CoSign: Exploring Co-occurrence Signals in Skeleton-based Continuous Sign Language Recognition

Peiqi Jiao[1,2], Yuecong Min[1,2], Yanan Li[3], Xiaotao Wang[3], Lei Lei[3], Xilin Chen[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Xiaomi Inc., China
{peiqi.jiao,yuecong.min}@vipl.ict.ac.cn
{liyanan3, wangxiaotao, leilei1}@xiaomi.com, xlchen@ict.ac.cn

Abstract

*The co-occurrence signals (e.g., hand shape, facial expression, and lip pattern) play a critical role in Continuous Sign Language Recognition (CSLR). Compared to RGB data, skeleton data provide a more efficient and concise option, and lay a good foundation for the co-occurrence exploration in CSLR. However, skeleton data are often used as a tool to assist visual grounding and have not attracted sufficient attention. In this paper, we propose a simple yet effective GCN-based approach, named CoSign, to incorporate **Co**-occurrence **Sign**als and explore the potential of skeleton data in CSLR. Specifically, we propose a group-specific GCN to better exploit the knowledge of each signal and a complementary regularization to prevent complex co-adaptation across signals. Furthermore, we propose a two-stream framework that gradually fuses both static and dynamic information in skeleton data. Experimental results on three public CSLR datasets (PHOENIX14, PHOENIX14-T and CSL-Daily) show that the proposed CoSign achieves competitive performance with recent video-based approaches while reducing the computation cost during training.*

1. Introduction

Sign languages, as the primary means of communication within the Deaf community, are naturally evolved and diversely structured systems in a rule-governed way [2]. Due to their unique physical transmission system, the grammar and vocabulary of sign languages differ greatly from that of spoken languages. To provide a convenient channel between the Deaf and hearing people, vision-based Sign Language Recognition (SLR) has attracted much attention [23] and recent works can be roughly divided into Isolated Sign Language Recognition (ISLR) [28, 18] and Continuous Sign Language Recognition (CSLR) [1, 30].

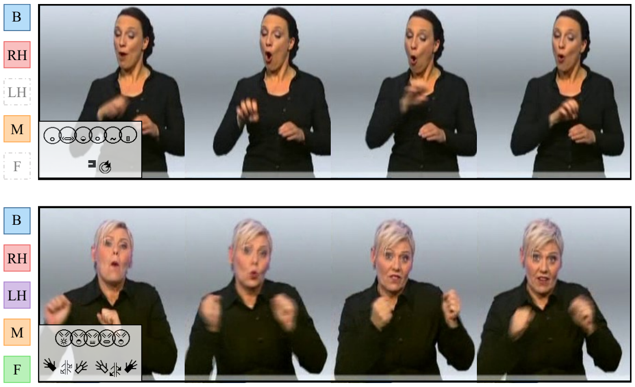


Figure 1. Two examples of signs, EUROPA (Europe, the upper) and UNWETTER (storm, the lower), from PHOENIX14 dataset. SignWriting [43] entries (lexical notations) are positioned at the bottom left-hand corner of images and the occurred signals (B, RH, LH, M and F represent body, right hand, left hand, mouth and face, respectively) are marked on the left side of the images. EUROPA is mainly signed with mouth, right hand and body, and nearly all signals occurred in UNWETTER.

Different to ISLR which predicts the corresponding gloss from a segmented sign video, CSLR aims to recognize a sequence of glosses from a continuous image sequence and is more common in real-life applications. Video-based CSLR develops rapidly in the last few years [13, 35, 9, 32, 5, 17, 30, 6]. However, video-based approaches are sensitive to background and illumination changes [21] and may introduce privacy concerns [3]. Meanwhile, the computation cost is a considerable problem because sign videos contain much visual redundancy and recent CSLR approaches usually extract visual features in a frame-by-frame way. Compared with videos, skeleton data provide a more concise and efficient representation for human body.

Recent single person pose estimation solutions [42, 16] have achieved high performance in complex scenes and are widely used in action recognition [48, 15]. Different to action recognition, CSLR focuses more on fine-grained information. However, recent skeleton-based methods [8, 33] achieve inferior performance compared with video-based approaches in CSLR, and some works [50, 8] leverage skeleton data to assist the learning of video data. These findings raise an interesting question: *what is the obstacle that prevents the utilization of skeleton data in CSLR?*

As shown in Fig. 1, sign language conveys information through both manual signals (hand shape, orientation, place of articulation and movement) and non-manual signals (lip pattern, facial expression, head and upper body orientation) [14, 24]. For skeleton data, these synchronous signals can be easily modeled by the interactions among different groups of keypoints, but directly treating all keypoints as a whole may prevent the model from learning these co-occurrence signals. Therefore, we argue that efficiently modeling these co-occurrence signals is the key to boost the performance of skeleton-based CSLR method.

To incorporate co-occurrence signals, we propose a simple yet effective Graph Convolution Network (GCN) based approach, named CoSign, which consists of a group-specific GCN and a complementary regularization. The group-specific GCN contains several customized modules to independently process different signals. Specifically, the estimated skeleton data are first divided into five groups: body, left hand, right hand, mouth, and face, and then the keypoints of each group are sent to the corresponding group-specific module to process each signal. Compared to directly treating the human skeleton as a whole, this design can better exploit the knowledge of each signal.

Due to the fast movement and heavy self-occlusion of hands, off-the-shelf estimators often miss or predict inaccurate keypoints, which may affect the accuracy of CSLR models. The proposed complementary regularization encourages the consistency between predictions based on two complementary subsets of signals, which can make better use of signals with different intensities and relieve the effects of inaccurate estimations. Moreover, we design a two-stream framework to explicitly capture static and dynamic information from skeleton data and gradually fuse them through an extra fusion branch.

