C^{2}ST: Cross-modal Contextualized Sequence Transduction for Continuous Sign Language Recognition

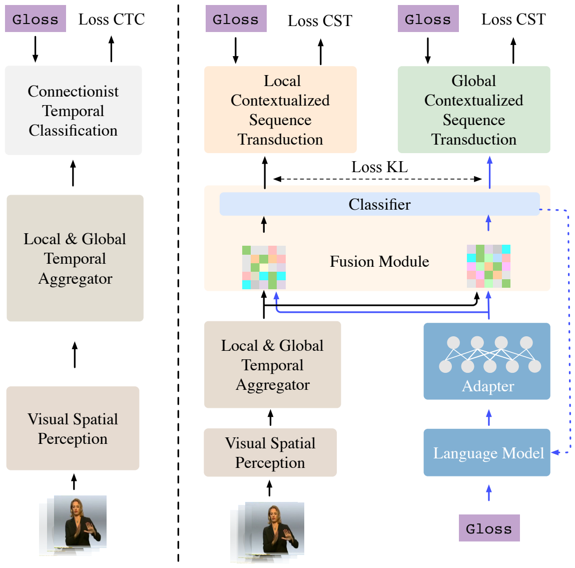
Huaiwen Zhang Zihang Guo Yang Yang Xin Liu De Hu*
College of Computer Science, Inner Mongolia University, Hohhot, 010021, China
huaiwen.zhang@imu.edu.cn, {zihang.guo, yangyang, xin.liu}@mail.imu.edu.cn, cshood@imu.edu.cn

Abstract

Continuous Sign Language Recognition (CSLR) aims to transcribe the signs of an untrimmed video into written words or glosses. The mainstream framework for CSLR consists of a spatial module for visual representation learning, a temporal module aggregating the local and global temporal information of frame sequence, and the connectionist temporal classification (CTC) loss, which aligns video features with gloss sequence. Unfortunately, the language prior implicit in the gloss sequence is ignored throughout the modeling process. Furthermore, the contextualization of glosses is further ignored in alignment learning, as CTC makes an independence assumption between glosses. In this paper, we propose a Cross-modal Contextualized Sequence Transduction (C^{2}ST) for CSLR, which effectively incorporates the knowledge of gloss sequence into the process of video representation learning and sequence transduction. Specifically, we introduce a cross-modal context learning framework for CSLR, in which the linguistic features of gloss sequences are extracted by a language model, and recurrently integrate with visual features for video modelling. Moreover, we introduce the contextualized sequence transduction loss that incorporates the contextual information of gloss sequences in label prediction, without making any independence assumptions between the glosses. Our method sets the new state of the art on three widely used large-scale sign language recognition datasets: Phoenix-2014, Phoenix-2014-T, and CSL-Daily. On CSL-Daily, our approach achieves an absolute gain of 4.9% WER compared to the best published results.

1. Introduction

Sign language, which utilizes signals like hand/arm positions, and body postures to aid individuals with hearing impairments globally, has become a powerful communication tool that enhances their quality of life. Due to the critical role of sign language and the increased availability of sign language datasets, continuous sign language recognition (CSLR) [28, 11, 40, 15, 14, 6] has gained significant attention in recent years, which enables communication between hearing-impaired people and persons without specific knowledge of sign language.


Gloss
Loss CTC
Connectionist Temporal Classification
Local & Global Temporal Aggregator
Visual Spatial Perception
Gloss
Loss CST
Gloss
Loss CST
Local Contextualized Sequence Transduction
Global Contextualized Sequence Transduction
Loss KL
Classifier
Fusion Module
Local & Global Temporal Aggregator
Adapter
Visual Spatial Perception
Language Model
Gloss

(a) Conventional (b) Ours C^{2}ST

Figure 1: Illustration of the framework of conventional sign language recognition methods (a) and our proposed Cross-modal Contextualized Sequence Transduction (C^{2}ST) (b). The proposed C^{2}ST incorporates the textual information to the original visual branch by the introduced fusion module. In addition, we introduce a novel Contextualized Sequential Transduction (CST) loss function to consider the relationship between the labels in training process.

With the aim of automatically recognizing gloss sequences (the smallest semantic unit in sign language) from untrimmed sign videos, various CSLR methods [15, 14] have been proposed. Fig.1a shows the mainstream CSLR framework, which consists of a spatial module and a temporal module to learn the spatial and temporal visual representations of sign language videos, and a Connectionist Temporal Classification (CTC) [10] loss to align the extracted visual features with the corresponding gloss sequences. For example, VAC [28] employs 2D Convolution Neural Network (CNN), 1D-TCN [23] and BiLSTM [34] to extract the visual features and uses CTC as the alignment module. C^{2}SLR [41] introduces a spatial attention consistency module and a temporal sentence embedding consistency module to learn a better spatial-temporal representation. SMKD [11] forces the visual and contextual module to focus on short-term and long-term information by the self-mutual knowledge distillation method.

Existing methods have achieved promising CSLR performance by utilizing spatial-temporal features for visual representation and CTC for alignment. However, the language prior implicit in the gloss sequence is neglected during the video modeling process. Without understanding the linguistic knowledge of glosses, existing models may perform unsatisfactorily in complex scenarios, for example, in long sequence prediction. Additionally, contextual information provided by the predicted glosses is also disregarded during alignment learning, as the widely-used CTC method makes the assumption of independence between glosses. Signs that present different semantics in different contexts may be misrecognized.

