A Multimodal Spatio-Temporal GCN Model with Enhancements for Isolated Sign Recognition

Yang Zhou[1], **Zhaoyang Xia**[1], **Yuxiao Chen**[1], **Carol Neidle**[2], **Dimitris N. Metaxas**[1]

[1] Rutgers University, [2] Boston University
[1] 110 Frelinghuysen Road, Piscataway, NJ 08854,
[2] Boston University Linguistics, 621 Commonwealth Ave., Boston, MA 02215
{eta.yang, zx149}@rutgers.edu, yc984@cs.rutgers.edu, carol@bu.edu, dnm@cs.rutgers.edu

Abstract

We propose a multimodal network using skeletons and handshapes as input to recognize individual signs and detect their boundaries in American Sign Language (ASL) videos. Our method integrates a spatio-temporal Graph Convolutional Network (GCN) architecture to estimate human skeleton keypoints; it uses a late-fusion approach for both forward and backward processing of video streams. Our (core) method is designed for the extraction—and analysis of features from—ASL videos, to enhance accuracy and efficiency of recognition of individual signs. A Gating module based on per-channel multi-layer convolutions is employed to evaluate significant frames for recognition of isolated signs. Additionally, an auxiliary multimodal branch network, integrated with a transformer, is designed to estimate the linguistic start and end frames of an isolated sign within a video clip. We evaluated performance of our approach on multiple datasets that include isolated, citation-form signs and signs pre-segmented from continuous signing based on linguistic annotations of start and end points of signs within sentences. We have achieved very promising results when using both types of sign videos combined for training, with overall sign recognition accuracy of 80.8% Top-1 and 95.2% Top-5 for citation-form signs, and 80.4% Top-1 and 93.0% Top-5 for signs pre-segmented from continuous signing.

Keywords: ASL, GCN, Gating module, Temporal action localization

1. Introduction

In the US, it is estimated that 28 million people are Deaf or hard of hearing (Lin et al., 2011), and that about 500,000 use American Sign Language (ASL) as their primary language (Mitchell et al., 2006). ASL is also the 3rd most studied non-native language (Looney and Lusin, 2021). Signed languages are full-fledged natural languages, with information expressed in the visual-gestural modality by movements of the arms, hands, head, and upper body, and by facial expressions. They generally lack a standardized written form.

Computer-aided sign language analytics and sign recognition from video have many potential applications, which include resources to provide/enhance access to digital materials for signers, and tools for sign language learners (including hearing parents of deaf children) and interpreters, for ASL-to-English translation, and for improved sign language research. Research in this area is challenging, however, in part because of the complexity and variability of sign production and the fact that information expressed across the relevant channels may differ in spatio-temporal scale. For example, grammatical information conveyed non-manually by facial expressions and head gestures may extend over phrasal domains, i.e., it may occur over a scope that includes more than one sign.

In this paper, we focus on the recognition of individual signs—both isolated, citation-form signs and signs pre-segmented from continuous signing. This is a critical step towards recognition of signs directly from sentences. Sign production in continuous signing differs somewhat from production of citation-form signs (Neidle, 2023), so it is particularly significant that we are able to achieve a high degree of success also for recognition of pre-segmented signs trained on the combined dataset.

One major challenge is the existence of both inter- and intra-signer variations in sign production. Another significant challenge results from the fact that different classes of signs (e.g., lexical signs, fingerspelled signs, and classifiers) have significantly different internal structures. Addressing these challenges requires extensive video datasets with diverse signers and consistent gloss labeling of signs, to train computational models effectively. We utilize multiple datasets shared on the Web by the American Sign Language Linguistic Research Project (ASLLRP) (Neidle et al., 2022b)—specifically, their collections of **isolated, citation-form signs** (ASLLVD (Neidle and Metaxas, 2023b), DSP (Neidle and Metaxas, 2023c), and RIT (Neidle and Metaxas, 2023e)), and of **signs pre-segmented from continuous signing** based on linguistic annotations that include information about the linguistic start and end points of these signs within sentences (ASLLRP Sentences (Neidle and Metaxas, 2023a) and DSP Sentences (Neidle and Metaxas, 2023d))—as well

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 408–419
25 May 2024.

isolated sign data from WLASL (Li et al., 2020), with annotations provided by ASLLRP to ensure consistent labeling (Neidle et al., 2022a; Neidle and Ballard, 2022). Taken together, this collection includes 21,083 videos with over 2,000 distinct signs from 119 signers, with consistent gloss labeling and a focus on lexical signs. This collection, which will be referred to in this paper as the ”ASLLRP Individual Sign Collection,” forms the basis for our experiments to advance sign recognition using deep learning techniques.

Prior to the advent of deep learning methods, traditional machine learning methods such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) were employed to capture the spatio-temporal aspects of sign language (Lafferty et al., 2001; Grobel and Assan, 1997; Dilsizian et al., 2014). Recent advances in deep learning, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), have opened new avenues towards the automated recognition of signs from large vocabularies without the manual identification of features in the video. However, several challenges remain. For example: (1) Many of the available video resources have poor spatio-temporal resolution; (2) There are many different types of signs, with different internal composition, and some types, such as classifier constructions (which incorporate some degree of iconicity) do not constitute a fixed vocabulary; (3) The size of the data is relatively small, compared to spoken-language datasets; and (4) There is no 1-1 correspondence between ASL signs and English words, and no agreed-upon convention for providing English-based gloss labels to uniquely identify ASL signs. In this paper, we present results for recognition of individual ASL lexical signs, using the largest-to-date dataset that includes both isolated signs and signs pre-segmented from continuous signing. For more precise recognition, we have also developed a new approach to detect the beginning and end of an isolated, citation-form sign within a video clip.

