Prior-aware Cross Modality Augmentation Learning for Continuous Sign Language Recognition

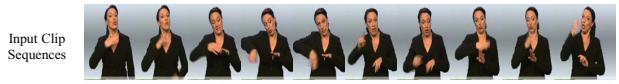
Hezhen Hu, Junfu Pu, Wengang Zhou, *Senior Member, IEEE,* Hang Fang, and Houqiang Li, *Fellow, IEEE*

***Abstract*—Continuous sign language recognition (CSLR) aims to map a sign video into a sentence of text words in the same order as the signs. Generally, word error rate (WER), *i.e.,* editing distance, is adopted as the main evaluation metric. Since this metric is not differentiable, current deep-learning-based CSLR methods usually resort to connectionist temporal classification (CTC) loss during optimization, which maximizes the posterior probability over the sequential alignment. Due to the optimization gap between CTC loss and WER, the decoded sequence with the maximum probability in CTC may not be the one with the lowest WER. To tackle this issue, we propose a novel prior-aware cross modality augmentation learning method. In our approach, we first generate the pseudo video-text pair by cross modality editing, *i.e.,* substitution, deletion and insertion on the paired real video-text data. To ensure the pseudo data quality, we guide the editing with both textual grammar prior and visual pose transition consistency prior. In this way, the generated pseudo video and text sentence follow the underlying distribution of the sign language data, and sever as more genuine hard examples for the cross modality representation learning of our CSLR task. Based on the real and generated pseudo data, we optimize our CSLR framework with three loss terms. We evaluate our approach on popular large-scale CSLR datasets and extensive experiments demonstrate the effectiveness of our method.**

***Index Terms*—cross modality augmentation learning, editing with prior incorporated, continuous sign language recognition.**

I. Introduction

SIGN language serves as a primary tool during communication among deaf people. It is a kind of visual language with specific grammar and lexicon, and conveys semantic meaning via manual and non-manual features. Specifically, the manual features include hand motion, orientation and position, *etc.*, while non-manual features refer to facial expressions and head movements, *etc.* These characteristics make it non-trivial for common people to master it, which leads to communication gap with deaf people. To this end, automatic continuous sign language recognition (CSLR) is widely studied, which maps the input sign video to the corresponding text in the same presenting order. Due to expensive annotation costs, current continuous sign videos are generally weakly labeled without frame-level alignment, which leads to more difficulty in learning correct sequence-to-sequence transformation.



Input Clip Sequences									
Ground Truth	__ON__	SONNTAG	SPEZIELL	SUEDOST	GEWITTER	NORD	MEHR	SONNE	WER
Candidate 1		SONNTAG	SPEZIELL	loc-SUEDOST	GEWITTER	loc-WEST	KOENNEN	SONNE	50.0%
Candidate 2		SONNTAG	SPEZIELL	loc-SUEDOST	GEWITTER	loc-WEST	MEHR	SONNE	37.5%
Candidate 3		SONNTAG	SPEZIELL	loc-SUEDOST	GEWITTER	loc-WEST		SONNE	50.0%
Candidate 4	__ON__	SONNTAG	SPEZIELL	loc-SUEDOST	GEWITTER	loc-WEST	KOENNEN	SONNE	37.5%
Candidate 5	__ON__	SONNTAG	SPEZIELL	loc-SUEDOST	GEWITTER	loc-WEST	MEHR	SONNE	25.0%

Fig. 1. A real example illustrating of the inconformity between the CTC objective and WER evaluation metric. We demonstrate the decoding result on a sample after CTC optimization, showing five predicted sentences with descent decoding probabilities (Candidate 1 to 5). The sentence with the fifth highest decoding probability exhibits the best performance. The box with the red background denotes the false prediction.

To address CSLR, early works [1], [2] feed hand-crafted features into the statistical sequential models, *e.g.* Hidden Markov Model (HMM). Recently, most state-of-the-art methods [3]–[6] utilize the advanced deep-learning-based techniques, *e.g.* Convolutional Neural Network (CNN) [7]–[9], Recurrent Neural Network (RNN) [10], [11] and Transformer [12], for representation learning. The learned deep features of sign videos are usually processed by connectionist temporal classification (CTC) [13] to deal with the sequence correspondence under weak-labeled data. For these CTC-based methods, beam search is utilized for iterative decoding, which produces the same number of candidates as the beam width. To evaluate the decoding quality on the CSLR task, word error rate (WER) is utilized as the evaluation metric, which is defined by the least operations, *i.e.,* substitution, deletion and insertion, to transform the prediction to the target sentence.

However, due to the inconsistency between the CTC loss and WER based evaluation metric, the candidate with the maximum decoding probability may not be the best one under the WER metric. As illustrated in Figure 1, the candidate with the fifth highest decoding probability (Candidate 5) exhibits the best performance on the WER metric. To quantitatively explore the universality of this problem, we calculate the top-K WER metric on the RWTH-Phoenix dataset. Top-K WER is defined by first choosing the candidate with the lowest WER out of K decoded candidates and calculating the average WER over the whole dataset. When all the sentences with the maximum decoding probability exhibit the lowest WER among K decoded sentences, the WER metric is equal to

This work was supported by the National Natural Science Foundation of China under Contract U20A20183 and 62021001. It was also supported by GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC.

Hezhen Hu, Wengang Zhou and Houqiang Li are with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230027, China e-mail: alexhu@mail.ustc.edu.cn, {zhwg, lihq}@ustc.edu.cn). Hang Fang is with Anhui University, Hefei, China. Junfu Pu is with the Tencent ARC Lab.

Corresponding authors: Wengang Zhou and Houqiang Li.

Top-K WER. To some extent, Top-K WER indicates the lower bound over the predicted decoding results. Based on the decoding results of the method [3], the WER on the test set of the RWTH-Phoenix dataset is 23.8%, while Top-5 WER further leads to better performance, *i.e.,* 19.7%. In order to minimize the performance gap, we attempt to make the candidate with the maximum decoding probability correspond to the lowest WER.

Based on the above motivation, we propose a novel cross modality augmentation learning paradigm to further boost the performance of continuous SLR. This paradigm first performs cross modality editing to generate pseudo video-text pair. For the pseudo video-text pair, we mimic the calculation process of WER and perform the same operations on the original video-text pair. To ensure the pseudo data quality, we incorporate prior for both the video and text editing process. Given the real and pseudo video-text pairs, our framework performs cross modality representation learning to minimize the inconsistency between CTC loss and WER evaluation metric. During this representation learning, we further design two extra loss terms apart from the necessary alignment loss, *i.e.,* semantic correspondence loss and real-pseudo discriminative loss. Semantic correspondence loss is adopted to map the video and text into a unified semantic space. In this unified space, real-pseudo discriminative loss enables the framework to distinguish the fine-grained difference between real and pseudo video-text pairs. Specifically, we divide features of these video-text pairs into two groups with real video and real text as the anchor, respectively. In each group, we attempt to minimize the distance between the anchor and the positive sample while maximizing the distance between the anchor and the negative one.

Our solution aims to meet the requirement of narrowing the gap between CTC loss and WER evaluation metric as follows. 1) It generates the pseudo video-text pair based on operations in WER calculation and further ensures its quality via inserting prior during editing. In this way, the generated video-text pair is a harder sample for the framework to distinguish from the perspective of WER evaluation. 2) The following training process adopts designed loss terms to embed the framework with awareness of the subtle difference between the real and pseudo video-text pair. This fine-grained awareness narrows the gap between CTC loss and WER evaluation and finally improves the recognition performance with notable gains.

This work is an extension of our conference paper [14]. Different from [14], during cross-modality editing, we incorporate prior to guide the editing process, *i.e.,* modeling the joint text distribution for text editing and constraining the transition smoothness for video editing. In this way, the generated pseudo text sentences are grammatical while the generated pseudo videos preserve the consistency in visual transition. Such quality-enhanced pseudo data improves the framework with a notable performance gain. Besides, we provide more experiments on benchmark datasets to validate its effectiveness. Furthermore, we present more discussion in related work as well as future work and the broader impact of our work.

The remainder of this paper is organized as follows. Section II reviews the related works about sign language research, data augmentation and contextual text representation modeling. Section III elaborates the detailed cross modality augmentation learning paradigm. In Section IV, we introduce the evaluation protocol and conduct extensive experiments and analysis. Finally, we conclude this work in Section V.

II. Related Work

In this section, we first briefly review the key components in continuous sign language recognition, *i.e.,* visual representation learning and sequence-to-sequence mapping, and summarize the research efforts towards relieving over-fitting in this area. Then we review the related data augmentation techniques. Finally, we introduce contextual text representation modeling.

A. Continuous Sign Language Recognition

Continuous sign language recognition aims to map the sign video to its corresponding text in the same presenting order. For continuous SLR, the visual encoder first extracts semantic representations from the sign video. Then the sequential module performs the mapping from the extracted semantics to the text sequence.