To address this issue, we propose a Cross-modal Contextualized Sequence Transduction (C^{2}ST) method for CSLR, which effectively incorporates the knowledge of gloss sequence into the process of video representation learning and sequence transduction. We first present the **cross-modal context learning** framework to equip visual representation with language prior of glosses. As shown in Fig.1b, we introduce a language model to extract linguistic features from gloss sequences, which are then combined with local and global temporal visual features with our recurrent cross-modal context fusion strategy. Specifically, the language model is first pre-trained on gloss sequences of training data. At the start of the training, the video and a blank gloss are as inputs to the spatial module and language model, respectively. Then, as an example shown in Fig.3, the gloss tokens predicted at each time step are collected and feed it into the language model recurrently. In addition, we present a novel **contextualized sequence transduction** method that further incorporates the context of gloss sequence into sequence transduction by making a dependence assumptions between the glosses. Specifically, we present a conditional gloss decoder that adopts all the predicted gloss as additional input for the prediction of the next gloss. We also introduce a sequence-level transduction calibration to counter the exposure bias [33] of sequence mapping methods with dependence assumptions.

Extensive experiments conducted on three large-scale sign language recognition datasets: Phoenix-2014 [20], Phoenix-2014-T [3], and CSL-Daily [39], demonstrate that our proposed C^{2}ST effectively utilizes gloss sequences and achieves a significant improvement over the state-of-the-art approach.

The main contributions are summarized as follows:

- A cross-modal context learning framework is proposed for CSLR, which effectively incorporates knowledge of gloss sequences into visual representations for better sign video modeling.
- A contextualized sequence transduction loss is introduced for CSLR, which leverages the contextual information of the previous gloss sequence to predict the current one, rather than making the conditional independence assumption in gloss prediction.
- The proposed C^{2}ST method achieves state-of-the-art performance on three large-scale sign language recognition datasets and outperforms state-of-the-art methods by a large margin.

2. Related work

2.1. Spatial-temporal Learning of CSLR

Continuous Sign Language Recognition (CSLR) aims to transcribe the signs of an untrimmed video into written words or glosses. Existing methods [6, 28, 11, 40, 41, 23, 25] usually consist of three modules, namely a visual spatial perception module, a temporal aggregation module, and an alignment module. Recent research has primarily focused on two directions to enhance certain tasks: incorporating external information such as hand gestures, mouth movements, and body language, as well as exploring better architectures to improve overall performance. For example, STMC [40] presents a multi-cue approach that models temporal correlations across multiple cues. C^{2}SLR [41] employs heatmaps to enhance video comprehension and utilizes a sentence embedding consistency constraint to align visual and sequential features at the sentence level, thereby boosting the representation power of both features simultaneously. In other work, VAC [28] employs auxiliary losses to improve the spatial module's performance while maintaining consistency with the temporal module. SMKD [11] proposes a knowledge distillation method that allows the visual and contextual modules to share classifier weights, thereby enhancing their classification capabilities. SEN [14] employs an improved convolution block to enhance the spatial module, which further boosts the performance.

Although these recent works have made significant strides in improving the accuracy of CSLR tasks and provide valuable insights for future research in this area. It

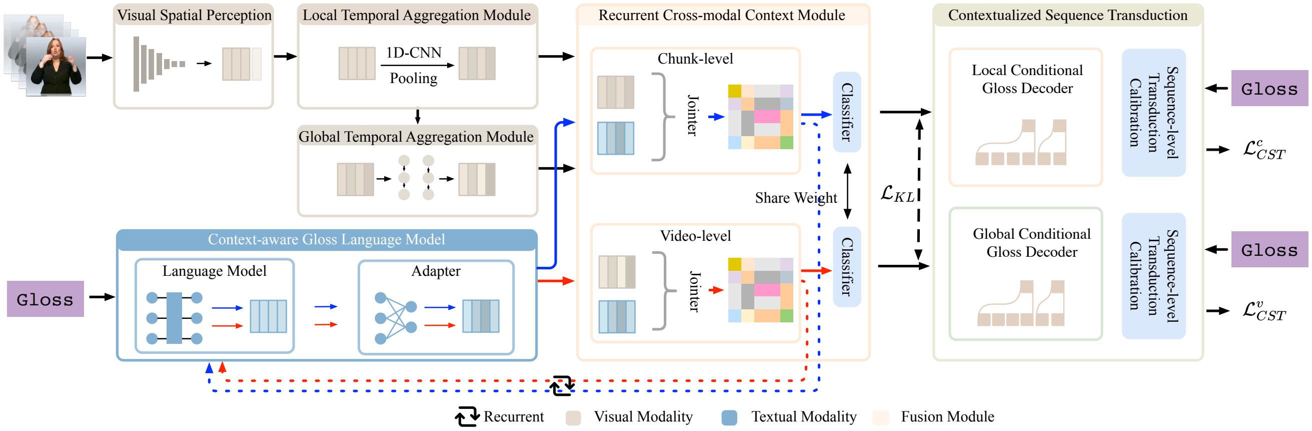

Visual Spatial Perception
Local Temporal Aggregation Module
1D-CNN
Pooling
Global Temporal Aggregation Module
Recurrent Cross-modal Context Module
Chunk-level
Jointer
Classifier
Share Weight
Video-level
Jointer
Classifier
$\mathcal{L}_{KL}$
Contextualized Sequence Transduction
Local Conditional Gloss Decoder
Sequence-level Transduction Calibration
Gloss
$\mathcal{L}^{c}_{CST}$
Global Conditional Gloss Decoder
Sequence-level Transduction Calibration
Gloss
$\mathcal{L}^{v}_{CST}$
Gloss
Context-aware Gloss Language Model
Language Model
Adapter
Recurrent
Visual Modality
Textual Modality
Fusion Module

Figure 2: The overall framework of C^{2}ST. 1) The visual spatial perception module is used to extract the spatial features from the original videos. 2) The temporal aggregation module is utilized to obtain the local and global temporal features. 3) The cross-modal context learning module involves a language model and a recurrent cross-modal context fusion module, which are used to extract textual features from the gloss sequences and recurrently integrate the gloss features with visual features. 4) The contextualized sequence transduction module is used to incorporate the contextual information of gloss features in label prediction at local-level and video-level.

can be seen that they commonly adopt a unimodal setting, which may not fully capture the continuous and cohesive nature of sign language videos. In contrast, text naturally provides clear boundaries that can assist in the alignment process due to its inherent semantic cohesion. Therefore, we propose a novel cross-modal contextualized sequence transduction method to better leverage these advantages.