2. Overview of our Approach

To achieve accurate sign recognition from video, we propose a deep learning approach based on skeletons. This method involves detecting start and end frames of the signs, and it leverages parameters from the skeleton data. Using a bidirectional learning framework within a Graph Convolutional Network (GCN) architecture, our method achieves notable accuracy on the ASLLRP Individual Sign Collection and WLASL data.

To improve sign recognition accuracy for the set of isolated signs, a Gating module designed to evaluate temporal weights has been embedded to enable the network to focus on the significant frames in the video clips, while avoiding frames that contain blurring or other artifacts often present in videos. To further enhance the feature extraction model, we designed an auxiliary multimodal branch network for temporal action localization based on an encoder and transformers. With training based on linguistic annotations of start and end frames in the ASLLVD and DSP isolated sign datasets, the auxiliary branch utilizes spatio-temporal features extracted by the GCN and the encoded handshape information, to detect the start and end points of isolated signs. The resulting improvements in sign recognition accuracy are shown in Section 5.3.3.

3. Related Work

Before the advent of deep learning techniques, sign language recognition research relied primarily on handcrafted features, such as the positioning and movement of hands relative to specific body parts (Tornay et al., 2020; Cooper et al., 2012; Badhe and Kulkarni, 2015; Xiaohan Nie et al., 2015), combined with standard classifiers like Support Vector Machines (SVMs), k-Nearest Neighbors (kNNs), Conditional Random Fields (CRFs), and Hidden Markov Models (HMMs) (Memiş and Albayrak, 2013; Dardas and Georganas, 2011; Yang, 2010; Metaxas et al., 2018; Tornay et al., 2020). However, these handcrafted features and underlying Gaussian distribution assumptions limited the systems' capabilities for generalization and scalability. Recently, deep neural network methods have made breakthroughs in computer vision tasks, such as action and gesture recognition; these methods have also been applied to sign language recognition, a more difficult problem given the complexity of linguistic structure (Rastgoo et al., 2021; Jiang et al., 2021).

Some recent research has used transfer learning methods for isolated sign recognition, since available sign language datasets have vocabularies that are small compared to those of general-purpose human motion databases like Kinetics400 (Carreira and Zisserman, 2017). Such approaches are discussed by Sandoval-Castañeda et al. (2023), who attained best results using a visual transformer pretrained first on human action videos in Kinetics400, and then on OpenASL (Shi et al., 2022) videos (following Wei et al. (2022)). They fine-tuned on the WLASL (Li et al., 2020) dataset—with modified glossing (as in Dafnis et al., 2022b; Neidle et al., 2022a; Neidle and Ballard, 2022). They also leveraged phonological features extracted from ASL-LEX 2.0 (Sevcikova Sehyr et al., 2021), to ”better characterize video models and pre-training tasks.” See further discussion in Section 5.4.

3.1. RGB-based Approaches

In sign recognition, RGB-based approaches have undergone a significant evolution with the rise of deep learning. Initially, these methods focused on extracting spatial features from RGB frames using traditional image processing techniques. The introduction of Convolutional Neural Networks (CNNs) marked a significant advance, allowing for more efficient and nuanced extraction of spatial features directly from RGB data.

Pioneering work by Krizhevsky et al. (2012) and Simonyan and Zisserman (2014) showcased the effectiveness of CNNs in automated image feature extraction (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), laying the groundwork for applying these networks to sign language recognition. These CNN models are adept at analyzing shapes, movements, and orientations of hands and body parts, critical for sign recognition. However, the challenge in sign recognition extends beyond spatial to temporal feature extraction. This led to the integration of CNNs with Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, known for their ability to capture temporal dynamics in sequences, as described by Hochreiter and Schmidhuber (1997). Further advances were achieved with 3D Convolutional Neural Networks (3D-CNNs), which, as explored by Ji et al. (2013), extract spatio-temporal features from video sequences, offering a more holistic approach to gesture recognition. More recent studies have investigated use of attention mechanisms, particularly in Transformer models (Vaswani et al., 2017), for sign recognition. These mechanisms focus on specific segments of video frames, enhancing recognition accuracy by highlighting critical sign language features.

Despite these technological advances, RGB-based methods still face challenges, in part because of sensitivity to lighting conditions, foreground-background complexities, and possible lack of focus on the important parts of the human body. This translates to an increased need for training data, which are unavailable in real-world settings. Our model-based approach aims to overcome these limitations, enhancing the robustness and applicability of sign language recognition systems in various real-world settings.