Visual representation learning. Since hand acts a dominant role in the expression of sign language, early works utilize hand-crafted features, *e.g.* HOG [15], [16], SIFT [17] and Grassmann covariance matrix (GCM) [18], to represent hand motion, shape and orientation. With the development of deep learning, Convolutional Neural Networks (CNNs) [7], [9], [19], [20] become the most powerful feature extractor. With this trend, researchers turn to explore the suitable CNN architecture to directly extract discriminative visual representation from the full video sequence [4], [14], [21]–[28]. There exist works utilizing 3DCNN [4], [21], [29], [30] and 2DCNN-TCN [3], [14], [28], [31] as the backbone to extract spatial-temporal discriminative cues simultaneously or sequentially, respectively. IAN [4] utilizes 3D-ResNet [20] for visual representation. DNF [3] subtly designs 2DCNN with the 1D temporal convolution, which has become one of the mainstream baseline methods. With its simplicity and effectiveness, we utilize 2DCNN-TCN for visual representation learning.

Sequence correspondence learning. Embedded with the visual representation, the sequential model attempts to learn the correspondence between the visual representation and sign gloss sequence. There exist three representative methods, *i.e.,* Hidden Markov Model (HMM) [32], [33], [33]–[35], encoder-decoder [36], [37] framework and Recurrent Neural Network (RNN) with Connectionist Temporal Classification (CTC) [3]–[5], [38], [39]. Oscar *et al.* [32] exploit the coordination of hybrid HMMs with the CNN-LSTM architecture, leveraging intermediate synchronisation constraints among multiple streams. Another representative method is encoder-decoder. Guo *et al.* [36] utilizes the encoder-decoder framework with hierarchical deep recurrent fusion to merge cues from RGB and skeleton modalities.

The effectiveness of RNN, *e.g.* Gated Recurrent Unit (GRU) [40] and Long-Short Term Memory (LSTM) [10],

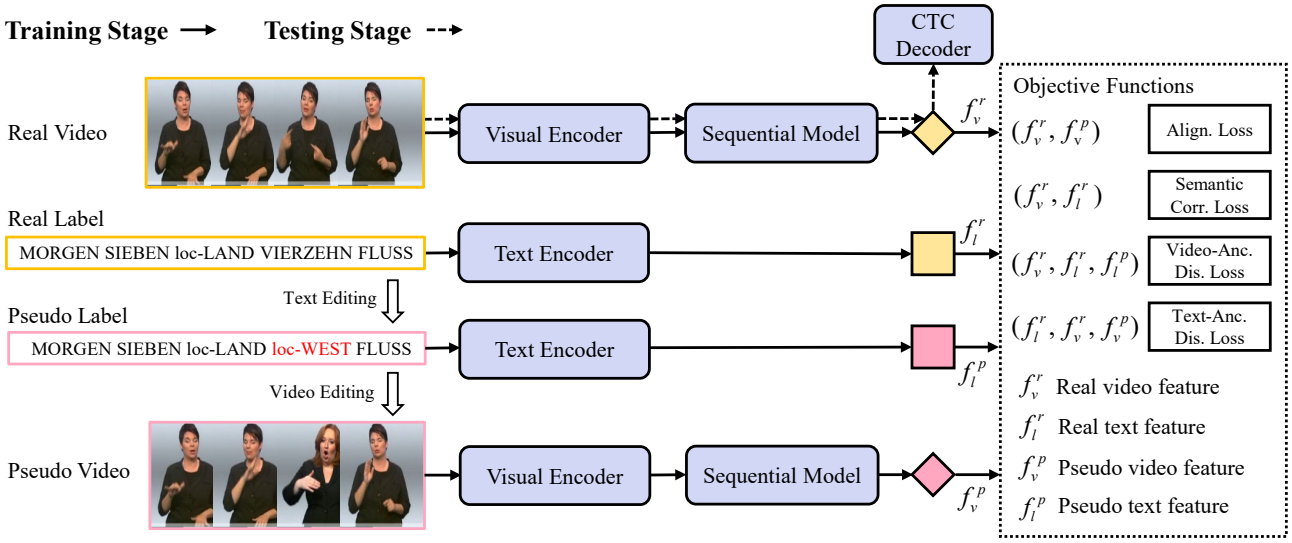


Fig. 2. Overview of our proposed framework. The framework consists of a CNN-TCN visual encoder, sequential model and text encoder. Jointly with the prior-guided cross modality editing, we design multiple loss terms to leverage both real and pseudo video-text pairs for boosting the performance on continuous SLR. During the testing stage, only the components connected in dash lines are used, *i.e.,* real video, visual encoder, sequential model and CTC decoder.

has been validated for modeling contextual information. CTC [13] is designed to deal with two unsegmented sequences without precise alignment, which has been successfully applied to speech recognition [41] and handwriting recognition [42], *etc.* Typically, for continuous SLR, bidirectional LSTM (BLSTM) is utilized with CTC to model the contextual cues from both forward and backward directions. BLSTM-CTC has become a commonly used sequential module [3], [5], [39]. In our work, we utilize BLSTM-CTC for sequential learning.

Efforts towards relieving over-fitting. Given the limited scale of annotated data, over-fitting is a common issue from the data perspective, which exists among current continuous SLR methods and leads to degraded performance. One common solution is iterative refinement [3], [4], [38]. It aims to extract the pseudo alignment between video clips and sign glosses. The alignment can serve as a label for refinement of the visual encoder. After refinement, the trained visual encoder can provide better initialization for end-to-end fine-tuning. In this way, this strategy is performed iteratively for boosting the recognition performance. Pu *et al.* [4] propose to utilize soft-DTW as the alignment constraint. DNF [3] proposes a pseudo alignment generation method, which aligns the probability matrix and corresponding gloss sequence via dynamic programming.

There exist methods to relieve this over-fitting issue from other perspectives. STMC [6] proposes a novel fusion strategy to merge multi-cue information under multi-task learning. SFLM [28] proposes a stochastic frame dropping mechanism and a gradient stopping method. Different from them, we notice the commonly existing inconsistency between generally adopted CTC loss and WER evaluation and aim to relieve over-fitting by solving this issue.

B. Data Augmentation

Data augmentation is a kind of technique to reduce over-fitting, which has been commonly utilized in deep learning [43], [44]. It performs transformation on the original data to cover the possible variants in real-world scenarios. With more data introduced, it relieves the deep-learning-based networks from falling into local optima, thus enhancing the performance on the downstream tasks. Different data augmentation techniques have been designed for different tasks. For image-based tasks, *e.g.* image classification and object detection, common augmentation skills include geometric transformation (*e.g.* random cropping and rotation), color jittering and random erasing, *etc.* For video-based tasks, *e.g.* action recognition and video tracking, more augmentation methods focus on the temporal dimension, such as temporal random sampling or padding. For the Natural Language Processing (NLP) task, augmentation methods include random insertion or deletion, synonym replacement, and sentence rotation, *etc.* Notably, these data augmentation techniques mainly aim to create augmented samples sharing the same label as the original one. In contrast, we generate pseudo data on both video sequence and its corresponding text sentence label as negative hard examples for discriminative learning. In this way, we build our new cross modality augmentation learning paradigm by leveraging both the real and pseudo video-text pair.

C. Contextual Text Representation Modeling

In related natural language processing (NLP), there exist works modeling contextual language representations via pre-training [45]–[49]. Context2Vec [45] learns contextual representations through a task to predict a single word from both left and right context based on LSTMs. With the emergence of the milestone work Transformer [12], it has become the generic block for modeling contextual text representation.

Based on Transformer, there exist works conducting pre-training, *e.g.* GPT [47], BERT [48], XLM [49] and *etc.* Among them, BERT [48] is the most popular one and benefits the downstream tasks. It designs two pre-training strategies, *i.e.,* masked language modeling (MLM) and Next Sentence Prediction (NSP). Different from BERT, we only adopt its pre-training strategy to model the context in the text corpora and guide our editing process.

III. Our Approach

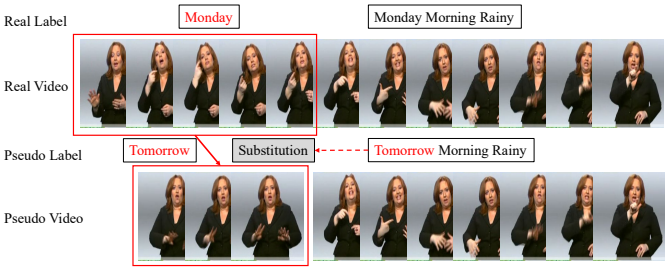
In this section, we first briefly introduce our whole framework. Then we elaborate the cross modality augmentation learning paradigm, including prior-guided cross modality editing, detailed framework architecture and cross modality optimization strategy.

Overview. As illustrated in Figure 2, during training, we first perform prior-guided cross modality editing, *i.e.,* substitution, deletion and insertion operations on the original data, to generate pseudo video-text pair. To incorporate prior during editing, we guide the text editing process by modeling the joint text distribution, while ensuring the pseudo video transition smoothness by the first-order pose derivative. Given the real and pseudo videos, they are first fed into the same visual encoder for high-dimensional semantic representations. Then the sequential learning module models the temporal dependency and performs mapping to the text sequence under the supervision of the alignment loss. Meanwhile, text modality is also mapped into the same semantic space as the video data. In this space, multiple loss terms are designed to make the framework aware of the fine-grained difference between real and pseudo video-text pairs. Notably, cross modality augmentation is only utilized during the training stage. During the inference stage, only the real sign video is fed into the visual encoder, sequential model and CTC decoding module to output the final predicted text sentence.