2.2. Sequence Mapping of CSLR

In early research, people mainly employed hand-crafted features [20, 35] and hidden Markov models (HMMs) [21, 16] as alignment modules for sequence mapping. Then, with the development of deep learning, 2D and 3D convolutional neural networks (CNNs) have been widely used as spatial module. Researchers have proposed various hybrid models for CSLR, including the “CNN+HMM” [21, 22] and “CNN+RNN+HMM” models [22]. However, the former has limited ability to comprehend complete sign language sentences [19], while the latter still requires HMMs and acquires frame-wise labels as a supervised signal [28]. To address these challenges, the connectionist temporal classification approach is widely employed to enable end-to-end training and decoding [37, 24], which only requires sentence-level annotations. As such, the “CNN+RNN+CTC” hybrid model has become the mainstream framework for CSLR [28, 11, 40, 15, 14].

We argue the contextualization of glosses is ignored in the alignment learning of CTC, as it makes an independence assumption between glosses. Existing models predict the gloss in a single step with only the visual input as conditions, which conflicts with the contextual nature of gloss sequences or sign language sentences. Each gloss in these sequences is contextualized and given a specific meaning in a particular context. To address this limitation, we propose a novel contextualized sequential transduction loss that considers the context of each frame as well as the predicted gloss to align sign language videos.

3. Problem Definition

The objective of the Continuous Sign Language Recognition (CSLR) task is to determine the most probable alignment between the video $X = (x_1, x_2, \cdots, x_N)$ with N frames, and the gloss sequence $Y = (y_1, y_2, \cdots, y_T)$ with T glosses. Existing CSLR methods [15, 25, 41] generally consist of three steps. 1) The visual spatial perception $\mathbf{V}$ is first utilized to extract the video feature $V = \mathbf{V}(X; \theta_v) \in \mathbb{R}^{N \times d_v}$, where θ_v is the parameter of $\mathbf{V}$, d_v is the feature dimension. 2) The local temporal features $L = \mathbf{L}(V; \theta_l) \in \mathbb{R}^{S \times d_s}$ are obtained using a local temporal aggregator, where the $S \leq N$ as the temporal pooling operation, and d_s is the feature dimension. A global temporal aggregator is then adopted to aggregate the global spatio-temporal features $G = \mathbf{G}(L; \theta_g) \in \mathbb{R}^{S \times d_s}$. 3) The video feature sequence G and the gloss sequence Y are finally aligned with connectionist temporal classification [10]:

$$P(Y|X) = P(Y|G) = \sum_{Z} p(Z|G) \tag{1}$$

where $Z = (z_1, z_2, \cdots, z_S)$ represents one of the possible alignments between video X and gloss sequence Y. Note that, when remove the blank labels $\varnothing$, Z will have the same length with gloss sequence Y. The $p(Z|G)$ is defined as:

$$p(Z|G) = \prod_{s} P(z_s|G) \tag{2}$$

We argue that the dominant CSLR methods completely overlook the context information of the gloss sequence Y. Besides, the CTC loss, which assumes that the label outputs are conditionally independent of each other, further breaks the contextual information of the potential alignment path. As illustrated in Fig.2, in this paper, we propose a novel cross-modal contextualized sequence transduction method for CSLR by incorporating the semantic and context of gloss sequences in both the video modeling and sequence transduction.

4. Cross-modal Contextualized Sequence Transduction

4.1. Basic CSLR Model

As with the mainstream approaches, our CLSR method includes the spatial perception module and the temporal aggregation module. Following [15, 25, 41], we build our basic CSLR model as follows:

Visual Spatial Perception Given the frame sequence $X = (x_1, x_2, \cdots, x_N)$ with N frames, the visual spatial perception module $\mathbf{V}$ is utilized to obtain the spatial feature: $\{v_n\}_{n=1}^{N} = \mathbf{V}(\{x_n\}_{n=1}^{N}; \theta_v)$, where $v_n \in \mathbb{R}^{d_v}$. Existing method [28, 15] usually adopt ResNet [12] as $\mathbf{V}$. In this paper, we explore various visual backbones, including ResNet, Visual Transformer [9] (ViT), and Swin Transformer [26] (SW-T/S).

Local & Global Temporal Aggregation Module The frame features are then fed into the hierarchical temporal aggregation modules. The 1D-Temporal Convolutional Network (1D-TCN) [23] is widely adopted as the local temporal aggregator $\mathbf{L}$. After convolution and pooling over temporal axis, video frames are divided into S chunks with representation: $\{l_s\}_{s=1}^{S} = \mathbf{L}(V; \theta_l)$, where $l_s \in \mathbb{R}^{d_s}$. The global temporal features $G = \mathbf{G}(L; \theta_g) \in \mathbb{R}^{S \times d_s}$ are then aggregated by global temporal aggregators $\mathbf{G}$, which are often constitute by LSTM [13]. Following existing approaches [14, 28, 15], we adopt 1D-TCN and Bi-LSTM as local and global temporal aggregators, respectively.