In conclusion, this paper focuses on the utilization of skeleton data in CSLR. We propose the CoSign to exploit the co-occurrence signals in sign language. Experimental results on several popular CSLR datasets show that the proposed CoSign can achieve comparable results with video-based approaches while reducing the training cost.

The main contributions are summarized as follows:

- Exploring the potential of skeleton data in CSLR, and attributing the key to utilizing co-occurrence signals.
- Proposing a group-specific GCN to exploit the knowledge of each signal in sign language independently.
- Proposing a complementary regularization to handle noisy skeleton input and co-adaptation across signals.
- Designing a two-stream framework to gradually fuse static and dynamic information in skeleton data.

2. Related Work

2.1. Continuous Sign Language Recognition

A general CSLR model can be roughly divided into two components: feature extractor and alignment module. For feature extractor, recent CSLR approaches usually adopt Convolutional Neural Networks (CNNs), such as 2D CNN [32], 2D CNN+1D CNN [13, 9, 17, 30] or 3D CNN [35, 6]. As CNNs only have local receptive field, some works append a contextual module like Recurrent Neural Network (RNN) [30, 17] or Transformer [32, 5]. The alignment module is employed to align frame-wise features with gloss sequence and provide supervision during training. Both Hidden Markov Models (HMMs) [26, 24] and Connectionist Temporal Classification (CTC) [13, 9] are explored to achieve this goal, and CTC becomes the mainstream method due to its simplicity. However, some works [35, 13] discover that CTC suffers from insufficient training, which largely harms the performance. The iterative training scheme [13] is proposed to relieve this problem, but it also increases the training time and complexity. After that, some works [30, 17, 52] find that increasing the consistency between the feature extractor and contextual module during training is a better solution: it can achieve competitive performance while reducing training cost.

Sign language conveys information through both manual and non-manual signals simultaneously. To exploit implicit collaboration of multiple visual signals, recent works [50, 52, 33, 8] explore the use of skeleton data in CSLR. Some works utilize skeleton data to guide the video feature extraction through an attention module [52] or an extra pose estimation branch [50]. Another series of works [8, 33] directly extract features from skeleton data and fuse them with visual features considering their complementary nature. However, there often exists a large performance gap between video-based and skeleton-based models, which is attributed to the inaccurate keypoints estimator [8]. In this paper, we explore the potential of skeleton data in CSLR and show that, with an off-the-shelf estimator, skeleton-based method could achieve competitive performance with video-based methods.

2.2. Skeleton-based Action Recognition

Recent skeleton-based action recognition approaches often adopt CNN-based or GCN-based architectures to exploit the topological structure of the skeleton. CNN-based

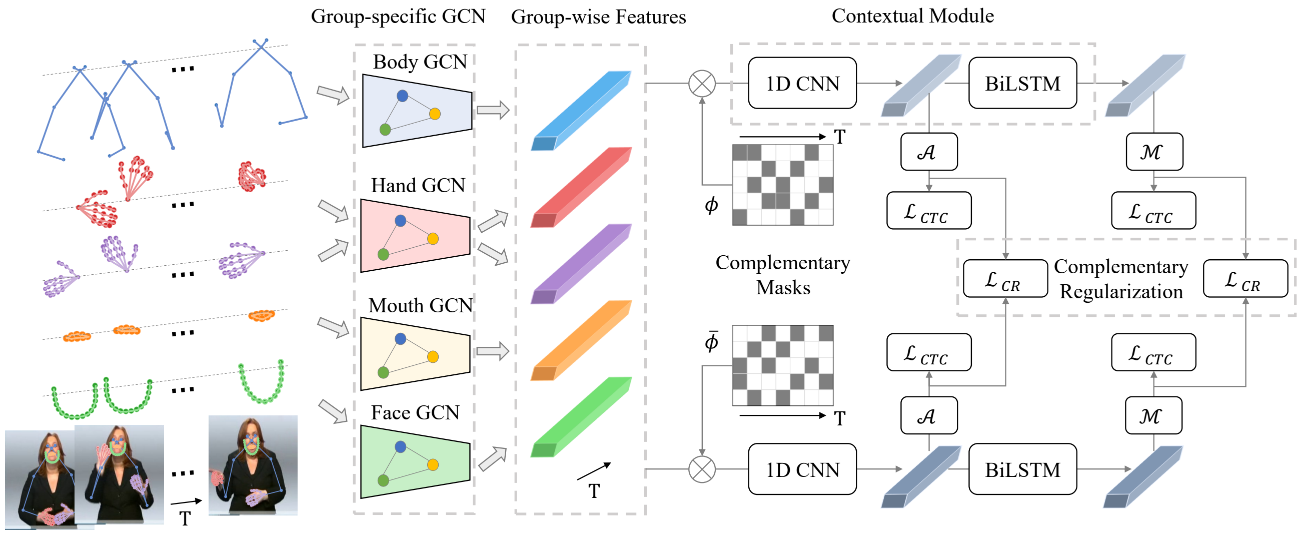


Figure 2. The framework of CoSign-1s, which contains a group-specific GCN, a contextual module and complementary regularization. The input skeleton data are first divided into five groups and the group-specific centralization is applied. A group-specific GCN consisting of four GCN modules is used to extract group-wise features from them. Two complementary masks are applied on the features before sending to the contextual module and the whole network is jointly supervised by CTC losses $\mathcal{L}_{CTC}$ and complementary regularization $\mathcal{L}_{CR}$.

methods usually need some special designs like representing the skeleton through stacking heatmaps of joints [15] or enhancing the topology feature by a cross-channel feature augmentation module [47]. GCN is naturally more suitable for skeleton data than CNN due to its graph structure. The pioneering work [48] proposes a Spatial Temporal Graph Convolutional Networks (ST-GCN) that directly aggregate keypoint information through fixed adjacency matrices. However, as the pre-defined graphs may not be optimal for specific classes, some works [29, 36, 7, 10] propose several kinds of adaptive graphs learned directly from data. Moreover, several works explore the role of bone and motion information in skeleton data by multi-stream GCNs [27, 37] with a simple late fusion mechanism.