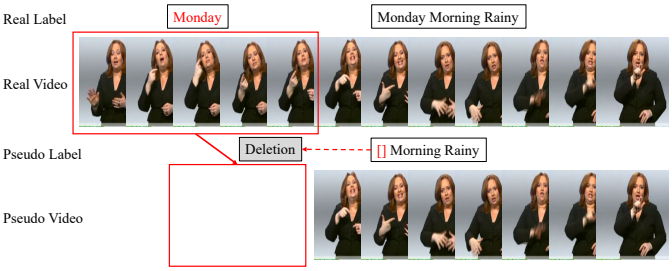
A. Prior-Guided Cross Modality Editing

Following the basic operations during WER calculation, *i.e.,* substitution, deletion and insertion, we edit the original video-text pair to generate a pseudo pair with prior incorporated. Figure 3 illustrates these editing operations. For the substitution operation shown in Figure 3(a), "Monday" is replaced with the word "Tomorrow", and the original sentence is modified as "Tomorrow Morning Rainy". Meanwhile, the video clip corresponding to "Monday" is replaced with the video clip with the meaning of "Tomorrow". For the deletion operation illustrated in Figure 3(b), the word "Monday" and its corresponding clip are simultaneously deleted. For the insertion operation in Figure 3(c), the "Light" word and corresponding video clip are inserted into the original video-text pair. The editing process repeats k times, and k ranges from 1 to K. With each editing described above, we obtain a pseudo video-text pair.

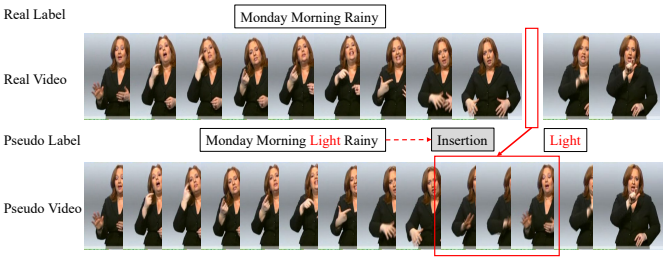
During editing, it is vital to choose a reasonable word and its corresponding video clip, such that the pseudo sentence label and pseudo video follow the same distribution as the original data, which will further benefit the cross modality



(a) Illustration of the "Substitution" operation.



(b) Illustration of the "Deletion" operation.



(c) Illustration of the "Insertion" operation.

Fig. 3. Illustration of different kinds of editing operations.

representation learning. To this end, we introduce grammar prior for pseudo sentence label generation and transition smoothness prior to pseudo video generation. The details are discussed in the following.

Text editing criteria. As shown in Figure 4, among these basic editing operations, substitution and insertion involve choosing new words, which are guided by high joint probabilities with the remaining text sequence.

To model the joint probabilities over the corpora, we get inspiration from the BERT framework and its masking strategy [48]. The gloss is mapped to the high-dimensional embedding f_g, summed with the position encoding f_p. After the summation, the embedded sequence is fed into the multi-layer Transformer encoder, whose block contains a multi-head attention model and a feed-forward network. The output of each layer retains the same size with the input as follows,

$$\begin{aligned} \mathbf{F}_0 &= \{f_g + f_p\}, \\ \widetilde{\mathbf{F}}_i &= LN(MHA(\mathbf{F}_{i-1}) + \mathbf{F}_{i-1}), \\ \mathbf{F}_i &= LN(FF(\widetilde{\mathbf{F}}_i) + \widetilde{\mathbf{F}}_i), \end{aligned} \tag{1}$$

where i indicates the i-th layer of the encoder. $LN(\cdot)$, $MHA(\cdot)$ and $FF(\cdot)$ represent the layer normalization, the multi-head self-attention and feed-forward network, respec-

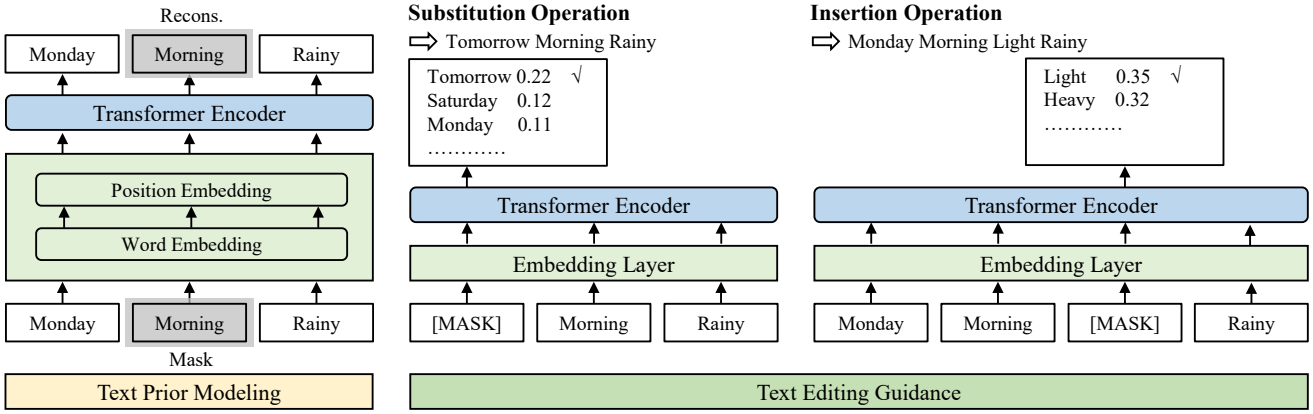


Fig. 4. Overview of the prior-guided text editing. We first model the joint text distribution via masking and reconstructing the gloss token. Then, during cross modality text editing, we guide the substitution and insertion operation by leveraging the prior incorporated into this model. For the substitution and insertion operation, we mask the original gloss token and insert a [MASK] token, respectively. Then we utilize the output gloss with the high probability at the corresponding location as the candidate.

tively. $\mathbf{F}_i$ represents the feature representation generated by the i-th layer.

The framework is trained via the masked modeling strategy on the text corpora of the training set to model its joint text distribution. Given a gloss sequence, we randomly choose 15% tokens. For the chosen token, we replace it with (1) [MASK] token 80% of the time, (2) a random token 10% of the time, and (3) the unchanged token 10% of the time. Then the framework predicts the chosen tokens by leveraging the bidirectional linguistic clues from the other unmasked tokens. Notably, this masking strategy is only adopted to model the text distribution. If we directly edit the original gloss sequence with this strategy, it may produce a biased text sequence, since [MASK] never exists in the original corpora and random replacing breaks the original grammar. The whole training objective is to maximize the log-likelihood of the correct gloss tokens g given the corrupted gloss sequence $\widetilde{g}$ as follows,

$$\max \sum_{g \in \mathcal{D}} \mathbb{E}_{\mathcal{M}} \left[\sum_{i \in \mathcal{M}} \log p\left(g_i \mid \widetilde{g}^{\mathcal{M}}\right) \right], \tag{2}$$

where $\mathcal{D}$ is the training corpus, $\mathcal{M}$ denotes the chosen token locations and this training objective is equivalent to maximizing the joint text distribution according to [50], [51]. This optimization totally lasts for 50 epochs and we utilize the Adam optimizer with the initial learning rate as 1e-3.

Finally, we deploy this model to recommend the words during cross modality editing. For the substitution and insertion operations, we replace the chosen word with the [MASK] token and insert the [MASK] token at the desired location of the input sequence, respectively. Then we randomly pick one word among Top-N_{tex} high probability candidates at the corresponding output location.

Video editing criteria. Consistent with the text editing operations, we perform the same operations on the original video data. To this end, we build the word-clip memory bank according to the alignment extracted in the refinement stage and consider pseudo video transition consistency after editing. In this work, we utilize MMPose [52] to extract the 2D upper body pose and build a criterion based on it.

Since signers have different body shapes and recording conditions, they have intrinsic spatial arm displacements among different videos. If we directly utilize the absolute metric as the criterion, *e.g.* spatial displacement, these intrinsic displacements will disturb the evaluation and make video clip selection less convincing. Therefore, we resort to the relative metric, which can somewhat filter these irrelevant factors from recording conditions and individual differences. The velocity of the arm somewhat filters the influence of intrinsic spatial displacements, but it may still be disturbed by individual differences. Besides, the velocity itself is jittering among the video. Therefore, it may be hard to build a metric based on the velocity. In this work, we resort to the inner should angle difference as the metric, which is calculated as follows,

$$\begin{gathered}
\mathcal{S} = \sum_{s} \delta(s), \ \delta = \sum_{i=1}^{2} \sum_{t} \beta_i(t) \\
\beta_i(t) = |\theta_i(t) - \theta_i(t-1) + 360 * K_i|, \ s.t. \ 0 \leq \beta_i(t) \leq 180, \\
tan\theta_1 = \frac{\overline{BA} \times \overline{BC}}{\langle \overline{BA}, \overline{BC} \rangle}, tan\theta_2 = \frac{\overline{CB} \times \overline{CD}}{\langle \overline{CB}, \overline{CD} \rangle},
\end{gathered} \tag{3}$$

where $\mathcal{S}$ is the final score. s represents the side (left or right). δ is the score for one side and $\beta_i(t)$ denotes one inner shoulder angle of one side at time t (the timestamp at front or back editing linkage) and it is measured in the 2D plane. K_i is the integer. A, B, C and D represent the hip, shoulder, elbow and wrist points, respectively. The lower $\mathcal{S}$ indicates better video transition consistency. During the substitution or insertion operation, we randomly pick one of the corresponding clips among the Top-N_{vid} lowest $\mathcal{S}$ to edit the video data.