3.2. Skeleton-based Approaches

Skeleton-based approaches for action and sign language recognition have significantly evolved, focusing on extracting and analyzing body keypoints or skeleton graphs. Facilitated by advanced human pose estimation technologies, this methodology prioritizes essential movement features while excluding irrelevant background noise. Initial research utilized Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture temporal aspects of actions (Soo Kim and Reiter, 2017; Liu et al., 2017). However, these models struggled with encoding spatial and temporal interactions between keypoints.

Addressing these limitations, Yan et al. (2018) introduced the Spatial Temporal Graph Convolutional Network (ST-GCN), showcasing the potential of Graph Convolutional Networks (GCNs) in learning skeleton dynamics. Despite this innovation, ST-GCNs, focusing on direct joint connections, overlooked critical indirect keypoint interactions, which are essential for comprehensive sign recognition. Efforts to surmount this challenge included Li et al.'s (2019a) exploration of latent connections and Shi et al.'s (2019b; 2020) multi-stream approaches that enhanced action recognition by integrating keypoints, bones, and their motion. Additionally, de Amorim et al. (2019) adapted the ST-GCN framework for sign recognition, achieving approximately 60% accuracy in recognizing a limited vocabulary of signs.

Further advances are exemplified by Jiang et al. 2021, which implemented a pose-based GCN with additional modalities like RGB frames and optical flow, resulting in significant progress in isolated sign recognition. Dafnis et al. (2022a) extended these approaches by incorporating forward and backward data streams with keypoints and bones acceleration, significantly improving recognition accuracy on the WLASL dataset.

4. Methodology

The human body can be represented as a graph with nodes consisting of the face, upper body, arms, and hands. For sign recognition, all these parts are important and need to be used. Therefore, our approach extracts this information from video based on the following three components: (1) a spatio-temporal Graph Convolutional Network (GCN) architecture, for detailed modeling of skeleton keypoints from a signer's video; (2) a late-ensemble technique to synergistically combine, in the GCN, the forward and backward video streams, for improved sign recognition; and (3) an Encoder and Transformer-based approach, for precise temporal motion localization of the beginning and end frames of a sign.

4.1. Spatio-temporal Graph Convolutional Network

Our goal is to capture and analyze the complex spatio-temporal movement dynamics of the arms and hands during signing. To achieve this, our method first extracts keypoints and bones from the torso, arms, and hands using Alphapose, as developed by Fang et al. (2017). This method is capable of estimating 136 keypoints for the entire body from single RGB images. Using this model, we constructed a skeletal graph consisting of 27 nodes. These keypoints and respective bones are integrated within a GCN using spatio-temporal graph convolutions. Our model's spatial convolutions are computed based on the spatial partitioning strategy described in the ST-GCN framework by Yan et al. (2018). The integration of spatio-temporal graph convolutions enables our model not only to capture the spatial relationships between keypoints and bones, but also to estimate their temporal evolution over time. This dual capability showcases the unique advantage of the ST-GCN framework in capturing both spatial intricacies and temporal variations. The spatial formulation of our GCN model is delineated as follows:

$$x_{\text{out}} = \Lambda^{-\frac{1}{2}}(I+A)\Lambda^{-\frac{1}{2}}x_{\text{in}}W, \qquad (1)$$

where x_{in} in the GCN input consists of keypoints, bones, and other related information, while x_{out} denotes the output feature matrix derived from the graph convolution process. Matrix A models the intra-body connections (bones), while the identity matrix I models self-connections (keypoints). Λ is a diagonal matrix derived from $(I + A)$, and W is the ST-GCN weight matrix (2018). For purposes of our proposed application, the spatial graph convolutions are modeled using 2D convolution operations; the result, $x_{\text{in}}W$, is then multiplied by the normalized term $\Lambda^{-\frac{1}{2}}(I+A)\Lambda^{-\frac{1}{2}}$ to compute x_{out}

The right of Figure 1 shows the ST-GCN network architecture. Notably, a Gating module is appended to the end of the network, specifically focusing on important frames in isolated sign videos. The middle of Figure 1 illustrates the architecture of each of the GCN Blocks. It is composed of a Decoupled Spatial GC, STC Attention, a Temporal GC, and a series of Batch Normalization (BN) layers along with ReLU activation functions. The entire GCN Block includes a tail concatenation in the form of a residual structure to preserve low-level feature information. Drop Graphs are used in certain locations to prevent overfitting. The left part of Figure 1 provides details of the STC Attention Block, which consists of three attention modules: Spatial Attention, Temporal Attention, and Channel Attention, each with a tail concatenation to model the residual structure.

The Gating module in our approach is designed to identify and remove frames that are not useful for recognizing the sign, such as those with blurring or extraneous movements. We achieve this by