4.2. Cross-modal Context Learning

Context-aware Gloss Language Model To incorporate contextual information from gloss, we first introduce the context-aware gloss language model $\mathbf{B}$ to capture the context information of gloss sequences. An ideal gloss language model would be pre-trained on the gloss sequence of the training dataset. However, due to the limited number of glosses in the CSLR dataset, model $\mathbf{B}$ is first pre-trained on the corpus of the corresponding language and then fine-tuned in the CSLR dataset. An $\mathrm{Adapter}$ layer is introduced to further adapt the language features to the CSLR model since the gloss sequence is slightly at variance with the human language grammar. Given a gloss sequence Y with length T, the gloss feature is calculated as follows:

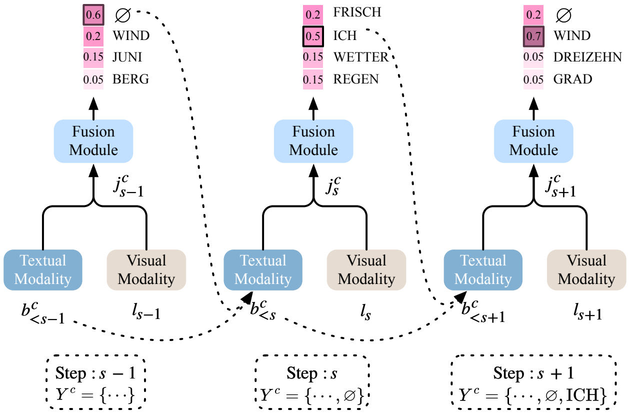


Figure 3: The illustration of the conditional gloss decoder. At each time step, all the predicted glosses from the previous step are used as the additional input to fuse with the current gloss. Note that we fuse the ∅ with the first gloss in the first iteration.

$$\{b_t\}_{t=1}^{T} = \mathrm{Adapter}(\mathbf{B}(y_i; \theta_b)) \in \mathbb{R}^{T \times d_b} \tag{3}$$

where d_b is the dimension of the gloss feature. During training, the parameters of the language model are frozen, with only the parameters of the adapter updating.

Recurrent Cross-modal Context Fusion To ensure that the context information of the gloss sequence is properly injected into the model, we design a recurrent cross-modal context fusion module to combine the video and gloss features. As shown in Fig.3, we feed the gloss sequence $Y^c_{<s} = \{y^c_i\}_{i=0}^{s-1}$ to the context-aware gloss language model to obtain the sequence representation $b^c_{<s}$, where $Y^c_{<s}$ is the gloss tokens predicted before time step s, and $y^c_0 = \varnothing$. Then visual and gloss features are fused as follows:

$$j^c_s = l_s + b^c_{<s} \tag{4}$$

where j^c_s is a cross-modal contextualized feature of chunk-level at time step s. The next gloss token y^c_s can be calculated as:

$$y^c_s = \mathrm{argmax}\, F(j^c_s) \tag{5}$$

Then the $Y^c_{<(s+1)} = Y^c_{<s} \cup y^c_s$ is ready for the next time step $s+1$. The local cross-modal contextualized features

$J^c = (j_1^c, j_2^c, \cdots, j_S^c)$ are obtained when step goes to final. The video-level cross-modal contextualized features J^v can be obtained similarly.

4.3. Contextualized Sequence Transduction

Conditional Gloss Decoder Different from CTC, which defines the $p(Z|G) = \prod_s P(z_s|G)$, we introduce the Conditional Gloss Decoder (CGD) which does not make a conditional independence assumption for label predictions. Specifically, we use the predicted gloss before step s, i.e., $z_{<s}$ as an additional input to the recurrent model when predicting the gloss at step s:

$$p(Z|J) = \prod_s P(z_s|z_{<s}, J) \tag{6}$$

Given the feature sequence $J = (j_1, j_2, \cdots, j_S)$ and the gloss sequence $Y = (y_1, y_2, \cdots, y_T)$, each possible alignment path Z is start from the initial step $s = 1$, and $t = 0$ and to the end of step $s = S + 1, t = T$. Similar as CTC [10], we define the forward variable $\alpha_s(t)$ as the probability of outputting glosses $(y_1, y_2, ..., y_t)$ up to step s, and backward variable $\beta_s(t)$ as the probability of outputting the glosses $(y_{t+1}, y_{t+2}, ..., y_T)$ starting from step s. The forward and backward variables can be calculated recursively as follows:

$$\begin{aligned}\alpha_s(t) =& \alpha_{s-1}(t-1) \times p\left(y_t|s-1, t-1\right) \\ &+ \alpha_{s-1}(t) \times p(\varnothing|s-1, t)\end{aligned} \tag{7}$$

$$\begin{aligned}\beta_s(t) =& \beta_{s+1}(t+1) \times p\left(y_{t+1}|s, t\right) \\ &+ \beta_{s+1}(t) \times p(\varnothing|s, t)\end{aligned} \tag{8}$$

where $p(y_{t+1}|s,t)$ is the probability of gloss y_{t+1} calculated by the softmax layer using the network state at node (s,t), and $p(\varnothing|s,t)$ be the probability of blank $\varnothing$. Thus, the probability for the target gloss sequence Y can be calculated by summing all the possible path, i.e., $\beta_0(0)$. The whole model can be trained by minimizing the negative log-likelihood:

$$\mathcal{L}_{CGD} = -\log(\sum_Z p(Z|J)) = -\log \beta_0(0) \tag{9}$$

Methods	Dev (%)		Test (%)	
	del/ins	WER ↓	del/ins	WER ↓
SubUNet [2]	14.6/4.0	40.8	14.3/4.0	40.7
IAN [32]	12.9/2.6	37.1	13.0/2.5	36.7
Re-Sign [21]	-	27.1	-	26.8
CNN+LSTM+HMM [19]*	-	26.0	-	26.0
SFL [29]	-	24.9	-	25.3
DNF [6]	7.8/3.5	23.8	7.8/3.4	24.4
DNF [6]*	7.3/3.3	23.1	6.7/3.3	22.9
CMA [31]	7.3/2.7	21.3	7.3/2.4	21.9
VAC [28]	8.3/3.1	21.2	8.8/3.2	21.3
SMKD [11]	6.8/2.5	20.8	6.3/2.3	21.0
STMC [40]*	7.7/3.4	21.1	7.4/2.6	20.7
C^{2}SLR [41]*	-	20.2	-	20.4
TLP [15]	7.9/2.5	19.7	8.4/2.6	20.2
SEN [25]	5.8/2.6	19.5	7.3/4.0	21.0
C^{2}ST	4.2/3.0	**17.5**	4.3/3.0	**17.7**