Another attractive feature in skeleton-based action recognition is the co-occurrence of joints in different actions, which is also known as “actionlet” [45]. For example, the actionlet of “eat meal” includes joints of the head, hand, and elbow. Many efforts persist to explore the co-occurrence of joints including designing more suitable spatial representation [22, 27], manually dividing human body into several parts [44, 40, 38] and designing proper mechanisms to adaptively focus on joints [51, 39, 20].

The co-occurrence of joints is more critical in CSLR due to the synchronous signals in sign language. Some CSLR works [11, 24, 50] explore the co-occurrence in RGB videos. Koller *et al.* [24] leverage multiple HMM streams to synchronize signals from the pose, mouth shape, and hand shape. Zhou *et al.* [50] adopt the estimated pose to guide the learning of different visual signals. Different from these works, we explore the co-occurrence characteristic of CSLR in skeleton data, which is a more suitable basis than videos due to its structural nature. Meanwhile, different from works in skeleton-based action recognition, we model each signal with respect to its own sequential nature.

3. Method

In this section, we first introduce the proposed single stream approach (CoSign-1s) to explore the co-occurrence signals within skeleton data in Sect. 3.1, which includes a group-specific GCN and a complementary regularization. Then we further design a two-stream framework named CoSign-2s to explore the potential of fusing skeleton and motion sequence in Sect. 3.2.

Background. The skeleton-based CSLR model aims to learn the monotonous alignment between the skeleton sequence $\mathbf{J} = \{\mathbf{J}_1, \cdots, \mathbf{J}_T\}$ and the corresponding gloss sequence $\mathbf{l} = \{l_1, \cdots, l_N\}$. Each skeleton frame contains K keypoints $\mathbf{J}_i = \{\mathbf{J}_{i,k} \in \mathbb{R}^2 | k = 1, \cdots, K\}$. Similar to general video-based CSLR framework [30], we design the group-specific GCN to extract the frame-wise features of each signal. Then a 1D CNN layer is followed to capture gloss-wise features by aggregating local features. We adopt a two-layer BiLSTM to learn long-term dependencies and CTC loss is utilized to provide supervision for the prediction $\mathbf{y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_{T'}\}$ through dynamic programming:

$$\begin{aligned} \mathcal{L}_{CTC}(\mathbf{y}) &= -\log p(\mathbf{l}|\mathbf{y}) \\ &= -\log_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi|\mathbf{y}), \end{aligned} \tag{1}$$

where $\mathcal{B}$ is a many-to-one mapping between predictions and labels, and π is a feasible alignment path.

3.1. Single Stream CoSign

The whole structure of our CoSign-1s is shown in Fig. 2, which consists of a group-specific GCN, a contextual module and a complementary regularization.

Group-specific GCN. Most recent CSLR datasets only provide video data. To obtain skeleton data, we utilize an off-the-shelf estimator (MMPose [12]) to obtain whole body keypoints from each frame of the sign language videos. For balancing the efficiency and exploiting group-specific knowledge, we empirically select 77 keypoints and divide them into five groups as shown in Fig. 2: 9 for body, 21 for each hand, 8 for mouth and 18 for face (denoted as $\mathcal{G}_B$, $\mathcal{G}_{LH}$, $\mathcal{G}_{RH}$, $\mathcal{G}_M$ and $\mathcal{G}_F$ respectively). Then group-specific centralization is applied to further decouple multi-grained motion information in skeleton data, which is implemented by aligning the root keypoints of each group across time:

$$\mathbf{J}_{t,k} = \mathbf{J}_{t,k} - \mathbf{J}_{t,r(g)}, k \in \mathcal{G}_g, \tag{2}$$

where $r(g)$ denotes the index of root keypoint of group g.

To exploit group-specific knowledge from keypoint groups, we design four GCN modules, where left and right hands share the same one. ST-GCN [48] is chosen as the basic building unit of each module due to its powerful modeling ability and lightweight structure. The basic ST-GCN layer consists of a spatial graph convolution and a temporal convolution. Given the keypoints of group g, the spatial features are aggregated through a group-specific adjacency matrix $\mathbf{A}_g$, which is constructed by connecting spatially adjacent keypoints of group g based on human anatomy. The spatial graph convolution operation on keypoints of group g can be formulated as:

$$\mathbf{f}_{out}(t,g) = \sum_k \mathbf{\Lambda}_{g,k}^{-\frac{1}{2}} \mathbf{A}_{g,k} \mathbf{\Lambda}_{g,k}^{-\frac{1}{2}} \mathbf{f}_{in}(t,g) \mathbf{W}_k, \tag{3}$$

where $\mathbf{f}_{in}(t,g)$ denotes the input feature vector of all keypoints of group g at timestep t. $\Lambda_{g,k}^{ii} = \sum_j A_{g,k}^{ij} + \epsilon$ is the normalized diagonal matrix and $\epsilon = 0.001$ to avoid empty rows in $\mathbf{A}_{g,k}$. We adopt the distance partition strategy ($k = 2, \mathbf{A}_0 = \mathbf{I}, \mathbf{A}_1 = \mathbf{A}$) and $\mathbf{W}_k$ is the weight matrix of the partition k.

ST-GCN layer gathers features of the same keypoint across frames within the temporal window Γ through a standard 2D convolution with a kernel size of $1 \times \Gamma$, which can help refine inaccurate estimations. Because the pre-defined spatial graph may not be the optimal one, the adjacency matrix $\mathbf{A}_g$ is parameterized and can be further optimized during training, which helps learn implicit correlations among keypoints within the graph.

The final group-specific GCN consists of a shared linear layer that maps coordinates of keypoints to the feature space and four modules that capture group-specific representation from five keypoint groups. Each module contains three ST-GCN layers. The output features of different groups are fused through a MLP layer to generate frame-wise features. A contextual module that consists of 1D CNN and BiLSTM layers is adopted to incorporate temporal information and make the prediction.