B. Framework Architecture

To perform cross modality augmentation learning, we carefully design our framework, as illustrated in Fig. 2, which consists of a visual encoder, a sequential model and a text encoder.

Visual encoder. It maps the raw RGB video to the semantic latent space. The raw RGB video goes through a spatial 2D-CNN and a temporal encoder sequentially for spatial-temporal representation. In our implementation, we keep the

same configuration with the commonly used baseline [3]. Specifically, the 2D-CNN is selected as GoogleNet [8]. The temporal encoder is implemented as the stack of the temporal convolution and pooling layers. The kernel size of temporal convolution and max pooling layers are set to 5 and 2, respectively, and their strides are all set to 1. Under these settings, the temporal encoder outputs the one-quarter temporal length of the input representations, with its receptive field as 16. The function of the visual encoder can be formulated as follows,

$$\mathbf{F} = E_v(\mathbf{V}), \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{C_1 \times T_1 \times H \times W}$ and $\mathbf{F} \in \mathbb{R}^{C_2 \times T_1/4}$.

Sequential model. The sequential model further captures the temporal dependency among the latent semantics and learns the correspondence with the gloss sequence. Specifically, the latent semantics are fed into BLSTM to capture the bidirectional feature representation. The sequential model is formulated as follows,

$$\mathbf{F}_v = E_b(\mathbf{F}), \tag{5}$$

where $E_b(\cdot)$ denotes the BLSTM and the output is $\mathbf{F}_v \in \mathbb{R}^{C_3 \times T_1/4}$.

Text encoder. The text modality also needs to be mapped to the same semantic latent space as the video modality. Specifically, we utilize a two-layer BLSTM as the text encoder, which is is formulated as follows,

$$\mathbf{F}_l = E_t(\boldsymbol{s}), \tag{6}$$

where $\mathbf{F}_l \in \mathbb{R}^{C_3 \times T_2}$.

C. Cross Modality Augmentation Optimization

To perform cross modality augmentation learning in a unified latent space, we utilize three kinds of loss terms during optimization, *i.e.,* alignment loss $\mathcal{L}_A$, semantic correspondence loss $\mathcal{L}_S$ and real-pseudo discriminative loss $\mathcal{L}_D$.

Alignment loss $\mathcal{L}_A$. The task of continuous sign language recognition aims to learn the correspondence between the video and text modality. We utilize the connectionist temporal classification (CTC) as the alignment constraint, which is designed to deal with two unsegmented sequences without precise alignment. It introduces a blank label for the cases of transition or silence without precise meaning. Typically, there exists many-to-one mapping from multiple input sequences to one target, which is achieved by removing repetition or blank labels. For each mapping path, CTC assumes the time independence and the probability of each path is calculated as follows,

$$p(\pi|\mathbf{V}) = \prod_{t=1}^{T} p(\pi_t|\mathbf{V}), \tag{7}$$

where π_t is the label at the timestamp t, and T is the input representation duration. Then the conditional probability of the target sequence $\boldsymbol{s}$ is calculated by summing that of all possible mapping paths, which is calculated as follows,

$$p(\boldsymbol{s}|\boldsymbol{V}) = \sum_{\pi \in \mathcal{B}^{-1}(\boldsymbol{s})} p(\pi|\boldsymbol{V}), \tag{8}$$

where $\mathcal{B}^{-1}$ is the inverse mapping of $\mathcal{B}$. The CTC loss is defined by the negative log probability of $p(\boldsymbol{s}|\boldsymbol{V})$ as follows,

$$\mathcal{L}_{\text{CTC}} = -\ln p(\boldsymbol{s}|\boldsymbol{V}). \tag{9}$$

In our work, we utilize the CTC loss to supervise both real and pseudo video streams, which is formulated the alignment loss as follows,

$$\mathcal{L}_A = \mathcal{L}^r_{\text{CTC}} + \mathcal{L}^p_{\text{CTC}}, \tag{10}$$

where $\mathcal{L}^r_{\text{CTC}}$ and $\mathcal{L}^p_{\text{CTC}}$ are the CTC loss for the real and pseudo video stream, respectively. This alignment loss aims to maximize the probabilities of all mapping paths between the input video and sign gloss sequence. During the inference stage, CTC obtains a set of predicted sentence as candidates using beam search and choose the one with the highest decoding probability as the final prediction.

Semantic correspondence loss $\mathcal{L}_S$. As illustrated in Figure 2, given the real and pseudo video-text pair, we denote the feature representations of real video, real text label, pseudo video, and pseudo text label as $\boldsymbol{f}^r_v$, $\boldsymbol{f}^r_l$, $\boldsymbol{f}^p_v$ and $\boldsymbol{f}^p_l$, respectively. One key issue before cross modality representation learning is building the distance metric between these feature representations from different sources and modalities.

Considering the temporal length variance between them, we utilize the dynamic time warping (DTW) as the distance indicator. DTW utilizes the dynamic programming technique to efficiently find the best alignment between two variable sequences with the lowest cost. Denote the cost between $\boldsymbol{f}^r_v$ at the timestamp i and $\boldsymbol{f}^r_l$ at the timestamp j as $d(i,j)$, DTW gradually calculates the distance $D_{i,j}$ of their sub-sequences, *i.e.,* $\boldsymbol{f}^r_v(1:i) = (\boldsymbol{f}^r_v(1), \boldsymbol{f}^r_v(2), \cdots, \boldsymbol{f}^r_v(i))$ and $\boldsymbol{f}^r_l(1:j) = (\boldsymbol{f}^r_l(1), \boldsymbol{f}^r_l(2), \cdots, \boldsymbol{f}^r_l(j))$, which is formulated as follows,

$$D_{i,j} = d(i,j) + \min(D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}). \tag{11}$$

Specifically, we calculate $d(i,j)$ based on the cosine distance as follows,

$$d(i,j) = 1 - \frac{\boldsymbol{f}^r_v(i) \cdot \boldsymbol{f}^r_l(j)}{||\boldsymbol{f}^r_v(i)|| \cdot ||\boldsymbol{f}^r_l(j)||}. \tag{12}$$

Since the DTW is not differentiable, we further utilize a continuous relaxation operator [53], [54] with the smoothing parameter $\gamma \geq 0$

$$\min{}^{\gamma}(a_1, \cdots, a_n) := \begin{cases} \min\limits_i a_i, & \gamma = 0. \\ -\gamma \log \sum\limits_i e^{-a_i/\gamma}, & \gamma \geq 0. \end{cases} \tag{13}$$

One important factor of cross modality representation learning is to ensure the video and text are mapped into the same latent semantic space. For the real video-text pair, their semantic distance should be as close as possible, which formulates the semantic correspondence loss as follows,

$$\mathcal{L}_S = \mathcal{D}(\boldsymbol{f}^r_v, \boldsymbol{f}^r_l) = D_{T,N}, \tag{14}$$

where T is the length of $\boldsymbol{f}^r_v$ and N is the length of $\boldsymbol{f}^r_l$.

Real-pseudo discriminative loss $\mathcal{L}_D$. This loss term aims at embedding the framework with the capability of distinguishing

TABLE I
STATISTICAL DATA ON RWTH-PHOENIX MULTI-SIGNER, SIGNER-INDEPENDENT, RWTH-PHOENIX-T, CSL AND CSL-DAILY DATASETS.

Statistics	RWTH-Phoenix						RWTH-Phoenix-T			CSL		CSL-Daily		
	Multi-Signer			Signer-Independent										
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Test	Train	Dev	Test
#signers	9	9	9	8	1	1	9	9	9	50	50	10	10	10
#frames	799,006	75,186	89,472	612,027	16,460	26,891	827,354	55,775	64,627	963,228	66,529	2,227,178	134,530	153,074
#duration (h)	8.88	0.84	0.99	6.80	0.18	0.30	9.19	0.62	0.72	10.70	0.74	20.62	1.25	1.42
#vocabulary	1,231	460	496	1,081	239	294	1,066	393	411	178	20	2,000	1,344	1,345
#videos	5,672	540	629	4,376	111	180	7,096	519	642	4,700	300	18,401	1,077	1,176

the fine-grained difference between real and pseudo video-text pairs. To this end, we divide these features into two groups, *i.e.,* real-video anchored group $(\boldsymbol{f}_v^r, \boldsymbol{f}_l^r, \boldsymbol{f}_l^p)$ and real-text anchored group $(\boldsymbol{f}_l^r, \boldsymbol{f}_v^r, \boldsymbol{f}_v^p)$. In each group, we aim to minimize the distance between the anchor and the positive sample while maximizing the distance between the anchor and the negative one. The real-video anchored loss term is formulated as follows,

$$\mathcal{L}_{D_v} = \mathcal{L}_{trip}(\boldsymbol{f}_v^r, \boldsymbol{f}_l^r, \boldsymbol{f}_l^p) = \max\left(\mathcal{D}(\boldsymbol{f}_v^r, \boldsymbol{f}_l^r) - \mathcal{D}(\boldsymbol{f}_v^r, \boldsymbol{f}_l^p) + \alpha, 0\right), \tag{15}$$

where $\mathcal{L}_{trip}(\cdot)$ represents the triplet loss [55], and α is a margin. For another group, the real text anchored discriminative loss is defined as follows,

$$\mathcal{L}_{D_l} = \mathcal{L}_{trip}(\boldsymbol{f}_l^r, \boldsymbol{f}_v^r, \boldsymbol{f}_v^p) = \max\left(\mathcal{D}(\boldsymbol{f}_l^r, \boldsymbol{f}_v^r) - \mathcal{D}(\boldsymbol{f}_l^r, \boldsymbol{f}_v^p) + \alpha, 0\right). \tag{16}$$