Table 1: Experimental results on the Phoenix-2014. "*" denotes that additional information is utilized in the approach. "del" and "ins" indicate the deletion error and insertion error, respectively. "-" means the no report provided in corresponding methods.

Sequence-level Transduction Calibration During training, the CSLR model optimizes the probability of a ground-truth gloss sequence Y by summing over all possible alignment paths. At the same time, the model also limits their vision to the training data distribution of correct glosses and suffers from exposure bias [33]. Once the CSLR model is used for inference with beam search, it may only see glosses branching off from the path learned during training and ignore all other possible sequences, even if the test distribution differs from the training distribution. Inspired by [33], we introduce the Sequence-level Transduction Calibration (STC) to our CGD loss by paying more attention to potential sequences predicted by the model that have a smaller sequence-level error:

$$\mathcal{L}_{CST} = -\log(\sum_Z \frac{P(Z|J)}{\mathrm{STC}(Z,Y)}) \tag{10}$$

where $P(Z|J)$ is the probability of an alignment Z as estimated by CGD and $\mathrm{STC}(Z,Y)$ is the sequence-level transduction calibration term, for example, the edit distance between the ground-truth sequence Y and Z after removing the blanks.

4.4. Training and Inference

Following [28], we adopt a KL divergence function $\mathcal{L}_{KL}$ to maintain the consistency between the chunk-level fusion feature J^c and the video-level fusion feature J^v. The final objective function is a combination of the CST loss and the KL divergence loss, which is expressed as:

$$\mathcal{L} = \mathcal{L}_{CST}^c + \mathcal{L}_{CST}^v + \mathcal{L}_{KL} \tag{11}$$

where $\mathcal{L}_{CST}^c$ and $\mathcal{L}_{CST}^v$ are the proposed contextualized sequence transductions at chunk-level and video-level, respectively. In the training phase, we can use Teacher Forcing [38] to update the cross-modal context learning module. In inference time, it is not necessary to run the conditional gloss decoder on all possible alignments. We can do beam search [36] decoding using the probability distribution estimated from the model at each time step.

Methods	WER ↓	
	Dev (%)	Test (%)
CNN-LSTM-HMM [19]	24.5	26.5
SFL [29]	25.1	26.1
Joint-SLRT [4]	24.6	24.5
CNN-LSTM-HMM [19]*	22.1	24.1
SMKD [11]	20.8	22.4
TLP [15]	19.4	21.2
STMC [40]*	19.6	21.0
SEN [25]	19.3	20.7
C^{2}SLR [41]*	20.2	20.4
C^{2}ST	**17.3**	**18.9**

Table 2: Experimental results on the Phoenix-2014-T. "*" denotes that additional information is utilized in the approach.

Methods	Dev (%)		Test (%)	
	del/ins	WER ↓	del/ins	WER ↓
LS-HAN [17]	14.6/5.7	39.0	14.6/2.8	39.4
FCN [5]	12.8/4.0	33.2	12.6/3.7	32.5
DNF [6]	-	32.8	-	32.4
Joint-SLRT [4]	-	33.1	-	32.0
BN-TIN [39]	13.9/3.4	33.6	13.5/3.0	33.1
SEN [25]	-	31.1	-	30.7
C^{2}ST	9.3/2.7	**25.9**	9.0/2.7	**25.8**

Table 3: Experimental results on the CSL-Daily.

5. Experiments

5.1. Experiment Settings

Dataset. We conduct experiments to evaluate the performance of our proposed method on three widely used datasets: Phoenix-2014 [20], Phoenix-2014-T [3], and CSL-Daily [39], from which Phoenix-2014 and Phoenix-2014-T consist of German corpora, while CSL-Daily consists of a Chinese corpus.

- **Phoenix-2014** [20] is recorded from the German TV weather forecasts and performed by nine hearing sign language interpreters. All recorded videos are at 25 frames per second with a frame size of 210 by 260 pixels. It contains 6,841 sentences and 1,295 distinct signs, which are divided into 5,672 training samples, 540 development samples, and 629 testing samples.
- **Phoenix-2014-T** [3] can be seen as an extension of Phoenix-2014, but does not overlap with it. It includes sign language videos, sign-gloss annotations, and German translations, all of which are divided into parallel sentences. It has 8,247 phrases totaling 1,085 signs in its lexicon. Specifically, 7,096, 519, and 642 samples are utilized for training, development, and testing, respectively.
- **CSL-Daily** [39] is a challenging Chinese sign language dataset collected from indoors with 20654 sentences, split into 18401 training samples, 1077 development (Dev) samples, and 1176 testing (Test) samples. The topic of the dataset revolves around people's daily life, such as travel, shopping, medical care, etc.

Evaluation Metric. The Word Error Rate (WER) is used to evaluate the similarity between the predicted sentence and the reference sentence. It calculates the minimum number of substitution (#sub), insertion (#ins), and deletion (#del) operations from the predicted sentence to the reference sentence. The reference (#ref) represents the total number of words in the gloss sequence. The metric is defined as:

$$\text{WER} = \frac{\#\text{sub} + \#\text{ins} + \#\text{del}}{\#\text{ref}} \tag{12}$$

The best performance is highlighted with **bold** in our following experiments.