Complementary Regularization. Although modeling each group independently can better exploit group-specific knowledge, it still faces the challenges of reducing impacts from estimation noise and preventing complex co-adaptation across signals. To explore co-occurrence in CSLR, we propose a complementary regularization that encourages the consistency between predictions based on two complementary subsets of signals.

When conveying the same information through different signals, the weak signals may be omitted. Inspired by the dropout method [41], we propose a group dropout mechanism to make better use of signals with different intensities. Concretely, for the outputs $\mathbf{v} \in \mathbb{R}^{T \times N \times C_{in}}$ of group-specific GCN, where N is the number of groups. The corresponding dropout mask $\boldsymbol{\xi} \in \mathbb{R}^{T \times N}$ is segmented into $\lceil T/\tau \rceil \times N$ clips with a pre-defined length τ. Each clip of dropout mask is independently sampled from a Bernoulli distribution $B(p)$. Then the dropout mask $\boldsymbol{\xi}$ is expanded to $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^{T \times N \times C_{in}}$, which has the same dimension as the group-wise feature $\mathbf{v}$. We apply the expanded mask on $\mathbf{v}$ and obtain the frame-wise feature $\tilde{\mathbf{v}} \in \mathbb{R}^{T \times C_{out}}$ through the fusion MLP layer $\mathcal{H}$:

$$\tilde{\mathbf{v}} = \mathcal{H}(\tilde{\boldsymbol{\xi}} \odot \mathbf{v}), \tag{4}$$

where $\odot$ denotes the Hadamard product, and the results of $\tilde{\boldsymbol{\xi}} \odot \mathbf{v}$ is reshaped to (T, NC_{in}).

The combination of group-specific GCN and group dropout provides a simple way to control the participated signals in CSLR. Inspired by the consistency regularization design of R-Drop [46], we further propose a complementary regularization to explore co-occurrence in CSLR. Specifically, we first generate a dropout mask $\boldsymbol{\xi}$ that is sampled from $B(2p)$, and then equally split it into two complementary masks $\boldsymbol{\phi}$ and $\bar{\boldsymbol{\phi}}$. As shown in Fig. 2, the group-wise features are fed to the contextual module twice with complementary masks ($\boldsymbol{\phi}$ and $\bar{\boldsymbol{\phi}}$) and two predictions are obtained from the classifier $\mathcal{C}$ (denoted as $\mathcal{P}_{\boldsymbol{\phi}}^{\mathcal{C}}$ and $\mathcal{P}_{\bar{\boldsymbol{\phi}}}^{\mathcal{C}}$). The complementary regularization is defined as the symmetrical Kullback-Leibler divergence between these predictions:

$$\mathcal{L}_{CR}(\mathcal{P}_{\boldsymbol{\phi}}^{\mathcal{C}}, \mathcal{P}_{\bar{\boldsymbol{\phi}}}^{\mathcal{C}}) = \frac{1}{2}\mathcal{D}_{KL}(\mathcal{P}_{\boldsymbol{\phi}}^{\mathcal{C}} || \mathcal{P}_{\bar{\boldsymbol{\phi}}}^{\mathcal{C}}) + \frac{1}{2}\mathcal{D}_{KL}(\mathcal{P}_{\bar{\boldsymbol{\phi}}}^{\mathcal{C}} || \mathcal{P}_{\boldsymbol{\phi}}^{\mathcal{C}}). \tag{5}$$

The intuition behind the proposed complementary regularization is simple: the use of complementary masks eliminates duplicate subsets and reduces shortcut solutions; and the regularization of Equ. 5 encourages the consistency of predictions from different signals, which can make the model more robust to noise.

Supervision. As shown in previous work [30] that adopting auxiliary loss can relieve overfitting, we attach an auxiliary classifier after the 1D CNN layer to provide supervision for

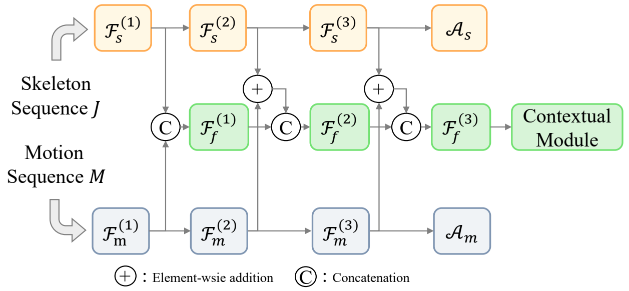


Figure 3. The structure of two-stream CoSign. The superscript (i) of $\mathcal{F}$ means the i-th block in it.

the group-specific GCN. Besides, we normalize the weight matrices of classifiers and the feature vectors, and share the weight matrix of the primary classifier $\mathcal{M}$ with the auxiliary classifier $\mathcal{A}$ as previous works do [17, 31]. The recognition loss $\mathcal{L}_{SLR}$ is composed of two CTC losses that applied on the auxiliary prediction $\mathcal{P}^{\mathcal{A}}$ and final prediction $\mathcal{P}^{\mathcal{M}}$:

$$\mathcal{L}_{SLR-1s}(\phi) = \mathcal{L}_{CTC}(\mathcal{P}^{\mathcal{A}}_{\phi}) + \mathcal{L}_{CTC}(\mathcal{P}^{\mathcal{M}}_{\phi}), \qquad (6)$$

where the superscript of prediction $\mathcal{P}$ denotes the classifier that generates it and the subscript denotes the applied mask.