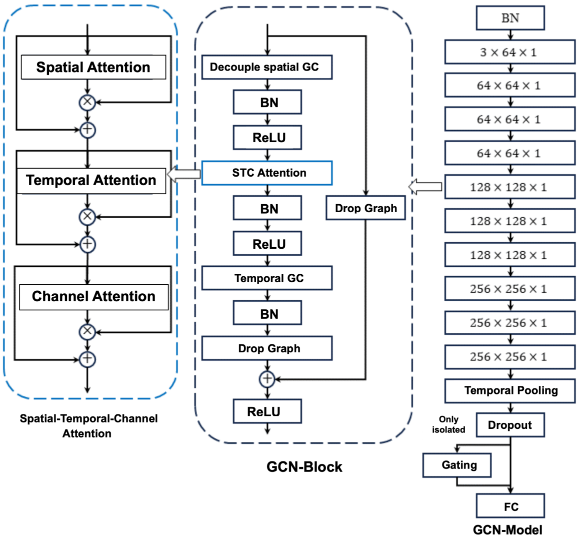

Figure 1: The ST-GCN Network Architecture

designing a multilayer convolution-based temporal attention module, to identify and remove those non-informative frames, as shown at the top of Figure 2. In this module, the skeleton feature dimension computed from the previous layers is reduced using a 3-layer stack of convolutions; a sequence of weights related to the temporal dimension is obtained by a temperature softmax layer (Hinton et al., 2015). The skeleton features computed from the previous layers are then multiplied with the output of the softmax layer in the Gating Block. Using this Gating Block, the network focuses, in the case of isolated signs, on those frames that carry valid information for sign recognition.

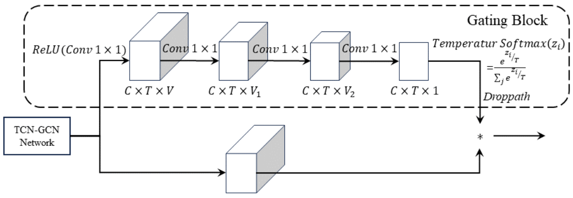

Figure 2: Gating Module Architecture

4.2. Bidirectional Stream GCN

Drawing inspiration from the multi-stream approach used in Shi et al. 2020, our methodology incorporates both forward and backward directions of video frame sequences for two types of data inputs: the location coordinates of the skeleton keypoints, and the bone vectors. To represent the bone vectors in our graph, we designate the nose as the root keypoint. Subsequent bone vectors are computed by tracing the connections between consecutive skeletal keypoints, starting from this root. As shown in Figure 3, the temporal data from the skeleton are processed with respect to two types of input: joints and bones; these are then input into the forward stream. Subsequently, the temporal dimension is reversed and input into the backward stream. Then an ensemble from the predictions of the four models gives rise to a final prediction for the sign, as shown in Figure 3.

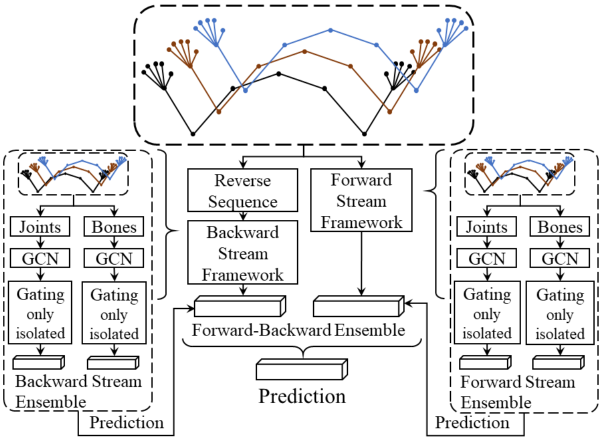

Figure 3: Bidirectional Stream GCN Architecture

4.2.1. Score Fusion

As mentioned previously, our proposed framework uses two types of information streams, specifically joints and bones. We use their forward and backward directions to arrive at an improved consolidated prediction. We first integrate the prediction scores from these streams within each direction by using the softmax scores from each stream, as described by Shi et al. (2019a; 2019b; 2020); Cai et al. (2021); and Dafnis et al. (2022a), to calculate an optimized weighted sum of the scores pertinent to each direction. This process is then replicated for the fusion of prediction softmax scores from both directions; an optimized weighted summation is computed to predict the sign labels.

4.2.2. Temporal Action Localization

To locate the start and end frames of isolated signs and thereby improve sign recognition, we design an auxiliary multimodal branch network. We train, using, in the loss function, linguistic annotations (which include the start and end frames of signs, and the handshapes in those frames), to learn to identify the start and end frames of a given isolated sign. As shown in Figure 4, the GCN network architecture is used to extract spatio-temporal features. Additionally, up to four types of handshapes for each sign video—Dominant start handshape, Dominant end handshape (and, for 2-handed signs, also Non-dominant start handshape and Non-dominant end handshape)—and the video are input into the network via a custom encoder. These are then processed through a transformer layer to improve the temporal positional dependence and interpretability of the handshapes. The extracted features are concatenated with the features extracted by the GCN using a Temperature Softmax to predict the start and end frames of the isolated sign.

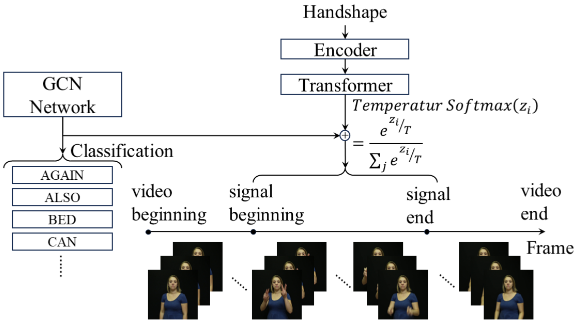

Figure 4: Auxiliary Multimodal Branch Architecture for Action Localization

5. Experiments

5.1. Data Preprocessing

Following the dataset partitioning strategy outlined in Dafnis et al. 2022b and Li et al. 2020, we divided the dataset into training, validation, and testing subsets. The division was carried out in a ratio of approximately 4:1:1 for each sign category; hence we further restrict these datasets to signs with at least 6 examples. For assessing the efficacy of sign recognition, we employed an evaluation metric based on the mean Top-K accuracy scores, where K is set to 1 and 5, applied across all instances of the signs.