Implementation Details. The video frames are resized to 256×256 and then cropped to resolution 224×224. In the training phase, we adopt random cropping, random flipping, and temporal scaling as data augmentation strategies. Only center cropping is adopted in the testing phase. We adopt the Swin-T [26] as the visual backbone. The Swin Transformer is pre-trained on ImageNet [7]. The layer numbers are set as $\{2, 2, 6, 2\}$, the head number to $\{3, 6, 12, 24\}$, the window size to 7 and the output size to 768. The 1D-TCN with Temporal Lift Pooling (TLP) [15] is utilized as a local temporal aggregator, which is composed of the $K5, P2, K5, P2$ layers, and the output dimension d_s is set to 1024. The K and P represent a 1D convolutional layer and a pooling layer, respectively. The number represents the size of the convolution kernel and the pooling kernel. The global temporal model consists of two BiLSTM layers. We adopt the pre-trained BERT-base [8] as the text module. Then the dimension is projected to $d_b = 1024$ by the followed adapter layer. The Adam [18] is adopted as the optimizer with a weight decay of 1e-3. The learning rate is initialized as 1e-4.

5.2. Comparison with the State-of-the-art.

We showcase the comparison results with several state-of-the-art approaches on all three datasets in Tab.1, Tab.2 and Tab.3. We observe that the proposed C^{2}ST outperforms other baselines and achieves the state-of-the-art. Specifically, our method exceeds the state-of-the-art approaches by 2.5% on test set of Phoenix-2014 and 1.5% of Phoenix-T 2014. In Phoenix-2014-T, we achieve a 3.0% improvement in the development set and 1.8% in the test set. We gain an enhancement of 5.2% in the development set and 4.9% in the test set. In particular, compared with the methods

CCL	CST	WER(%) ↓	
		Dev	Test
×	×	19.2	19.9
√	×	18.5	19.0
×	√	18.8	19.1
√	√	**17.5**	**17.7**

Table 4: Ablation studies on cross-modal context learning (CCL) and contextualized sequence transduction (CST).

$\mathcal{L}_{CTC}$	$\mathcal{L}_{CGD}$	$\mathcal{L}_{CST}$	WER(%) ↓	
			Dev	Test
√	×	×	18.5	19.0
×	√	×	17.9	18.3
×	×	√	**17.5**	**17.7**

Table 5: Ablation studies on the different loss functions in Phoenix-2014. The $\mathcal{L}_{CTC}$ indicates the original CTC loss function.

Chunk-level Fusion	Video-level Fusion	WER(%) ↓	
		Dev	Test
×	×	18.8	19.1
√	×	18.6	18.8
×	√	17.9	18.2
√	√	**17.5**	**17.7**

Table 6: Ablation studies on the proposed chuck-level and video-level cross-modal fusion.

VSP	WER(%) ↓	
	Dev	Test
ResNet18	18.1	18.3
Swin-T	**17.5**	**17.7**
Swin-S	17.8	18.2

Table 7: Ablation studies of the Visual Spatial Perceptions (VSPs) in Phoenix-2014.

LM	WER(%) ↓	
	Dev	Test
-	18.8	19.1
Word Emb	17.6	18.3
GloVe	17.7	18.2
BERT-Base	**17.5**	**17.7**

Table 8: Ablation studies on the proposed chuck-level and video-level cross-modal fusion.

(labeled with "*") that use information like hands, posture, and face, our method achieves a significant improvement, which proves the effectiveness of language prior of glosses.

5.3. Ablation study

In this section, we conduct ablation studies to validate the effectiveness of our contributions. All experiments are conducted on Phoenix-2014.

Ablation on the proposed components. Ablation studies on cross-modal context learning (CCL) and contextualized sequence transduction loss (CST) are presented in Tab.4. By removing the CCL and CST, the model is degenerated to a TLP method with Swin-Transformer as the visual spatial perception module. As shown in row 2, by introducing cross-modal context learning, the model gains 0.9% WER in the test set, which demonstrate the importance of gloss contextual information in sign language recognition. The proposed contextualized sequence transduction (row 3) brings a considered improvement in both development and test set, which verifies that the dependency relationships between glosses are necessary for better CSLR.

Ablation on different loss functions. In Tab.5, we conduct the ablation studies on the loss functions. Specifically, with the help of cross-modal context learning, the conventional CTC also achieves a competitive recognition performance (row 1). Considering the relationship between the glosses, the proposed conditional gloss decoder achieves the significant recognition performance in row 2. We introduce Sequence-level Transduction Calibration (STC) into our $\mathcal{L}_{CGD}$, termed as $\mathcal{L}_{CST}$, which focuses more on the potential sequences with small sequence-level errors predicted by the model, further boosting the performance.

Ablation on Cross-modal fusion. We conduct the ablation studies on the cross-modal fusion strategy in the cross-modal context learning, in Tab.6. We observe that the chunk-level constraint alone did not perform well, potentially due to the lack of global information in the training process. It is note that although adopting the video-level constraint alone results in a considerably better performance than using only the Chunk-level fusion, further improvement can be achieved by adopting both Chunk-level and video-level fusion. The experimental results demonstrate that both Chunk- and Video-level fusion are complementary and the combination of them could brings superior recognition performance.

Ablation on the visual spatial perceptions. In Tab.7, we investigate the influence of different visual backbones in sign language recognition. Specifically, we adopt ResNet [12], and Swin Transformers [26] with different scales as the visual spatial perception module, respectively. We observe that the best performance is shown at the Swin-T, which may be caused by overfitting of the alignment module, leading to insufficient training of the feature extractors in CSLR training [28].