As we feed $\mathbf{v}$ to the contextual module twice with different masks, two recognition losses are calculated. The final loss of CoSign-1s can be formulated as:

$$\begin{aligned}\mathcal{L} = {} & \frac{1}{2}\mathcal{L}_{SLR-1s}(\phi) + \frac{1}{2}\mathcal{L}_{SLR-1s}(\bar{\phi}) \\ & + \alpha\mathcal{L}_{CR}(\mathcal{P}^{\mathcal{A}}_{\phi}, \mathcal{P}^{\mathcal{A}}_{\bar{\phi}}) + \beta\mathcal{L}_{CR}(\mathcal{P}^{\mathcal{M}}_{\phi}, \mathcal{P}^{\mathcal{M}}_{\bar{\phi}}),\end{aligned} \qquad (7)$$

where α and β denote the loss weights of complementary regularizations on auxiliary and primary predictions.

3.2. Two-stream CoSign

Both static and dynamic information plays a critical role in CSLR. We argue that directly modeling them is more efficient compared with only depending on the temporal convolution layers in group-specific GCN. Thus we obtain the bidirectional movement $\mathbf{M}_{t,k}$ of the keypoint $\mathbf{J}_{t,k}$ by calculating the coordinate differences in two consecutive frames:

$$\mathbf{M}_{t,k} = [\mathbf{J}_{t,k} - \mathbf{J}_{t-1,k}, \mathbf{J}_{t+1,k} - \mathbf{J}_{t,k}], \qquad (8)$$

where $[\cdot, \cdot]$ means concatenation.

To leverage the intermediate presentations of both streams, we propose a two-steam framework that consists of three branches (referred to as skeleton, motion and fusion branches), and each branch contains a group-specific GCN (denoted as $\mathcal{F}_s$, $\mathcal{F}_m$ and $\mathcal{F}_f$ respectively). As shown in Fig. 3, the skeleton and motion branches take skeleton sequence $\mathbf{J}$ and motion sequence $\mathbf{M}$ into account independently, and the fusion branch incorporates the intermediate features of them gradually. Similar to CoSign-1s, the group-wise features from the fusion branch are fused to the frame-wise features and further sent to the contextual module to get the final prediction.

Due to the increased capacity of the CoSign-2s and different convergence rates of branches, we first pre-train the skeleton and motion-based CoSign-1s independently for several epochs with Equ. 6. After that, we load the pre-trained weights of the corresponding branches and start the training of the CoSign-2s.

The supervision of Equ. 6 is applied to the training of CoSign-2s with slight modifications. We attach two auxiliary classifiers on the skeleton and motion branches (denoted as $\mathcal{A}_s$ and $\mathcal{A}_m$) as shown in Fig. 3 and adopt group dropout in each branch. For the dropout mask ϕ, the recognition loss of CoSign-2s is:

$$\begin{aligned}\mathcal{L}_{SLR-2s}(\phi) = {} & \mathcal{L}_{CTC}(\mathcal{P}^{\mathcal{A}_f}_{\phi}) + \mathcal{L}_{CTC}(\mathcal{P}^{\mathcal{M}_f}_{\phi}) \\ & + \lambda\left(\mathcal{L}_{CTC}(\mathcal{P}^{\mathcal{A}_s}_{\phi}) + \mathcal{L}_{CTC}(\mathcal{P}^{\mathcal{A}_m}_{\phi})\right),\end{aligned} \qquad (9)$$

where λ is the loss weight of skeleton and motion branches, $\mathcal{A}_f$ and $\mathcal{M}_f$ represent the auxiliary and primary classifiers of the fusion branch.

Similar to single stream approach, we apply complementary masks ϕ and $\bar{\phi}$ on the group-wise features of all branches, and only apply complementary regularization on fusion branch for simplify. The total loss has a similar format as Equ. 7:

$$\begin{aligned}\mathcal{L} = {} & \frac{1}{2}\mathcal{L}_{SLR-2s}(\phi) + \frac{1}{2}\mathcal{L}_{SLR-2s}(\bar{\phi}) \\ & + \alpha\mathcal{L}_{CR}(\mathcal{P}^{\mathcal{A}_f}_{\phi}, \mathcal{P}^{\mathcal{A}_f}_{\bar{\phi}}) + \beta\mathcal{L}_{CR}(\mathcal{P}^{\mathcal{M}_f}_{\phi}, \mathcal{P}^{\mathcal{M}_f}_{\bar{\phi}}),\end{aligned} \qquad (10)$$

where α and β are the same hyper-parameters as Equ. 7.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the proposed method on three popular CSLR datasets: PHOENIX14 [25], PHOENIX14-T [4] and CSL-Daily [49]. Word Error Rate (WER) is adopted as the evaluation metric for all experiments and a lower WER indicates a more accurate recognition. All ablation studies are conducted on PHOENIX14.

- **PHOENIX14** is a German sign language dataset which has a vocabulary of 1295 glosses and contains 5672, 540, and 629 samples performed by 9 signers for training, dev and test sets, respectively.
- **PHOENIX14-T** is an extension to PHOENIX14 with both gloss and translation annotations. It has a vocabulary of 1085 glosses and is divided into three parts: 7096 samples for training, 519 for development, and 642 for testing.
- **CSL-Daily** is a Chinese sign language dataset, it has a vocabulary of 2000 glosses and the number of samples of training, dev and test sets is 18401, 1077 and 1176.

Baseline. The adopted baseline considers all keypoints as a whole and stacks three ST-GCN [48] layers with the distance partition strategy as the frame-wise feature extractor.

Table 1. Performance comparison (WER, %) on PHOENIX14/14-T and CSL-Daily. We divide these methods into three groups according to their input type. The **best** results in each group are highlighted.