We have used different combinations of the datasets for different tasks.

- To recognize isolated and pre-segmented sign videos, we combined video clips from all six datasets as follows: the **isolated** sign collections (WLASL (19,666 video clips), RIT (12,197 video clips), ASLLVD (9,746 video clips), DSP (2,935 video clips)); and the **pre-segmented** sign collections: ASLLRP (17,222 video clips) and DSP Sentences (hereafter referred to as DSP_S, 3,136 video clips); totaling 64,902 video clips. After imposing a requirement of at least 6 available example video clips per sign, we arrived at a total of 56,681 distinct video clips corresponding to 2,377 distinct signs.
- To recognize isolated sign videos, the four isolated sign datasets just listed were used, with a total of 44,544 video clips. With the same restriction on example count, this yielded 41,597 distinct video clips corresponding to 2,295 distinct signs. We use the whole video clip, without estimating the beginning and the end frames of the sign.

- To train for recognition of the start and end frames of isolated signs, we merged the two isolated datasets for which we had ground truth annotations for the start and end frames of signs—ASLLVD and DSP—with a total of 12,681 distinct video clips corresponding to 748 distinct signs.

The process of graph construction begins with the normalization of keypoint coordinates within the range of [-1,1]. We then apply a variety of data augmentation techniques, including random sampling, mirroring, rotation, scaling, and shifting. Considering the varying lengths of the videos, we standardize all videos to a uniform length of 200 frames. For videos exceeding this frame count, only the initial 200 frames are used. This truncation does not result in any significant loss of information because of the nature and length of the signs in our datasets. Conversely, for videos shorter than 200 frames, we pad zeros to the end of the temporal dimension to fill up to 200 frames.

5.2. Training Details

We employ Pytorch version 1.7.0 alongside a NVIDIA Quadro RTX8000 graphics card for all computational operations. The Graph Convolutional Network (GCN) models, encompassing both forward and backward streams, are trained under specific parameter settings. The training uses the Cross-Entropy loss function, with a finely-tuned weight decay parameter set to 1×10^{-4}. For optimization, Stochastic Gradient Descent (SGD) with Nesterov Momentum was the chosen method, where the momentum is maintained at 0.9. We initiated the learning rate at 0.1, reducing it by a factor of 10 at the 100th and 150th epoch milestones, culminating the training at 200 epochs.

With respect to batch processing, the batch size is uniformly set at 64 across both the training and testing stages. Each training iteration involves the random selection of 64 videos as inputs, ensuring a varied and comprehensive exposure of the dataset in each epoch. This strategy is pivotal in incorporating every video in the dataset into the training process, thus enhancing the robustness and diversity of the model training.

5.3. Results

5.3.1. GCN Performance

The sign recognition accuracy achieved using the combination of methods described in this paper is presented in Tables 1, 2, and 3.

5.3.2. Improvements in Performance Resulting from Use of Gating & Fusion

The score fusion of the forward and backward streams enhances overall sign recognition, as does the use of Gating for isolated sign video clips.

	WLASL	ASLLVD	RIT	DSP	**Comb.**
Top-1	79.59%	85.53%	75.98%	80.73%	**79.98%**
Top-5	95.32%	96.57%	93.22%	95.70%	**95.04%**

Table 1: Recognition accuracy for isolated signs trained on the combined isolated sign collections

	WLASL	ASLLVD	RIT	DSP	**Comb.**
Top-1	81.32%	86.70%	75.31%	79.97%	**80.76%**
Top-5	95.41%	96.95%	93.38%	95.28%	**95.18%**

Table 2: Recognition of isolated signs trained on the combined isolated & pre-segmented datasets

	ASLLRP	DSP_S	**Comb.**
Top-1	81.58%	73.86%	**80.39%**
Top-5	93.39%	90.62%	**92.96%**

Table 3: Recognition of pre-segmented signs trained on the combined isolated & pre-segmented datasets

This is shown in Table 4. The Bidirectional model's Top-1 and Top-5 performance using forward and backward streams of joints and bones is presented in that table. The first four columns show recognition of isolated signs—based on training on the combined isolated sign collections—with and without Gating. The last two columns show results for recognition of signs from (and trained on) the combined isolated and pre-segmented datasets. It should be noted that the Gating module is not needed for our pre-segmented sign videos, since the start and end frames of these videos had been determined based on linguistic annotations of the start and end points of these signs.

5.3.3. Temporal Action Localization

In this section, we report (1) the accuracy of identification of the start and end frames of signs in isolated video clips, and then (2) the resulting improvement in sign recognition accuracy.