Ablation on the language models. We investigate the influence of different language models, e.g. Word Embedding [1], GloVe [30], and BERT [8], in extracting the representations of gloss sequences. Specifically, the predicted gloss sequence is represented by [CLS] in BERT and the average of gloss tokens for Glove and Word Embedding. As shown in Tab.8, BERT demonstrates superior performance, due to its massive data training and ability to capture contextual word relationships effectively. The GloVe and word embedding also shows competitive performance, which indicates the effectiveness of cross-modal context learning.

5.4. Visualization

The visualization results of alignment process. In Fig.4, we visualize the alignment case between glosses and video

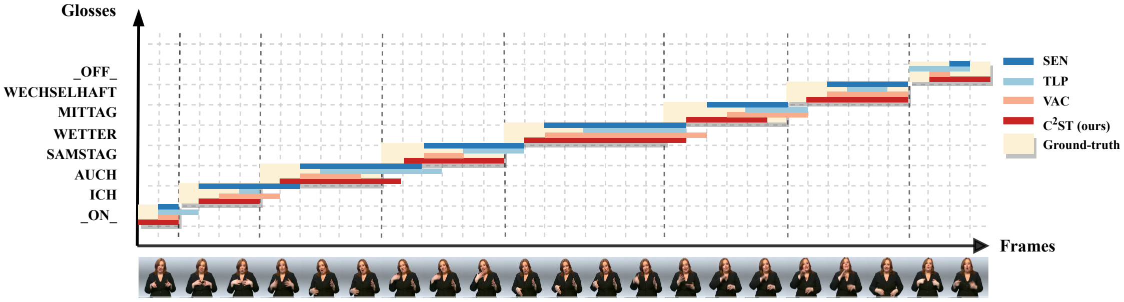


Figure 4: An visualization example of the alignment process on the Phoenix-2014 test set. The horizontal and vertical axis represent the frames and gloss sequences, respectively. Our method achieves better alignment results compared to the other methods.

Methods	Params	Throughput Train/Infer	GFLOPs Train/Infer	WER(%) Train/Infer
VAC$_{Resnet}$	**34.3**	**22.6/17.0**	562/**567**	21.2/21.3
SEN$_{Resnet}$	34.5	18.7/15.5	568/578	19.5/21.0
TLP$_{Resnet}$	59.5	18.5/17.0	563/573	19.7/20.2
C^{2}ST$_{Resnet+WE}$	60.6	18.8/15.3	**560**/570	18.2/18.6
C^{2}ST$_{SW+WE}$	78.2	7.5/4.4	1343/1368	17.6/18.3
C^{2}ST$_{SW+BERT}$	187.8	7.3/3.0	1635/1384	**17.5/17.7**

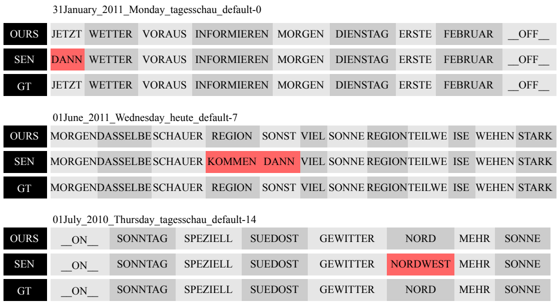
Table 9: Method efficiency analysis on Phoenix2014 with THOP[27] tool. Throughput (videos/s) is measured on a A100 card with batch size 1.

frames in our and some state-of-the-arts. All baselines are based on CTC to align the gloss sequences and video. The visualization results show that the alignment by C^{2}ST outperforms all baselines and more closer to the ground-truth. The reason may be that our C^{2}ST leverages the cross-modal context learning to fully exploit the prior knowledge from gloss sequences.

The visualization results of recognition. In Fig.5, we show some qualitative results of our method compared to the state-of-the-art approach SEN [25]. All examples are selected from the Phoenix-2014 dataset. The conditional independence assumption in CTC overlooks the context of the gloss sequence and exploits insufficient knowledge to predict, resulting in inferior performance in SEN. Our C^{2}ST could give a more accurate prediction is attributed to the contextual information from glosses.

6. Conclusion and Limitation

In this paper, we present a Cross-modal Contextualized Sequence Transduction (C^{2}ST) method for Continuous Sign Language Recognition (CSLR), which integrate the knowledge of gloss sequences into the process of video representation learning and sequence transduction. Specifically, we present a cross-modal context learning framework for CSLR, which equip the visual representation of sign



31January_2011_Monday_tagesschau_default-0

OURS	JETZT	WETTER	VORAUS	INFORMIEREN	MORGEN	DIENSTAG	ERSTE	FEBRUAR	__OFF__
SEN	DANN	WETTER	VORAUS	INFORMIEREN	MORGEN	DIENSTAG	ERSTE	FEBRUAR	__OFF__
GT	JETZT	WETTER	VORAUS	INFORMIEREN	MORGEN	DIENSTAG	ERSTE	FEBRUAR	__OFF__

01June_2011_Wednesday_heute_default-7

OURS	MORGEN	DASSELBE	SCHAUER	REGION	SONST	VIEL	SONNE	REGION	TEILWE	ISE	WEHEN	STARK
SEN	MORGEN	DASSELBE	SCHAUER	KOMMEN	DANN	VIEL	SONNE	REGION	TEILWE	ISE	WEHEN	STARK
GT	MORGEN	DASSELBE	SCHAUER	REGION	SONST	VIEL	SONNE	REGION	TEILWE	ISE	WEHEN	STARK

01July_2010_Thursday_tagesschau_default-14

OURS	__ON__	SONNTAG	SPEZIELL	SUEDOST	GEWITTER	NORD	MEHR	SONNE
SEN	__ON__	SONNTAG	SPEZIELL	SUEDOST	GEWITTER	NORDWEST	MEHR	SONNE
GT	__ON__	SONNTAG	SPEZIELL	SUEDOST	GEWITTER	NORD	MEHR	SONNE

Figure 5: Visualization examples from the Phoenix-2014 test set. For each example, we show our method's results in row 1, while rows 2 and 3 display the results of SEN and ground-truth, respectively. Each example is labeled with its name above, and contain gloss units as gray blocks, with red blocks indicating incorrect predictions.

video with language prior of glosses. Moreover, we propose a contextualized sequence transduction loss function to fuse the contextual information in gloss sequences when alignment. Extensive experimental results on three large-scale CSLR datasets show that the proposed C^{2}ST outperforms state-of-the-art CSLR methods by a large margin.