Method	Input Type		PHOENIX14		PHOENIX14-T		CSL-Daily	
	Raw Video	Skeleton	Dev	Test	Dev	Test	Dev	Test
DNF [13]	✓		23.8	24.4	-	-	32.8	32.4
VAC [30]			21.2	22.3	-	-	33.3	32.6
CMA [34]			21.3	21.9	-	-	-	-
TwoStream-SLR [8]			22.4	23.3	21.1	22.4	28.9	28.5
SMKD [17]			20.8	21.0	20.8	22.4	**28.4**	**27.5**
TLP [19]			19.7	20.8	**19.4**	**21.2**	-	-
RadialCTC [31]			**19.4**	**20.2**	-	-	-	-
STMC [50]	✓	✓	21.1	20.7	19.6	21.0	-	-
C2SLR [52]			20.5	20.4	20.2	20.4	-	-
SignBERT+ [18]			19.9	20.0	18.8	19.9	-	-
TwoStream-SLR [8]			**18.4**	**18.8**	**17.7**	**19.3**	**25.4**	**25.3**
SignBERT+ [18]		✓	34.0	34.1	32.9	33.6	-	-
TwoStream-SLR [8]			28.6	28.0	27.1	27.2	34.6	34.1
Baseline			24.3	24.4	23.6	23.8	31.6	31.2
CoSign-1s			20.9	21.2	20.4	20.6	29.5	29.1
CoSign-2s			**19.7**	**20.1**	**19.5**	**20.1**	**28.1**	**27.2**

Table 2. Efficiency comparison on PHOENIX14 without pose estimation stage taking into account. The computational cost is measured in FLOPs (FLoating-point OPerations).

Method	FLOPs	Parameters	Training Time	Training Memory	Training Batch Size	Inference Speed
SMKD [17]	183.2 G	31.6 M	∼19.5 h	21.5 GB	2	11.2 seq/s
Baseline	5.3 G	18.8 M	∼1.0 h	5.4 GB	8	26.1 seq/s
CoSign-1s	5.8 G	21.4 M	∼2.0 h	6.5 GB	8	18.8 seq/s
CoSign-2s	30.1 G	28.2 M	∼4.5 h	18.0 GB	8	12.7 seq/s

The adjacency matrices in all blocks are built based on human anatomy and optimized during training. The 1D CNN layer is made up of C3-P2-C3-P2, where Cθ and Pθ represent the 1D convolutional layer and max pooling layer with a kernel size of θ, respectively. The hidden states of the two-layer BiLSTM are set to 2x512 dimension. The supervision consists of two CTC losses that are applied on auxiliary and primary predictions, respectively.

Implementation Details. We use the 2D coordinates along with a confidence score for each joint as input. For single stream, both α and β in Equ. 7 and Equ. 10 are set to 2 and we train all models for 40 epochs with a mini-batch size of 8. AdamW optimizer is used with an initial learning rate of 4×10^{-4} and divided by 10 after 20 and 35 epochs. For two-stream approach, we pre-train skeleton and motion branches for 10 epochs and train the whole model for 40 epochs. The loss weight λ in Equ. 9 is set to 0.5. Mini-batch size and optimizer keep the same as single stream. For pose estimation, we adopt 256×256 spatial resolution for PHOENIX14 and PHOENIX14-T, and 512×512 for CSL-Daily. We only use random temporal scaling ($\pm 20\%$) for augmentation.

Table 3. Performance comparison (WER, %) under PHOENIX14-SI setting.

Method	Dev	Test
ReSign [26]	45.1	44.1
DNF [13]	36.0	35.7
CMA [34]	34.8	34.3
SMKD [17]	34.5	34.2
VAC [31]	36.7	33.8
RadialCTC [31]	**33.8**	**32.2**
Baseline	40.7	39.0
CoSign-1s	32.1	31.8
CoSign-2s	**31.1**	**29.6**

4.2. Comparison with State-of-the-arts

To show the effectiveness of the proposed method, we first compare CoSign with state-of-the-art methods on three popular CSLR datasets: PHOENIX14, PHOENIX14-T and CSL-Daily. As shown in Table 1, the adopted baseline already achieves superior recognition results than previous skeleton-based work [8], and we attribute it to the skeleton graphs and lightweight architecture, which can clearly distinguish occluded keypoints of different groups and prevent overfitting. Based on this simple yet strong baseline, the proposed CoSign-1s and CoSign-2s can further reduce WER by 3.4%/3.2% and 4.6%/4.3% on the Dev/Test sets of POHENIX14, respectively. Similar results can also be observed on the other two datasets. Although the adopted pose estimator sometimes predicts inaccurate keypoints, the proposed CoSign-2s can achieve competitive performance with

Table 4. Ablation results (WER, %) of group-specific GCN.

Group-specific Module	Centralization	Dev	Test
		24.3	24.4
✓		22.8	22.9
✓	✓	**21.8**	21.9

Table 5. Ablation results (WER, %) of two-stream fusion.

Method	Stream		Dev	Test
	Skeleton	Motion		
CoSign-1s	✓		21.8	21.9
		✓	22.8	23.5
Late Fusion	✓	✓	21.0	21.1
CoSign-2s	✓	✓	**20.7**	20.5

the best video-based methods on all three datasets, which demonstrates the potential of skeleton data in CSLR. However, there still exists a performance gap between CoSign-2s and the sota TwoStream-SLR [8], which also reveals the potential of adopting more powerful pose estimators and more efficient fusion methods for RGB and skeleton data.

Skeleton data are more robust to appearance changes than RGB data. Therefore, we evaluate the robustness of CoSign to signer changes and present experimental results under the PHOENIX14-SI setting [26], which excludes signer 5 from the training set for signer-independent experiments. As shown in Table 3, the proposed CoSign significantly improves the performance compared to baseline and achieves a new state-of-the-art result, which outperforms the previous best method [31] by 2.7%/2.6% and verifies the robustness of CoSign to signer changes.

Moreover, we also report both the training and inference efficiency on a NVIDIA GeForce RTX 3090 GPU with data cached. We do not take pose estimation stage into account because the proposed method only loads the video data and estimates skeleton data once before training. As shown in Table 2, CoSign models are training friendly, which have lower FLOPs (calculated under a sequence of 100 frames), smaller model size and faster training speed (the average sequence per second on PHOENIX14 dev and test sets with a batch size of 1) than SMKD [17]. With the development of pose estimation, CoSign is easy to deploy and compatible with other skeleton-based applications.