Notably $\mathcal{D}(\cdot)$ in Equation 15 and 16 follow the the same calculation process as the Equation 14. The real-pseudo discriminative loss $\mathcal{L}_D$ is calculated by summing these two loss terms as follows,

$$\mathcal{L}_D = \mathcal{L}_{D_v} + \mathcal{L}_{D_l}. \tag{17}$$

The final optimization loss is the weighted summation of the aforementioned loss terms as follows,

$$\mathcal{L} = \lambda\mathcal{L}_A + (1-\lambda)(\mathcal{L}_D + \mathcal{L}_S), \tag{18}$$

where λ indicates the weighting factor. Since $\mathcal{L}_D$ and $\mathcal{L}_S$ have the same distance metrics, we group them together and perform weighted summation with $\mathcal{L}_A$.

IV. EXPERIMENTS

In this section, we first introduce the detailed experimental setup, which includes datasets, evaluation protocol and implementation details. Then we perform an ablation study to demonstrate the effectiveness of each part in our framework. Finally, we make comparison with state-of-the-art methods.

A. Experiment Setup

Datasets. We perform extensive experiments on popular benchmark datasets, *i.e.,* RWTH-Phoenix (multi-signer setting) [15], RWTH-Phoenix signer-independent (signer-independent setting), CSL [21], RWTH-Phoenix-T [56] and CSL-Daily [57]. **RWTH-Phoenix** dataset obtains sources from the public weather broadcast, which is recorded by a monocular camera at 25 frames per second (fps) with the resolution of 210 $\times$ 260. This dataset contains two settings, *i.e.,* multi-signer and signer-independent. Each setting divides the data into 3 independent sets, for training, validation and testing. The multi-signer setting contains totally 6,841 sentences with 9 different signers appearing across all sets. The signer-independent setting chooses 8 signers for training and leaves 1 signer for evaluation. **RWTH-Phoenix-T** is an extended version of the RWTH-Phoenix dataset, which has no overlap with it. It contains two-stage annotations for different tasks, *i.e.,* sign gloss annotation for continuous SLR and translation annotation for sign language translation, respectively. **CSL** is a Chinese continuous SLR dataset containing 100 sentences performed by 50 signers. It is divided under the unseen sentence setting, *i.e.,* the sentences appearing in the testing set do not appear in the training set. **CSL-Daily** mainly revolves around the daily life of the deaf community, which is the current largest Chinese sign language dataset. Similar to RWTH-Phoenix-T, it also contains two-stage annotations for two tasks. For RWTH-Phoenix-T and CSL-Daily, we only utilize its sign gloss annotation for continuous SLR. The detailed statistics are illustrated in Table I.

TABLE II
EFFECTS OF THE NUMBER K OF MAXIMUM EDITING OPERATIONS ON RWTH-PHOENIX MULTI-SIGNER DATASET (THE LOWER THE BETTER).

K	1	2	3	4	5	6
del	7.7	6.5	5.9	7.4	7.5	7.7
ins	2.8	3.2	3.2	2.8	2.6	2.7
WER	20.7	20.6	20.2	20.9	20.7	21.0

Evaluation protocol. We utilize the word error rate (WER) as the main evaluation metric. It is defined by the least operations, *i.e.,* substitution, deletion and insertion, to transform the predicted sentence to the reference one as follows,

$$WER = \frac{n_i + n_d + n_s}{L}, \tag{19}$$

where n_i, n_d, and n_s are the number of operations for insertion, deletion, and substitution, respectively. L denotes the length of the reference sequence. Besides, following the previous works [4], [36], we provide additional metrics on CSL dataset. Additional metrics include Acc-w (the ratio of correct words to the reference words) and some metrics from Natural Language Processing (NLP), including BLEU [58], METEOR [59], CIDEr [60] and ROUGE-L [61].

Implementation details. In this work, our framework first adopts the clip-level alignment from the baseline method [3]. Given this alignment, we first utilize the clip-label pairs for

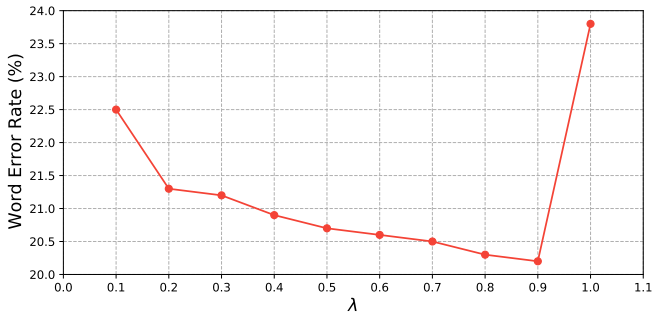


Fig. 5. Effects of different hyper parameter λ in Equation (18) on RWTH-Phoenix multi-signer dataset (the lower the better).

TABLE III
EFFECTS OF THE NUMBER N_{vid} OF CANDIDATES FOR TEXT EDITING ON RWTH-PHOENIX MULTI-SIGNER DATASET (THE LOWER THE BETTER). WE RANDOMLY CHOOSE ONE OF THEM DURING VIDEO EDITING.

N_{vid}	1	3	5	7	9
del	8.1	7.1	8.1	7.5	7.5
ins	2.4	3.0	2.4	2.7	2.8
WER	21.2	20.9	21.0	21.1	21.3

classification to pre-train the visual encoder. We add a fully-connected layer on top of the visual encoder and supervise it with the cross-entropy loss. The input length is set to 16 and stochastic gradient descent (SGD) is set as the optimizer. The training lasts 40 epochs and the initial learning rate is set to 5e-3 with 10x reduction when loss saturates. The batch size and weight decay are set to 32 and 1e-4, respectively. Adopted from the pre-trained parameter of the visual encoder, our framework trains in an end-to-end method, which is supervised by the loss in Equation 18. The text encoder contains a two-layer BLSTM, whose hidden size is set to 1024. γ and α are set to 0.01 and 10, respectively. We utilize Adam to optimize the whole framework, with the learning rate and batch size set as 5e-3 and 3, respectively. Our whole framework is implemented on the PyTorch framework and all the experiments are performed on NVIDIA RTX 3090.

Besides, we utilize data augmentation to relieve over-fitting. During training, we perform augmentation on both the spatial and temporal dimensions of the real and pseudo videos. For spatial augmentation, the video is randomly cropped at the same spatial location along the time dimension, with the resolution of 224 $\times$ 224. Then the whole video is randomly flipped horizontally with the probability of 0.5. For temporal augmentation, we randomly drop 20% of the total frames. For the testing stage, the video is center cropped with the resolution of 224 $\times$ 224. All the frames are fed into the framework.

B. Ablation Study

In this subsection, we study the impact of hyper parameters (λ, K, N_{vid} and N_{tex}), pseudo data modality and prior-guided editing. The ablation experiments are performed on

TABLE IV
EFFECTS OF THE NUMBER N_{tex} OF CANDIDATES FOR TEXT EDITING ON RWTH-PHOENIX MULTI-SIGNER DATASET (THE LOWER THE BETTER). WE RANDOMLY CHOOSE ONE OF THEM DURING TEXT EDITING.

N_{tex}	1	5	10	20	30
del	7.1	8.4	7.4	5.9	8.0
ins	3.0	2.8	2.9	3.2	2.5
WER	20.9	21.2	20.8	20.2	20.9

TABLE V
EFFECTS OF CROSS MODALITY EDITING SETTINGS ON RWTH-PHOENIX MULTI-SIGNER DATASET (THE LOWER THE BETTER). WE DEMONSTRATE THE EFFECTIVENESS OF DIFFERENT PSEUDO DATA MODALITIES AND PRIOR-GUIDED EDITING. THE FIRST LINE CORRESPONDS TO THE BASELINE METHOD. "PSEUDO DATA" REPRESENTS THE PSEUDO DATA MODALITY UTILIZED IN THE FRAMEWORK. "PRIOR" DENOTES GUIDING THE CROSS MODALITY EDITING WITH PRIOR. NOTE THAT PRIOR-GUIDED EDITING STRATEGY IS BINDED WITH PSEUDO DATA MODALITY AND WE DEMONSTRATE THE EFFECTIVENESS OF PRIOR WHEN THE CORRESPONDING DATA MODALITY IS UTILIZED.