1. Accuracy of Temporal Action Localization

To validate the accuracy of detection of start and end frames, we use the ASLLVD and DSP datasets—for which we have linguistic annotations of the start and end frames for signs. Table 5 presents the Mean Absolute Deviation (MAD), computed separately for the start and end frames as follows:

$$\text{MAD}_{\text{start}} = \frac{1}{N} \sum_{i=1}^{N} |p_{s_i} - g_{s_i}| \tag{2}$$

$$\text{MAD}_{\text{end}} = \frac{1}{N} \sum_{i=1}^{N} |p_{e_i} - g_{e_i}| \tag{3}$$

where, p_{s_i} and p_{e_i} are the predicted start and end frames for the i-th segment, while g_{s_i} and g_{e_i} are

	Isolated (no Gating)		Isolated (with Gating)		Isolated and Pre-segmented	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Forward stream of joints	74.04%	93.06%	74.82%	93.26%	75.59%	92.15%
Forward stream of bones	74.17%	92.63%	75.33%	93.12%	75.07%	92.52%
Backward stream of joints	73.36%	91.82%	73.96%	91.81%	74.02%	91.40%
Backward stream of bones	72.52%	92.44%	75.09%	92.71%	75.49%	92.28%
Fusion	79.24%	94.89%	79.98%	95.04%	80.61%	94.96%

Table 4: Recognition performance for forward and backward streams, where the isolated signs shown in the first 4 columns had been trained on the combined isolated sign data, and the combined isolated and pre-segmented signs in the final 2 columns had been trained on that total dataset

the annotated start and end frames for the i-th examples, and N is the total number of examples.

This is a measure of the deviation between the annotations and predictions for start and end frames of signs in videos with a frame rate of about 30 fps. However, it should be noted that in some cases, there is minimal difference in the images of the annotated and predicted frames; and in some other cases, the prediction may actually be more accurate than the annotation.

	start frame	end frame
ASLLVD	3.03	3.00
DSP	3.93	5.33
Comb.	3.24	3.56

Table 5: Mean Absolute Deviation between annotated and predicted start and end frames

2. Resulting Improvement in Sign Recognition

When our auxiliary multimodal branch network is used to segment signs in our isolated sign datasets, this results in some improvement in sign recognition rates. All video clips were subjected to segmentation processing prior to being input into the GCN model. Table 6 presents the recognition results for the isolated sign datasets, trained on the combined isolated sign datasets, by the GCN model WITH (row [2]) and WITHOUT (row [1]) prior segmentation.

		WLASL	ASLLVD	RIT	DSP	Comb.
[1]	*Top-1*	79.41%	85.35%	75.72%	80.62%	**79.78%**
	Top-5	95.15%	96.53%	93.11%	95.58%	**94.92%**
[2]	*Top-1*	79.59%	85.53%	75.98%	80.73%	**79.98%**
	Top-5	95.32%	96.57%	93.22%	95.70%	**95.04%**

Table 6: Sign recognition accuracy from isolated sign video clips: rows in [1] WITHOUT – and rows in [2] WITH – prior segmentation based on detected sign start and end frames

Although sign segmentation results directly in only a very slight improvement, there are additional ways in which we plan to leverage the ability to identify the start and end frames of lexical signs, specifically with respect to explicit detection of handshapes. As demonstrated by Dilsizian et al. (2014), e.g., it is possible to exploit the linguistic dependencies that hold between start and end handshapes and between the handshapes on the two hands of lexical signs, to improve handshape recognition, which is an important component of sign recognition. They showed that incorporation of statistical information about such handshape dependencies, which can be derived from our annotated corpora, results in significant improvements in isolated sign recognition for lexical signs. This is planned for future research.

5.4. Comparisons of Overall Isolated Sign Recognition Accuracy

Table 7 compares the accuracy of our proposed model against state-of-the-art methods for recognition of signs from the WLASL dataset (Li et al., 2020). The overview at the top is taken from Xiao et al. (2023), Table 2 "Recognition performance comparison for different learning methods in WLASL dataset;" it shows results from [1] (Vinyals et al., 2016); [2] (Snell et al., 2017); [3] (Sung et al., 2018); [4] (Ravi and Larochelle, 2016); [5] (Mishra et al., 2017); [6] (Finn et al., 2017); [7] (Cai et al., 2018); [8] (Gidaris and Komodakis, 2018); [9] (Gordon et al., 2018); [10] (Qiao et al., 2018); [11] (Gidaris and Komodakis, 2019); [12] (Garcia and Bruna, 2017); [13] (Li et al., 2019b); [14] (Liu et al., 2018); and their own [15] (Xiao et al.). These studies used the WLASL dataset, which contains 21,083 video clips with about about 2,000 ASL signs.

As shown at the bottom of the table, our model secured the highest recognition rates for both Top-1 and Top-5. However, it should be noted that Dafnis et al. (2022b) and our own research used a partial but substantial subset of the WLASL data, consisting of 19,672 video examples, reglossed to ensure consistency of labeling (both internal to the WLASL dataset and across our other datasets (Neidle et al., 2022a; Neidle and Ballard, 2022)).