4.3. Ablation Study

Group-specific GCN. We first evaluate the effects of group-specific GCN. As shown in Table 4, both group-specific module and centralization can significantly reduce WER. We attribute the effects of group-specific module to the better exploration of the group-specific knowledge in sign language, which will be further discussed. For centralization, it allows a further decomposition of common information among different groups, *e.g.*, the tiny finger motion and the upper body motion, which can make the model focus on the unique signals inside each group.

Table 6. Ablation results (WER, %) of clip length in dropout mask. T represents the clip length is the same as the video length.

Clip Length	Dev	Test
12	21.5	21.7
25	**21.2**	21.4
50	21.3	21.6
100	21.4	21.2
T	21.6	21.6

Table 7. Ablation results (WER, %) of complementary regularization. G-drop, CR and C-mask denote group dropout, complementary regularization, and complement masks, respectively.

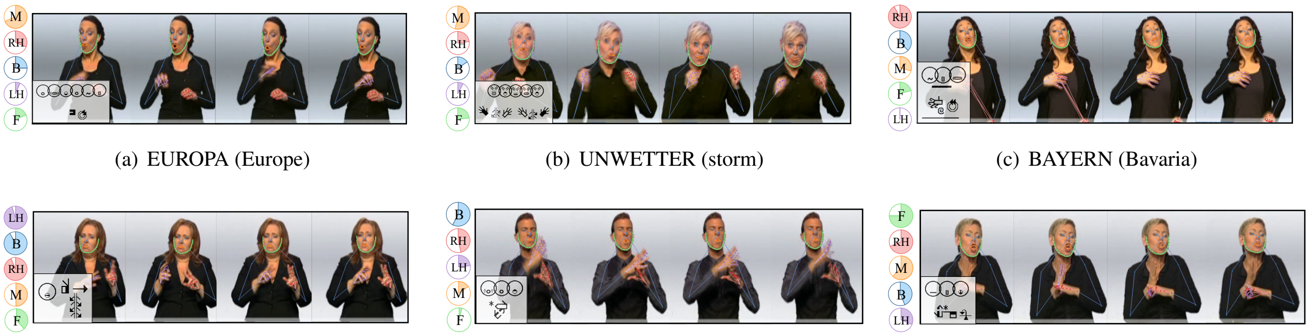
Method	G-drop	CR	C-mask	Dev	Test
CoSign-1s				21.8	21.9
	✓			21.2	21.4
	✓	✓		21.0	21.4
	✓	✓	✓	**20.9**	21.2
CoSign-2s				20.7	20.5
	✓			20.3	20.4
	✓	✓		20.2	20.2
	✓	✓	✓	**19.7**	20.1

Two-stream Fusion. Table 5 presents the evaluation results of different skeleton formats and fusion approaches. Fusing both skeleton formats can bring further improvement, which indicates they are complementary to each other. The fusion approach comparison also reveals that keeping the independence of each branch and gradually fusing features of two branches can better take advantage of them.

Clip Length in Dropout Mask. Because the sequence features are temporally correlated, the effects of proposed $\mathcal{L}_{CR}$ rely on the clip length in dropout mask: adopting short masks may fail to prevent co-adaptation across signals, and adopting long masks will reduce the diversity of signals within the sequence. We evaluate different mask lengths with a fixed probability of 0.2. Experimental results in Table 6 support our assumption and adopting the lengths in dropout mask ranging from 25 to 100 achieves comparable results. We adopt the length of 25 as the default setting because it is slightly longer than the approximate length of a single sign in PHOENIX14.

Complementary Regularization. We evaluate the effectiveness of the complementary regularization and present results in Table 7. Adopting group-wise dropout achieves lower WERs on both single and two-stream CoSign, which indicates the existence of co-adaptation among different signals in CSLR. It is also worth noting that the complementary masks play a critical role in the complementary regularization, because it eliminates the duplicate subsets of signals and improves the efficiency of regularization.

Co-occurrence Signals. To verify the existence of co-occurrence signals in sign language, we evaluate the performance of different *trained* models on a specific group by masking keypoints of other groups and fine-tuning the model with *frozen* feature extractor. The fine-tuning re-



(a) EUROPA (Europe)

(b) UNWETTER (storm)

(c) BAYERN (Bavaria)

(d) WECHESLHAFT (variable)

(e) SCHON (already)

(f) MAL (just)

Figure 4. Visualization of sign examples, the corresponding precision of different signals (*e.g.*, ◔ represents the precision of 25% for this sign) and corresponding SignWritting entries. B, RH, LH, M and F represent body, right hand, left hand, mouth and face, respectively.

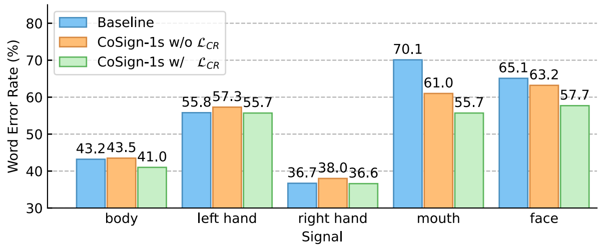
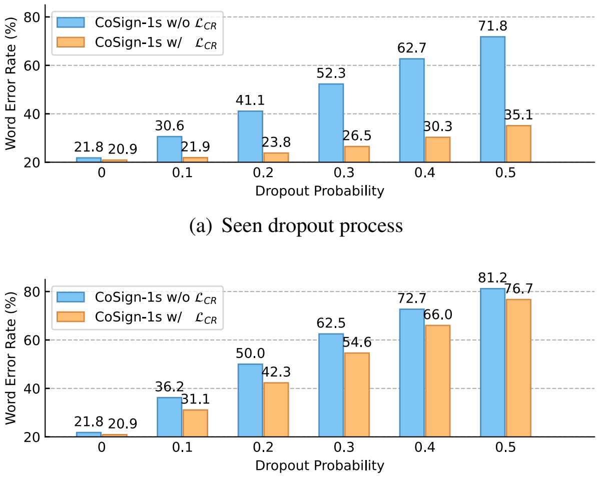