Pseudo Data		Prior		Dev	
Video	Text	Video	Text	del / ins	WER
				7.8 / 3.5	23.8
✓				7.7 / 3.0	22.0
✓		✓		8.4 / 2.5	21.4
	✓			8.1 / 2.8	21.9
	✓		✓	8.1 / 2.6	20.8
✓	✓			7.3 / 2.7	21.3
✓	✓	✓		7.7 / 2.6	20.8
✓	✓		✓	6.8 / 2.9	20.5
✓	✓	✓	✓	5.9 / 3.2	**20.2**

RWTH-Phoenix multi-signer dataset and we utilize the WER performance on the dev set as the indicator.

Impact of loss weighting parameter λ. As illustrated in Figure 5, we study the impact of the loss weighting factor λ. When λ grows, the WER is gradually decreased to the lowest value and then bounces back. The best WER performance is achieved when λ is equal to 0.9. Notably, there exists no result when λ is equal to 0, since the alignment loss is needed for continuous SLR. In the following, we set the default λ parameter as 0.9 unless stated.

Impact of maximum editing operations K. We demonstrate the impact of maximum editing operations in Table II. "K" indicates the maximum number of operations, *i.e.,* the actual editing number ranges from 1 to K. It can be observed that the best WER performance is obtained when K is equal to 3. This is partially attributed to the fact that a relatively small number of editing operations enforce the framework to capture fine-grained differences between real and pseudo data.

Impact of N_{vid} and N_{tex}. As shown in Table III, we first demonstrate the effects of N_{vid}. As mentioned in Section III-A, we randomly choose one word among Top-N_{vid} high probability candidates. It can be observed that it achieves the best WER performance when N_{vid} is equal to 3. The above experiment demonstrates the effects of N_{vid} when N_{tex} is set to 1. We further demonstrate the effects of N_{tex} in Table IV. The WER performance achieves the best when N_{tex} is equal to 20.

Impact of pseudo data modality. As shown in Table V, we

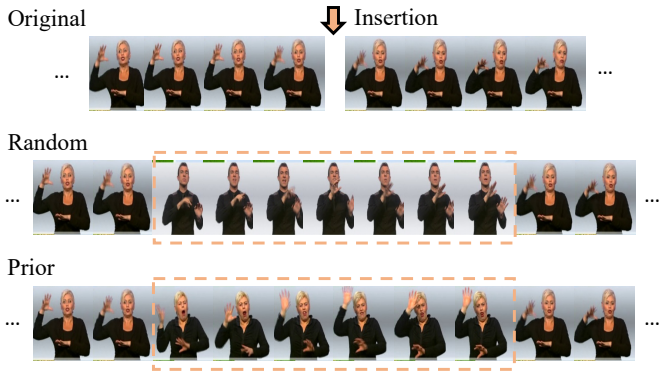
Fig. 6. Qualitative illustration of the effectiveness of prior-guided cross modality editing. We demonstrate the prediction of one sample on the dev set of RWTH-Phoenix multi-signer dataset. The predictions are produced by the baseline, the method without and with prior-guided editing ("Baseline", "CMA" and "Ours"). The first row shows the raw input RGB video. The medium three rows exhibit the sign gloss with the maximum probability of each time step. The bottom four rows demonstrate the final predicted sentence. Red symbols represent false prediction. "D" and "S" denote deletion and substitution, respectively.

TABLE VI
QUALITATIVE RESULTS OF GENERATED PSEUDO TEXT DATA. "ORIGINAL", "RANDOM" AND "PRIOR" DENOTE THE ORIGINAL TEXT, GENERATED TEXT UNDER RANDOM AND PRIOR-GUIDED EDITING, RESPECTIVELY. WE EXHIBIT THE ENGLISH TRANSLATION CORRESPONDING TO EACH GERMAN GLOSS FOR EACH TEXT SENTENCE. BLUE HIGHLIGHTS THE DIFFERENCE WITH THE ORIGINAL GLOSS SEQUENCE.

Original:	_on_ mittwoch sonne wechselhaft regen nord gewitter schwer sturm
	(on Wednesday sunny change rainy north thunderstorm heavy storm)
Random:	_on_ mittwoch sonne wechselhaft **fahren** nord gewitter schwer sturm
	(on Wednesday sunny change **drive** north thunderstorm heavy storm.)
Prior:	_on_ mittwoch sonne wechselhaft **wolke** nord gewitter schwer sturm
	(on Wednesday sunny change **cloudy** north thunderstorm heavy storm.)
Original:	morgen loc-region sechszehn bis zwanzig grad warm
	(tomorrow region sixteen to twenty degree warm)
Random:	morgen **stark** loc-region sechszehn bis zwanzig grad warm
	(tomorrow **strong** region sixteen to twenty degree warm.)
Prior:	morgen **nord** loc-region sechszehn bis zwanzig grad warm
	(tomorrow **north** region sixteen to twenty degree warm.)
Original:	morgen kuehl achtzehn bis drei zwanzig grad
	(tomorrow cool eighteen to three twenty degree)
Random:	morgen **frei** kuehl **drehen** bis drei zwanzig grad
	(tomorrow **free** cool **rotate** to three twenty degree)
Prior:	morgen **region** kuehl **fuenfzehn** bis drei zwanzig grad
	(tomorrow **region** cool **fifteen** to three twenty degree)

demonstrate the effectiveness of each pseudo data modality and prior editing. "Pseudo Data" denotes the modality we utilize during cross modality augmentation learning. It can be observed that each modality brings performance improvement, and pseudo text data brings a relatively larger gain than pseudo video. Besides, the effectiveness of each modality is complementary. When both modalities are utilized, the WER performance will further get improved.

Effectiveness of prior-guided editing. As shown in Table V, "Prior" represents that we guide the editing process with prior incorporated. Compared with random editing, the prior-incorporated scheme outperforms it under all pseudo data modality settings. It may be attributed to the fact that random editing severely breaks the original video and text distribution, which makes training biased. Besides, the recognition performance gain from the text prior is more significant than that from the video prior.



(a) Qualitative results of generated pseudo video data under the "Insertion" operation.

(b) Qualitative results of generated pseudo video data under the "Substitution" operation.

Fig. 7. Qualitative results of generated pseudo video data. "Original", "Random" and "Prior" denote the original video, generated video under random and prior-guided editing, respectively. It can be observed that our prior-guided video editing well preserves the video transition consistency during editing.

TABLE VII
EVALUATION ON RWTH-PHOENIX MULTI-SIGNER DATASET (THE LOWER THE BETTER).

Methods	Venue	Stream				Dev		Test	
		full	hand	face	pose	del / ins	WER	del / ins	WER
CMLLR [15]	CVIU'15		✓	✓	✓	21.8 / 3.9	55.0	20.3 / 4.5	53.0
1-Million-Hand [62]	CVPR'16		✓	✓	✓	16.3 / 4.6	47.1	15.2 / 4.6	45.1
CNN-Hybrid [63]	BMVC'16		✓			12.6 / 5.1	38.3	11.1 / 5.7	38.8
RCNN [38]	CVPR'17		✓			13.7 / 7.3	39.4	12.2 / 7.5	38.7
Re-Sign [33]	CVPR'17	✓				-	27.1	-	26.8
SubUNets [5]	ICCV'17	✓	✓			14.6 / 4.0	40.8	14.3 / 4.0	40.7
Hybrid CNN-HMM [35]	IJCV'18		✓			-	31.6	-	32.5
CTF [64]	MM'18	✓				12.8 / 5.2	37.9	11.9 / 5.6	37.8
Dilated [30]	IJCAI'18	✓				8.3 / 4.8	38.0	7.6 / 4.8	37.3
IAN [4]	CVPR'19	✓				12.9 / 2.6	37.1	13.0 / 2.5	36.7
CNN-LSTM-HMM [32]	TPAMI'19	✓	✓			-	26.0	-	26.0
DNF (RGB) [3]	TMM'19	✓				7.8 / 3.5	23.8	7.8 / 3.4	24.4
SL-Trans. [65]	CVPR'20	✓				5.8 / 4.7	23.1	5.4 / 4.6	24.2
SFLM [66]	ECCV'20	✓				10.3 / 4.1	24.9	10.4 / 3.6	25.3
FCN [28]	ECCV'20	✓				-	23.7	-	23.9
PiSLTRc-R [67]	TMM'21	✓				8.1 / 3.4	23.4	7.6 / 3.3	23.2
STMC [6]	TMM'21	✓	✓	✓	✓	7.7 / 3.4	21.7	7.4 / 2.6	20.7
VAC [39]	ICCV'21	✓				7.9 / 2.5	21.2	8.4 / 2.6	22.3
CMA [14]	MM'20	✓				7.3 / 2.7	21.3	7.3 / 2.4	21.9
Ours	This work	✓				5.9 / 3.2	**20.2**	6.0 / 2.8	**20.0**

As demonstrated in Table VI, we illustrate some generated text samples under different editing schemes (random and prior-guided editing). The three parts exhibit the generated sentence under the substitution, insertion and multiple operations, respectively. It can be observed that random editing contains unreasonable operations, *e.g.* changing the part of speech and adding some irrelevant words, *etc.* As a result, it breaks the original syntax and becomes a sample out of distribution. In contrast, our prior-aware editing is able to perform reasonable substitution ("rainy" to "cloudy" and "eighteen" to "fifteen") and insertion (adding "north" before "region"). These operations enforce the framework aware of these fine-grained cues, which are crucial for accurate recognition.

