)	90.0
SignBERT (Hu et al., 2021a)	96.6
HMA (Hu et al., 2021b)	95.9
RGB-based	
3D-R50 (Qiu et al., 2017)	95.1
GLE-Net (Hu et al., 2021c)	96.8
SignBERT (Hu et al., 2021a) (R+P)	97.6
BEST (Zhao et al., 2023) (R+P)	97.7
Ours	**98.1**

Table 7: Comparison with other methods on the SLR500 dataset. 'R+P' denotes the fused results of RGB and pose modalities.

cluding MSASL100, MSASL200, MSASL500 and MSASL1000. The upper and bottom parts of tab. 6 compare our method against other skeleton-aware methods and RGB-based methods, respectively. It's observed that our method achieves much better accuracy in both metrics across all four subsets than other skeleton-aware approaches. Moreover, our method outperforms RGB-based approaches in most cases.

SLR500 and NMFs-CSL. Tab. 7 and tab. 8 compare our method against other skeleton-aware and RGB-based methods on the SLR500 and NMFs-CS datasets, respectively. As a skeleton-aware approach, our method beats both skeleton-aware methods and RGB-based methods, verifying its effectiveness.

4.4. Visualizations

Fig. 5 depicts the edges computed by the graph correlation module to demonstrate its efficacy in capturing human dynamics for the "Clothes" and "No" signs. Only the edges with top 5% weights are displayed. It is observed that the generated edges primarily connect the center joint (neck) with
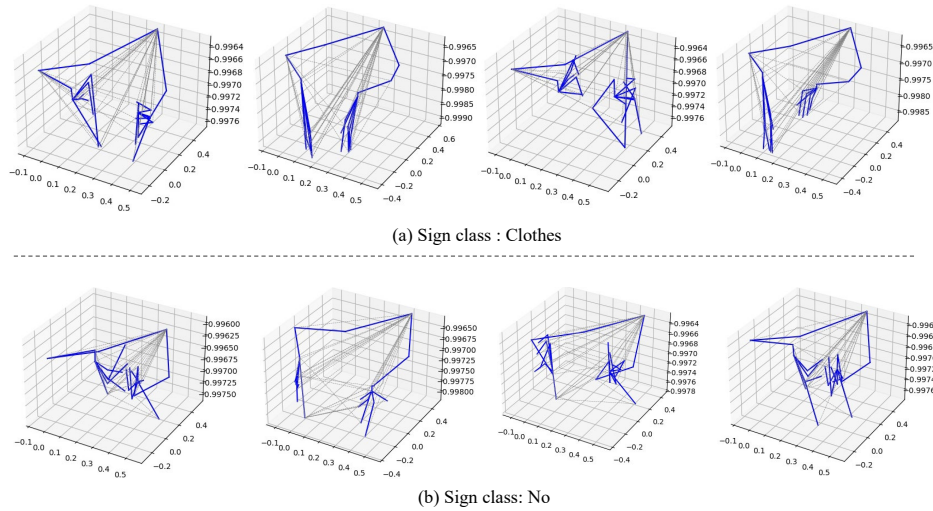
5456

(a) Sign class : Clothes

(b) Sign class: No

Figure 5: Visualizations of learned edges in the graph correlation module for two signs. Only the top 5% active connections are plotted. It's observed that our method learns to build dynamic joint relationships for different input samples, and especially pays attention to the joints of both hands to capture sign movements.

Methods	Top-1(%)
Skeleton-aware	
ST-GCN (Yan et al., 2018)	59.9
HMA (Hu et al., 2021b)	71.7
SignBERT (Hu et al., 2021a)	74.9
BEST (Zhao et al., 2023)	68.5
RGB-based	
I3D (Carreira and Zisserman, 2017)	64.4
TSM (Lin et al., 2019)	64.5
Slowfast (Feichtenhofer et al., 2019)	66.3
GLE-Net (Hu et al., 2021c)	69.0
HMA (Hu et al., 2021b) (R+P)	75.6
Ours	**77.9**

Table 8: Comparison with state-of-the-art methods on the NMFs-CSL dataset. 'R+P' denotes the fused results of RGB and pose modalities.

joints located in both hands for both signs. Additionally, our approach learns to establish connections between the joints of the left and right hands to create distant relationships. We conclude that our method can develop dynamic joint relationships for various input samples and, in particular, focuses on the joints of both hands to capture sign movements.