While the proposed method demonstrates superior performance, it does not appear to effectively balance accuracy and efficiency, as shown in Tab. 9. Striving to maintain high efficiency while enhancing performance remains challenge.

Acknowledgement This work is supported by National Natural Science Foundation of China (No. 62206137, 62201297, 62206200, 62106262, 62276257 and 62036012), Inner Mongolia Natural Science Foundation (No.2022MS06025), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (No.NJYT23105), Tianjin Natural Science Foundation (No.22JCQNJC00940), Beijing Natural Science Foundation (No.L221004).

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146, 2017. 7

[2] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *ICCV 2017*, pages 3075–3084. IEEE Computer Society, 2017. 5

[3] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society, 2018. 2, 6

[4] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR 2020*, pages 10020–10030. Computer Vision Foundation / IEEE, 2020. 6

[5] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *ECCV 2020*, volume 12369, pages 697–714. Springer, 2020. 6

[6] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Trans. Multim.*, 21(7):1880–1891, 2019. 1, 2, 5, 6

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, pages 248–255. IEEE Computer Society, 2009. 6

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 6, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021*. OpenReview.net, 2021. 4

[10] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML 2006*, volume 148, pages 369–376. ACM, 2006. 2, 3, 5

[11] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *ICCV 2021*, pages 11283–11292. IEEE, 2021. 1, 2, 3, 5, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 4, 7

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 4

[14] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Self-emphasizing network for continuous sign language recognition. *CoRR*, abs/2211.17081, 2022. 1, 2, 3, 4

[15] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Temporal lift pooling for continuous sign language recognition. *CoRR*, abs/2207.08734, 2022. 1, 3, 4, 5, 6

[16] Chung-Lin Huang and Wen-Yi Huang. Sign language recognition using model-based tracking and a 3d hopfield neural network. *Mach. Vis. Appl.*, 10(5/6):292–307, 1998. 3

[17] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *AAAI 2018*, pages 2257–2264. AAAI Press, 2018. 6

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR 2015*, 2015. 6

[19] Oscar Koller, Necati Cihan Camgöz, Hermann Ney, and Richard Bowden. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9):2306–2320, 2020. 3, 5, 6

[20] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Comput. Vis. Image Underst.*, 141:108–125, 2015. 2, 3, 6

[21] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *CVPR 2017*, pages 3416–3424. IEEE Computer Society, 2017. 3, 5

[22] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *BMVC 2016*. BMVA Press, 2016. 3

[23] Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *ECCV 2016*, volume 9915, pages 47–54, 2016. 2, 4

[24] Hongzhu Li and Weiqiang Wang. Reinterpreting CTC training as iterative fitting. *Pattern Recognit.*, 105:107392, 2020. 3

[25] Zekang Liu Lianyu Hu, Liqing Gao and Wei Feng. Self-emphasizing network for continuous sign language recognition. In *Thirty-seventh AAAI conference on artificial intelligence*, 2023. 2, 3, 4, 5, 6, 8

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV 2021*, pages 9992–10002. IEEE, 2021. 4, 6, 7

[27] Lyken17, HaoKang-Timmy, lvmingzhe, and ttumiel. Thop: Pytorch-opcounter. `https://github.com/Lyken17/pytorch-OpCounter`, 2019. 8

[28] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV 2021*, pages 11522–11531. IEEE, 2021. 1, 2, 3, 4, 5, 7

[29] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *ECCV 2020*, volume 12361, pages 172–186. Springer, 2020. 5, 6

[30] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP 2014*, pages 1532–1543. ACL, 2014. 7

[31] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *ACMMM 2020*, pages 1497–1505. ACM, 2020. 5

[32] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *CVPR 2019*, pages 4165–4174. Computer Vision Foundation / IEEE, 2019. 5

[33] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR 2016*, 2016. 2, 5

[34] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 2

[35] Chao Sun, Tianzhu Zhang, Bing-Kun Bao, Changsheng Xu, and Tao Mei. Discriminative exemplar coding for sign language recognition with kinect. *IEEE Trans. Cybern.*, 43(5):1418–1428, 2013. 3

[36] Christoph Tillmann and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Comput. Linguistics*, 29(1):97–133, 2003. 5

[37] Chengcheng Wei, Jian Zhao, Wengang Zhou, and Houqiang Li. Semantic boundary detection with reinforcement learning for continuous sign language recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 31(3):1138–1149, 2021. 3

[38] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280, 1989. 5

[39] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR 2021*, pages 1316–1325. Computer Vision Foundation / IEEE, 2021. 2, 6

[40] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *IAAI 2020*, pages 13009–13016. AAAI Press, 2020. 1, 2, 3, 5, 6

[41] Ronglai Zuo and Brian Mak. C^{2}slr: Consistency-enhanced continuous sign language recognition. In *CVPR 2022*, pages 5121–5130. IEEE, 2022. 2, 3, 4, 5, 6