Besides, qualitative results of generated pseudo video data are demonstrated in Figure 7. During pseudo video editing, we do not put the constraint that the video clip is selected from the video with the same background and human. It can be observed that our utilized inner shoulder angle metric well preserves the video transition consistency after editing. Since we adopt the off-the-shelf detector to extract the 2D upper body pose, noise is inevitably introduced for all metrics based on the pose. Other factors are also desirable to explore to serve as more effective metrics for pseudo video clip selection in the future.

To further demonstrate the effectiveness of prior-guided editing on recognition, we qualitatively illustrate one sample in Figure 6. In this figure, the first line denotes the raw RGB video. The middle part exhibits the decoding result at each time step. The gray and red box denotes the blank and false prediction, respectively. The bottom part represents the final predicted sentence. We exhibit the result of the baseline, and our framework with random and prior-guided cross modality editing. It can be observed that the incorporated prior during training is beneficial for improving performance on the testing inference stage. The WER of the shown sample is gradually improved with the addition of cross modality augmentation learning and prior guidance. To some extent, the prior makes the framework better capture the discriminative cues, which improves the recognition performance.

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

We conduct extensive experiments, make comparisons with state-of-the-art methods and perform analysis on existing benchmarks, *i.e.,* RWTH-Phoenix multi-signer and signer-independent, RWTH-Phoenix-T, CSL and CSL-Daily datasets. "CMA" and "Ours" denote the results of our previous and current extended work, respectively.

Evaluation on RWTH-Phoenix multi-signer dataset. As shown in Table VII, we perform comparison on the most popular continuous SLR benchmark, *i.e.,* RWTH-Phoenix multi-signer dataset. CMLLR [15] is one of the representative methods based on hand-crafted features and traditional HMM models. SubUNets [5] jointly solves alignment and recognition via supervising the CNN-BLSTM framework by the CTC loss. DNF [3] utilizes 2DCNN-TCN-BLSTM as the backbone, in cooperation with the iterative refinement strategy. SL-Trans. [65] utilizes the Transformer to model bidirectional sequential information. Even compared with the most challenging STMC [6], our framework still achieves the new state-of-the-art performance with only one stream utilized during inference.

Evaluation on RWTH-Phoenix signer-independent dataset. We also perform comparison on the signer-independent setting of RWTH-Phoenix dataset in Table IX. This setting evaluates the generalization capability of methods over unseen signers. It can be observed that this setting is a more challenging one, since the average WER performance is 10% worse than the multi-signer setting. Our method achieves new state-of-the-art

TABLE VIII
EVALUATION ON CSL DATASET. (↑ INDICATES THE HIGHER THE BETTER, WHILE ↓ INDICATES THE LOWER THE BETTER.)

Methods	Venue	Acc-w ↑	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	CIDEr ↑	ROUGE-L ↑	METEOR ↑	WER ↓
ELM [68]	ICPR'14	0.175	0.376	0.381	0.142	0.028	0.120	0.388	0.987
LSTM & CTC [10], [13]	-	0.332	0.343	0.124	0.039	0.241	0.362	0.111	0.757
S2VT (3-layer) [69]	ICCV'15	0.461	0.475	0.265	0.145	0.477	0.465	0.186	0.652
HLSTM-attn [36]	AAAI'18	0.506	0.508	0.330	0.207	0.605	0.503	0.205	0.641
HRF-Fusion [37]	TIP'19	0.445	0.450	0.238	0.127	0.398	0.449	0.171	0.672
IAN [4]	CVPR'19	0.670	0.724	-	-	**3.946**	0.716	0.383	0.327
CMA [14]	MM'20	0.747	**0.784**	0.635	0.514	3.006	0.782	0.390	0.245
Ours	This work	**0.755**	0.780	**0.664**	**0.571**	3.512	**0.815**	**0.406**	**0.225**

TABLE IX
EVALUATION ON RWTH-PHOENIX SIGNER-INDEPENDENT DATASET (THE LOWER THE BETTER).

Methods	Venue	Dev		Test	
		del / ins	WER	del / ins	WER
Re-Sign [33]	CVPR'17	-	45.1	-	44.1
DNF [3]	TMM'19	9.2 / 4.3	36.0	9.5 / 4.6	35.7
CMA [14]	MM'20	11.1 / 2.4	34.8	11.4 / 3.3	34.3
Ours	This work	10.6 / 3.1	**31.4**	9.8 / 3.5	**30.4**

TABLE X
EVALUATION ON RWTH-PHOENIX-T DATASET (THE LOWER THE BETTER). (V: VIDEO, M: MOUTH, F:FACE, H:HAND, P:POSE)

Methods	Venue	Dev	Test
CNN-LSTM-HMM (v) [32]	TPAMI'19	24.5	26.5
SL-Trans. (v) [65]	CVPR'20	24.9	24.6
SFLM (v) [66]	ECCV'20	25.1	26.1
FCN (v) [28]	ECCV'20	23.3	25.1
PiSLTRc-R (v) [67]	TMM'21	21.8	22.9
CNN-LSTM-HMM (v+m) [32]	TPAMI'19	24.5	25.4
CNN-LSTM-HMM (v+m+h) [32]	TPAMI'19	22.1	24.1
STMC (v+h+f+p) [6]	TMM'21	19.6	21.0
CMA (v) [14]	MM'20	20.3	21.2
Ours (v)	This work	**18.8**	**20.0**

TABLE XI
EVALUATION ON CSL-DAILY DATASET (THE LOWER THE BETTER).

Methods	Venue	Dev		Test	
		del / ins	WER	del / ins	WER
SubUNets [5]	ICCV'17	14.8 / 3.0	41.4	14.6 / 2.8	41.0
LS-HAN [21]	AAAI'18	14.6 / 5.7	39.0	14.8 / 5.0	39.4
DNF [3]	TMM'19	12.8 / 3.3	32.8	12.5 / 2.7	32.4
SL-Trans. [65]	CVPR'20	10.3 / 4.4	33.1	9.6 / 4.1	32.0
FCN [28]	ECCV'20	12.8 / 4.0	33.2	12.6 / 3.7	32.5
SignBT [57]	CVPR'21	13.9 / 3.4	33.6	13.5 / 3.0	33.1
CMA [14]	MM'20	14.3 / 2.4	30.5	13.4 / 2.4	29.9
Ours	This work	13.0 / 2.6	**29.4**	12.1 / 2.6	**28.7**

performance, with a larger performance gain (3.4% and 3.9% on the dev and test set) over our previous work.

Evaluation on RWTH-Phoenix-T dataset. As demonstrated in Table X, we make comparisons on RWTH-Phoenix-T dataset. This dataset further introduces the spoken German annotation corresponding to the sign gloss annotation. CNN-LSTM-HMM [32] utilizes the spoken German annotation, which serves as auxiliary information to infer weak mouth shape labels. Besides, it also leverages multi-cue information, *e.g.* mouth and hand, to further enhance the recognition performance. STMC [6] utilizes different multi-cue streams to boost the performance. Our method only utilizes one full video stream during inference and surpasses STMC [6] with a notable gain, achieving 18.8% and 20.0% on the dev and test set, respectively.

Evaluation on CSL dataset. Extensive experimental results on CSL dataset are demonstrated in Table VIII. We compare our framework with other competitive methods, *e.g.* HRF-Fusion, HLSTM-attn, and IAN. This dataset introduces additional metrics to evaluate the semantic correspondence. HLSTM-attn [36] treats it as a translation task and utilizes the encoder-decoder architecture with the temporal attention mechanism for performance boosting. IAN [4] also utilizes the encoder-decoder framework for sequence learning, jointly with the iterative refinement strategy. Our method still achieves state-of-the-art performance on most metrics.

Evaluation on CSL-Daily dataset. In Table XI, we make comparison with other state-of-the-art methods on CSL-Daily dataset, which is the current largest Chinese sign language corpora. The CSL-Daily dataset covers a wide range of topics in daily life, involving family, school and shopping, *etc.* On this dataset, our framework surpasses all methods with a notable gain, achieving 29.4% and 28.7% on the dev and test set, respectively.

D. Analysis & Future Work

The incorporated prior guides the cross modality editing process, which generates the pseudo video-text pair with more realness. Compared with the randomly generated one, our prior-guided produced pseudo data better matches the original data distribution, *i.e.,* conforming to the text grammar and video transition consistency. In this way, it will serve as a more genuine hard sample for discriminative representation learning. During optimization, the objective enforces the framework to distinguish more fine-grained differences between the real and pseudo data, which further improves the performance on continuous SLR.

We outline the future work as follows. More effective prior-guided editing methods are desirable to design. Such

improvement can focus on improving the quality of generated pseudo video-text pair.

Sign language text, *i.e.,* gloss, is substantially different from the common spoken language text, since it contains its unique grammar and lexicon. Gloss needs expert annotation and is hard to collect a large-scale one from the available open source such as the Internet for each target language. When a large-scale gloss corpus are available, it may help generate better pseudo text, which fertilizes the final recognition model.