Figure 5. Comparison (WER, %) of different signals with different *finetuned* models.

sults can reflect the representation ability of the feature extractor for a specific signal. Fig. 5 visualizes the comparison results among different settings and we can observe that adopting group-specific GCN achieves lower WERs for some weak signals like mouth and face. With the complementary regularization, signals from almost all groups are better explored, which demonstrates that CoSign can better explore the co-occurrence signals. In Fig. 4, we further visualize the average precision of different signals from CoSign on six different signs from PHOENIX14 dev and test sets. For example, the sign BAYERN in Fig. 4(c) is mainly signed with right hand, mouth and body, which are successfully captured by CoSign. Besides, the diverse contributions from different signals verify the necessity of exploring co-occurrence in CSLR.

Robustness to Noise. The complementary regularization can also reduce impacts from estimation noise. To simulate the effects of estimation noise, we apply groupwise dropout on keypoints (unseen) or intermediate features (seen) with different dropout probabilities and visualize recognition results in Fig. 6. For the seen dropout process in Fig. 6(a), CoSign with $\mathcal{L}_{CR}$ can still achieve acceptable performance even when half of groups are dropped. Besides, CoSign with $\mathcal{L}_{CR}$ can steadily reduce the performance drop at different dropout probabilities of unseen dropout process. However, dropping keypoints still leads to severe performance degeneration and how to handle estimation noise requires further exploration.



(a) Seen dropout process

(b) Unseen dropout process

Figure 6. Comparison (WER, %) with simulated estimation noise on PHOENIX14 Dev set.

5. Conclusion

In this study, we focus on the utilization of skeleton data in CSLR and attribute the key to the utilization of co-occurrence signals. To explore the potential of skeleton data, we employ two techniques: the group-specific GCN aims to exploit the knowledge of each signal independently and the complementary regularization handles the co-adaptation across signals and noisy skeleton input. In addition, we design a two-stream framework to fuse static and dynamic information from both skeleton and motion sequence. Experimental results show that our CoSign can achieve a competitive performance with video-based methods and proof the effectiveness of modeling co-occurrence signals and reducing effects from estimation noise and co-adaptation across signals. Besides the performance, our CoSign models are training friendly with fewer FLOPs and smaller model size. We hope our approach can inspire future studies on co-occurrence signals in CSLR and promote the development of skeleton-based CSLR approaches.

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*, pages 35–53. Springer, 2020. 1

[2] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31, 2019. 1

[3] Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. Exploring collection of sign language datasets: Privacy, participation, and model performance. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2020. 1

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. 5

[5] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10023–10033, 2020. 1, 2

[6] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5120–5130, 2022. 1, 2

[7] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13359–13368, 2021. 3

[8] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022. 2, 6, 7

[9] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *European Conference on Computer Vision*, pages 697–714. Springer, 2020. 1, 2

[10] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022. 3

[11] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3056–3065, 2017. 3

[12] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 4

[13] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, 2019. 1, 2, 6

[14] Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. Speech recognition techniques for a sign language recognition system. In *Eighth Annual Conference of the International Speech Communication Association*, 2007. 2

[15] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2, 3

[16] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[17] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11303–11312, 2021. 1, 2, 5, 6, 7

[18] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pretraining for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 6

[19] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *European Conference on Computer Vision*, pages 511–527. Springer, 2022. 6

[20] Xiaohu Huang, Hao Zhou, Bin Feng, Xinggang Wang, Wenyu Liu, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Graph contrastive learning for skeleton-based action recognition. In *International Conference on Learning Representations*, 2023. 3

[21] Youngjoon Jang, Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, Joon Son Chung, and In So Kweon. Signing outside the studio: Benchmarking background robustness for continuous sign language recognition. *Proceedings of the British Machine Vision Conference*, 2022. 1

[22] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. 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, 2017. 3

[23] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020. 1

[24] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2306–2320, 2019. 2, 3

[25] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. 5

[26] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4297–4305, 2017. 2, 6, 7

[27] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 786–792, 2018. 3

[28] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020. 1

[29] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 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, 2019. 3

[30] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11542–11551, 2021. 1, 2, 3, 4, 6

[31] Yuecong Min, Peiqi Jiao, Yanan Li, Xiaotao Wang, Lei Lei, Xiujuan Chai, and Xilin Chen. Deep radial embedding for visual sequence learning. In *European Conference on Computer Vision*, pages 240–256. Springer, 2022. 5, 6, 7

[32] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *European Conference on Computer Vision*, pages 172–186. Springer, 2020. 1, 2

[33] Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. Spatio-temporal graph convolutional networks for continuous sign language recognition. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8457–8461, 2022. 2

[34] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1497–1505, 2020. 6

[35] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4165–4174, 2019. 1, 2

[36] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 3

[37] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020. 3

[38] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *European conference on computer vision*, pages 103–118. Springer, 2018. 3

[39] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 3

[40] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *proceedings of the 28th ACM international conference on multimedia*, pages 1625–1633, 2020. 3

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 2

[43] Valerie Sutton. *Lessons in sign writing*. SignWriting, 1990. 1

[44] Kalpit Thakkar and PJ Narayanan. Part-based graph convolutional network for action recognition. In *Proceedings of the British Machine Vision Conference*, 2018. 3

[45] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012. 3

[46] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021. 4

[47] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. 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, 2022. 3

[48] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on Artificial Intelligence*, 2018. 2, 3, 4, 5

[49] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1325, 2021. 5

[50] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016, 2020. 2, 3, 6

[51] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 3

[52] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5131–5140, 2022. 2, 6