OVERVIEW from Xiao et al. (2023)

Method	Top-1	Top-5
Metric-based		
Matching Nets [1]	41.22%	50.26%
Prototypical Nets [2]	47.61%	65.13%
Relation Net [3]	45.26%	63.21%
Meta-based		
MetaLSTM [4]	41.56%	60.38%
SNAIL [5]	42.18%	53.77%
MAML [6]	46.21%	59.15%
MMNet [7]	52.13%	65.06%
Dynamic-Net [8]	54.21%	70.21%
Generation-based		
VERSA [9]	49.11%	61.19%
Param Predict [10]	55.36%	73.28%
wDAE [11]	55.05%	70.12%
Graph-based		
GNN [12]	52.02%	63.89%
CovaMNet [13]	51.18%	66.39%
TPN [14]	52.15%	65.22%
SL-GCN [15]	56.15%	73.26%
COMPARE WITH		
Dafnis et al. 2022b	77.43%	94.54%
Ours	79.59%	95.32%

Table 7: Performance on the WLASL dataset (which contains isolated signs)

Sandoval-Castañeda et al. (2023) also used this subset of the WLASL dataset, with the same revised glosses. Using a very different approach (summarized in Section 3), they obtained similar results, with 79.02 % Top-1 recognition accuracy; Top-5 accuracy was not reported.

Table 8 compares performance of our model, with training on our isolated sign collection, and that of Dafnis et al. (2022b) on the same combined WLASL and ASLLVD dataset. We attained an improvement of 2.86% in Top-1 accuracy.

Combined	WLASL & ASLLVD	
	Top-1	Top-5
Dafnis et al. 2022b	78.70%	94.79%
Ours	81.56%	95.73%

Table 8: Performance on the same combined WLASL & ASLLVD datasets

6. Conclusions

We introduce here a comprehensive framework for recognition of individual ASL signs. Although most prior related research has focused on isolated, citation-form signs, we successfully extend our recognition to include signs pre-segmented from continuous signing. Our method relies on spatio-temporal GCNs, enhanced by bidirectional stream processing, and, for isolated signs, introduction of a Gating module and an auxiliary multimodal branch for temporal action localization. Our methodology addresses many of the inherent challenges of sign language recognition.

The application of our framework to an extensive collection of different datasets results in a high degree of recognition accuracy. For present purposes, we have used only a limited set of information from facial expressions (i.e., skeleton keypoints), to establish a baseline. In future work we will explore adding more complete information from facial expressions, as this has been shown to improve sign recognition accuracy (von Agris et al., 2008).

We achieve state-of-the-art performance across various metrics, with overall sign recognition accuracy of 80.8% Top-1 and 95.2% Top-5 for citation-form signs, and 80.4% Top-1 and 93.0% Top-5 for signs pre-segmented from continuous signing, when using the combined isolated and pre-segmented sign datasets for training.

Performance enhancements are achieved through use of a bidirectional approach to harness the full temporal context of sign videos; and, for isolated sign clips, of both a Gating module, to filter out non-informative frames and an auxiliary multimodal branch for temporal action localization, to identify the start and end frames of signs. Temporal action localization is a critical step towards ASL recognition from fluent signing.

7. Acknowledgments

We are grateful to the many people who have helped with the collection, linguistic annotation, and sharing of the ASL data upon which we have relied for this research. In particular, we are endebted to the many ASL signers who have contributed to our database; to Gregory Dimitriadis at the Rutgers Laboratory for Computer Science Research, the principal developer of SignStream®, our software for linguistic annotation of video data (`https://www.bu.edu/asllrp/SignStream/3/`); to Matt Huenerfauth and his team for data collection at RIT; to DawnSignPress for sharing video data; to the many who have helped with linguistic annotations (especially Carey Ballard and Indya Oliver); and to Augustine Opoku, for development and maintenance of our Web-based database system for providing access to the linguistically annotated video data (`https://dai.cs.rutgers.edu/dai/s/dai`). This work was supported in part by NSF grants #2235405, #2212302, #2212301, and #2212303, but any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. Bibliographical References

Purva C Badhe and Vaishali Kulkarni. 2015. Indian sign language translator using gesture recognition algorithm. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pages 195–200. IEEE.

Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu. 2021. JOLO-GCN: Mining joint-centered light-weight information for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2735–2744.

Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. 2018. Memory matching networks for one-shot image recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4080–4088.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

HM Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13:2205–2231.

Konstantinos M Dafnis, Evgenia Chroni, Carol Neidle, and Dimitris Metaxas. 2022a. Bidirectional skeleton-based isolated sign recognition using graph convolution networks and transfer learning. In *13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 7328–7338, Marseille, France. European Language Resources Association (ELRA).

Konstantinos M Dafnis, Evgenia Chroni, Carol Neidle, and Dimitris Metaxas. 2022b. Isolated sign recognition using ASL datasets with consistent text-based gloss labeling and curriculum learning. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 13–20, Marseille, France. European Language Resources Association (ELRA).

Nasser H Dardas and Nicolas D Georganas. 2011. Real-time Hand Gesture Detection and Recognition using Bag-of-Features and Support Vector Machine Techniques. *IEEE Transactions on Instrumentation and measurement*, 60(11):3592–3607.

Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. 2019. Spatial-temporal graph convolutional networks for sign language recognition. In *International Conference on Artificial Neural Networks*, pages 646–657. Springer.

Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris Metaxas. 2014. A new framework for sign language recognition based on 3D handshape identification and linguistic modeling. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1924–1929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1126–1135.