4.5. Efficiency

4.5.1. Model parameters

Tab. 9 presents a comparison between our proposed method and two recent skeleton-aware approaches, namely SAM-CLR-v2 (Jiang et al., 2021a) and SPOTER (Boháček and Hrúz, 2022).
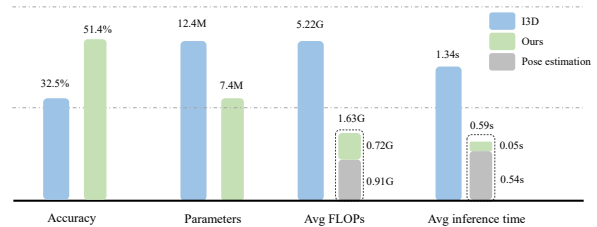

Figure 6: Comparison of our method (with joint stream only) and I3D (Li et al., 2020a) upon accuracy, parameters, average FLOPs, and average inference time on the WLASL2000 dataset.

We observe that while the parameters of our method fall between those of the other approaches, our method significantly outperforms them in terms of accuracy. This demonstrates the effectiveness and efficiency of our proposed approach.

4.5.2. Comparison with RGB-based methods

Fig. 6 presents a comparison between our method and a popular RGB-based approach, I3D (Li et al., 2020a), using various metrics such as accuracy, parameters, FLOPs, and inference time on the WLASL2000 dataset to highlight the advantages of our method in real-world applications. The figure shows that our method outperforms I3D in terms of accuracy, while having lower parameters, much fewer average FLOPs per video, and less average inference time per video. Compared to RGB-based approaches, our method demonstrates great advantages in both accuracy and efficiency by taking skeleton data as input, demonstrating its superiority in real-life scenarios.

Methods	Parameters(M)	Top-1(%)
SAM-CLR-v2	4.3	45.61
SLGTformer	9.3	47.42
Ours	7.4	**51.44**

Table 9: Comparison of our method with other skeleton-aware approaches upon both accuracy and parameters on the WLASL2000 dataset with joint stream only.

5. Conclusion

In this paper, we address the shortcomings of previous skeleton-aware sign language recognition methods by introducing two novel approaches. Firstly, we dynamically construct joint relationships to gather useful spatial features. Secondly, we present a parallel temporal convolution module to capture intricate human dynamics. Our method achieves new state-of-the-art results on four widely-used benchmarks and even outperforms RGB-based approaches in terms of both effectiveness and efficiency in most cases.

6. Acknowledgements

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer.

Matyáš Boháček and Marek Hrúz. 2022. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 182–191.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channelwise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368.

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

Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. 2007. Speech recognition techniques for a sign language recognition system. *hand*, 60:80.

Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Xiang Gao, Wei Hu, Jiaxiang Tang, Jiaying Liu, and Zongming Guo. 2019. Optimized skeleton-based action recognition via sparsified graph regression. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 601–610.

Al Amin Hosain, Panneer Selvam Santhalingam, Parth Pathak, Huzefa Rangwala, and Jana Kosecka. 2021. Hand pose guided 3d pooling for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3429–3439.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021a. Signbert: pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11087–11096.

Hezhen Hu, Wengang Zhou, and Houqiang Li. 2021b. Hand-model-aware sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1558–1566.

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

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

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021b. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3413–3423.

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

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Lipeng Ke, Kuan-Chuan Peng, and Siwei Lyu. 2022. Towards to-at spatio-temporal focus for skeleton-based action recognition. *arXiv preprint arXiv:2202.02314*.

Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. 2017. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297.

Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He. 2017. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 601–604. IEEE.

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

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. 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.

Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. 2020b. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6205–6214.

Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3595–3603.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018b. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5457–5466.

Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093.

Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*.

Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer.

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152.

Sylvie CW Ong and Surendra Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(06):873–891.

Diana Pagliari and Livio Pinto. 2015. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. *Sensors*, 15(11):27569–27589.

Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale

dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545.

Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. 2019. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1227–1236.

Neil Song. 2022. Slgtformer: An attention-based approach to sign language recognition. *arXiv preprint arXiv:2212.10746*.

Yukun Su, Guosheng Lin, and Qingyao Wu. 2021. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13328–13338.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703.

Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan Wachs. 2021. Pose-based sign language recognition using gcn and bert. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 31–40.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. 2014. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656.

Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. 2022. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2866–2874.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*.

Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. 2021. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13423–13433.

Jihai Zhang, Wengang Zhou, Chao Xie, Junfu Pu, and Houqiang Li. 2016. Chinese sign language recognition with adaptive hmm. In *2016 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE.

Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126.

Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1112–1121.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. Best: Bert pre-training for sign language recognition with coupling tokenization. *arXiv preprint arXiv:2302.05075*.