Victor Garcia and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.

Spyros Gidaris and Nikos Komodakis. 2018. Dynamic few-shot visual learning without forgetting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4367–4375.

Spyros Gidaris and Nikos Komodakis. 2019. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–30.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. 2018. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*.

Kirsti Grobel and Marcell Assan. 1997. Isolated sign language recognition using Hidden Markov Models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 162–167. IEEE.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458.

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

Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo. 2019b. Distribution consistency based covariance metric networks for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8642–8649.

Frank R Lin, John K Niparko, and Luigi Ferrucci. 2011. Hearing loss prevalence in the United States. *Archives of Internal Medicine*, 171(20):1851–1853.

Hong Liu, Juanhui Tu, and Mengyuan Liu. 2017. Two-stream 3D Convolutional Neural Network for skeleton-based action recognition. *arXiv preprint arXiv:1705.08106*.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 2018. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*.

Dennis Looney and Natalia Lusin. 2021. Enrollments in Languages other than English in United States Institutions of Higher Education, Summer 2016 and Fall 2016 report. In *Modern Language Association*.

Abbas Memiş and Songül Albayrak. 2013. A Kinect Based Sign Language Recognition System using Spatio-temporal Features. In *Sixth International Conference on Machine Vision (ICMV 2013)*, volume 9067, pages 179–183. SPIE.

Dimitris Metaxas, Mark Dilsizian, and Carol Neidle. 2018. Linguistically-driven framework for computationally efficient and scalable sign recognition. In *11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1711–1718, Miyazaki, Japan. European Language Resources Association (ELRA).

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*.

Ross E Mitchell, Travas A Young, Bellamie Bachelda, and Michael A Karchmer. 2006. How many people use ASL in the United States? why estimates need updating. *Sign Language Studies*, 6(3):306–335.

Carol Neidle. 2023. Challenges for Linguistically-driven Computer-based Sign Recognition from Continuous Signing for American Sign Language. *arXiv preprint arXiv:2311.00762*, pages 1–32.

Carol Neidle and Carey Ballard. 2022. Why alternative gloss labels will increase the value of the wlasl dataset. Report no. 21, American Sign Language Linguistic Research Project.

Carol Neidle and Dimitris Metaxas. 2023a. ASLLRP Continuous Signing Corpora, version 1. `https://dai.cs.rutgers.edu/dai/s/signbank`. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.

Carol Neidle and Dimitris Metaxas. 2023b. Boston University American Sign Language Lexicon Video Dataset (ASLLVD), version 7. `https://dai.cs.rutgers.edu/dai/s/signbank`. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.

Carol Neidle and Dimitris Metaxas. 2023c. DawnSignPress (DSP) Collection, version 1. `https://dai.cs.rutgers.edu/dai/s/signbank`. American Sign Language Linguistic Research

Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.

Carol Neidle and Dimitris Metaxas. 2023d. DawnSignPress (DSP) Sentences Collection, version 2. https://dai.cs.rutgers.edu/dai/s/signbank. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.

Carol Neidle and Dimitris Metaxas. 2023e. Rochester Institute of Technology (RIT) Collection, version 4. https://dai.cs.rutgers.edu/dai/s/signbank. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.

Carol Neidle, Augustine Opoku, Carey M Ballard, Konstantinos M Dafnis, Evgenia Chroni, and Dimitris Metaxas. 2022a. Resources for computer-based sign recognition from video, and the criticality of consistency of gloss labeling across multiple large ASL video corpora. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 165–172, Marseille, France. European Language Resources Association (ELRA).

Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022b. ASL video corpora & Sign Bank: Resources available through the American Sign Language Linguistic Research Project (ASLLRP). *arXiv preprint arXiv:2201.07899*, pages 1–20.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan Yuille. 2018. Few-shot image recognition by predicting parameters from activations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7229–7238.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. ZS-SLR: Zero-Shot Sign Language Recognition from RGB-D Videos. *arXiv preprint arXiv:2108.10059*.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *International conference on learning representations*.

Marcelo Sandoval-Castañeda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. Self-supervised video transformers for isolated sign language recognition. *arXiv preprint arXiv:2309.02450*.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in American Sign Language. *Journal of Deaf Studies and Deaf Education*, 26(2):263–277.

Bowen Shi, Diane Brentari, Greg Shakhnarovic, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.1287*.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019a. Skeleton-based action recognition with directed graph neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7904–7913.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019b. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVPR Conference on Computer Vision and Pattern Recognition*, pages 12026–12035.

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.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tae Soo Kim and Austin Reiter. 2017. Interpretable 3D human action analysis with temporal convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 20–28.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

Sandrine Tornay, Oya Aran, and Mathew Magimai Doss. 2020. An HMM approach with inherent model selection for sign language and gesture recognition. In *12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 6049–6056, Marseille, France. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. 29.

Ulrich von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2022. Masked feature prediction for self-supervised visual pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14648–14658.

Qinkun Xiao, Lu Li, and Yilin Zhu. Skeleton-based few-shot sign language recognition. *Available at SSRN 4334054*.

Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. 2015. Joint action recognition and pose estimation from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Quan Yang. 2010. Chinese sign language recognition based on video sequence appearance modeling. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1537–1542. IEEE.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818.