For pseudo video editing, it is also possible to select video clips by video-text retrieval. Specifically, with our extracted gloss-video alignment, we can train a video-text retrieval model. With this model, we can retrieve the video clips corresponding to a gloss from a large amount of unlabeled long sign videos on the Internet. This can enrich our built gloss-video alignment bank, which may further enhance the model performance or robustness.

Besides, more effective objective loss terms on the cross modality augmentation learning can be explored. More related tasks involving the video-text alignment can be explored.

V. Conclusion

In this paper, we aim to tackle the inconsistency between the WER evaluation metric and CTC objective function for continuous sign language recognition. To this end, we propose a novel prior-aware cross modality augmentation learning paradigm. Following the operations during WER calculation, *i.e.,* substitution, insertion and deletion, we edit the real video-text pair to generate the pseudo pair. To ensure the generated pseudo data quality, we incorporate prior during the editing process to recommend suitable candidates by following the text grammar and video transition consistency. Given the real and pseudo video-text pairs, we jointly feed them into the same framework. We optimize the framework via three types of loss terms, *i.e.,* alignment loss, semantic correspondence loss and real-pseudo discriminative loss. These loss terms embed the framework with the discriminative capability of the fine-grained difference between the real and pseudo pair. Extensive experiments are conducted on popular large-scale benchmarks and validate the effectiveness of our framework.

VI. Broader Impact

As reported officially by the World Health Organization (WHO), there are around 466 million people with hearing loss. It is estimated this number will increase to over 900 million by 2050. The hearing loss will directly bring communication difficulties, which may lead to social frustration or some other mental issues.

Our designed technique will promote the development of automatic sign language recognition (SLR), which bridges the communication gap between the deaf and the community. Besides, it will raise social awareness for people with disabilities and encourage the equal distribution of health care and resources for all communities. Our built system is not intended for the potential privacy issue, such as surveillance on the talk using sign language. Besides, failure recognition may lead to potential misunderstanding. Thus current recognition system may still serve as an auxiliary tool for communication.

References

[1] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 12, pp. 1371–1375, 1998.

[2] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *IEEE International Conference on Multimedia & Expo (ICME)*, 2016, pp. 1–6.

[3] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 7, pp. 1880–1891, 2019.

[4] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4165–4174.

[5] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 3075–3084.

[6] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for sign language recognition and translation," *IEEE Transactions on Multimedia (TMM)*, pp. 1–13, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[14] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 1497–1505.

[15] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding (CVIU)*, vol. 141, pp. 108–125, 2015.

[16] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2961–2968.

[17] T. Pfister, J. Charles, and A. Zisserman, "Large-scale learning of sign language by watching TV (using co-occurrences)." in *British Machine Vision Conference (BMVC)*, 2013, pp. 1–11.

[18] H. Wang, X. Chai, and X. Chen, "A novel sign language recognition framework using hierarchical grassmann covariance matrix," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 11, pp. 2806–2814, 2019.

[19] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 056–12 065.

[20] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 5533–5541.

[21] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 2257–2264.

[22] H. Hu, W. Zhou, and H. Li, "Hand-model-aware sign language recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 1558–1566.

[23] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, "SignBERT: Pre-training of hand-model-aware representation for sign language recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 087–11 096.

[24] W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, "BEST: BERT pre-training for sign language recognition with coupling tokenization," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023, pp. 1–9.

[25] L. Liu, W. Zhou, W. Zhao, H. Hu, and H. Li, "Multi-modal sign language spotting by multi/one-shot learning," in *ECCV 2022 Workshops*, 2023, pp. 256–270.

[26] H. Hu, W. Wang, W. Zhou, W. Zhao, and H. Li, "Model-aware gesture-to-gesture translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 428–16 437.

[27] N. M. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, and *et al.*, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Transactions on Multimedia (TMM)*, pp. 1–14, 2021.

[28] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 697–714.

[29] H. Hu, W. Zhou, J. Pu, and H. Li, "Global-local enhancement network for nmf-aware sign language recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 3, pp. 1–19, 2021.

[30] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 885–891.

[31] H. Hu, J. Pu, W. Zhou, and H. Li, "Collaborative multilingual continuous sign language recognition: A unified framework," *IEEE Transactions on Multimedia (TMM)*, pp. 1–12, 2022.

[32] O. Koller, C. Camgoz, H. Ney, and R. 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 (TPAMI)*, vol. 42, no. 9, pp. 2306–2320, 2019.

[33] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4297–4305.

[34] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 8, pp. 1583–1597, 2016.

[35] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," *International Journal of Computer Vision (IJCV)*, vol. 126, no. 12, pp. 1311–1325, 2018.

[36] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical LSTM for sign language translation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 6845–6852.

[37] D. Guo, W. Zhou, A. Li, H. Li, and M. Wang, "Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 1575–1590, 2019.

[38] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7361–7369.

[39] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 11 542–11 551.

[40] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, 2014, pp. 103–111.

[41] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multitask learning," *IEEE Transactions on Multimedia (TMM)*, vol. 23, pp. 1–11, 2020.

[42] J. Zhang, J. Du, and L. Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia (TMM)*, vol. 21, no. 1, pp. 221–233, 2018.

[43] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[44] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," in *Findings of the Association for Computational Linguistics (ACL Findings)*, 2021, pp. 968–988.

[45] O. Melamud, J. Goldberger, and I. Dagan, "Context2Vec: Learning generic context embedding with bidirectional LSTM," in *Special Interest Group on Natural Language Learning (SIGNLL)*, 2016, pp. 51–61.

[46] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018, pp. 2227–2237.

[47] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *arxiv*, pp. 1–12, 2018.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[49] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 7059–7069, 2019.

[50] A. Wang and K. Cho, "BERT has a mouth, and it must speak: BERT as a Markov random field language model," in *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 2019, pp. 30–36.

[51] K. Goyal, C. Dyer, and T. Berg-Kirkpatrick, "Exposing the implicit energy networks behind masked language models via metropolis–hastings," in *International Conference on Learning Representations (ICLR)*, 2022.

[52] M. Contributors, "OpenMMLab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[53] M. Cuturi and M. Blondel, "Soft-DTW: A differentiable loss function for time-series," in *International Conference on Machine Learning (ICML)*, 2017, pp. 894–903.

[54] C.-Y. Chang, D.-A. Huang, Y. Sui, L. Fei-Fei, and J. C. Niebles, "D^3TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3546–3555.

[55] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*, 2015, pp. 84–92.

[56] N. Cihan Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7784–7793.

[57] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1316–1325.

[58] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[59] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[60] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[61] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *ACL workshop on Text summarization branches out*, 2004, pp. 74–81.

[62] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3793–3802.

[63] O. Koller, O. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid CNN-HMM for continuous sign language recognition," in *British Machine Vision Conference (BMVC)*, 2016.

[64] S. Wang, D. Guo, W.-g. Zhou, Z.-J. Zha, and M. Wang, "Connectionist temporal fusion for sign language translation," in *ACM International Conference on Multimedia (ACM MM)*, 2018, pp. 1483–1491.

[65] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 023–10 033.
[66] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 172–186.
[67] P. Xie, M. Zhao, and X. Hu, "Pisltrc: Position-informed sign language transformer with content-aware convolution," *IEEE Transactions on Multimedia (TMM)*, pp. 1–13, 2021.
[68] X. Chen and M. Koskela, "Using appearance-based hand features for dynamic RGB-D gesture recognition," in *International Conference on Pattern Recognition (ICPR)*, 2014, pp. 411–416.
[69] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *International Conference on Computer Vision (ICCV)*, 2015.

Hezhen Hu is currently pursuing the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering and Information Science, from the University of Science and Technology of China (USTC). His research interests include sign language understanding, self-supervised pre-training, human-centric visual understanding, and computer vision.

Junfu Pu received the B.E. degree in Electronic Information Engineering and the Ph.D degree in Information and Communication Engineering from the University of Science and Technology of China (USTC) in 2015 and 2020, respectively. He is a senior researcher at ARC Lab (Applied Research Center), Tencent. His research interests include sign language recognition/translation/generation, multimedia understanding and applications, vision-language pretraining and search.

Wengang Zhou (S'20) received the B.E. degree in electronic information engineering from Wuhan University, China, in 2006, and the Ph.D. degree in electronic engineering and information science from the University of Science and Technology of China (USTC), China, in 2011. From September 2011 to September 2013, he worked as a postdoc researcher in Computer Science Department at the University of Texas at San Antonio. He is currently a Professor at the EEIS Department, USTC.

His research interests include multimedia information retrieval, computer vision, and computer game. In those fields, he has published over 100 papers in IEEE/ACM Transactions and CCF Tier-A International Conferences. He is the winner of National Science Funds of China (NSFC) for Excellent Young Scientists. He is the recepient of the Best Paper Award for ICIMCS 2012. He received the award for the Excellent Ph.D Supervisor of Chinese Society of Image and Graphics (CSIG) in 2021, and the award for the Excellent Ph.D Supervisor of Chinese Academy of Sciences (CAS) in 2022. He won the First Class Wu-Wenjun Award for Progress in Artificial Intelligence Technology in 2021. He served as the publication chair of IEEE ICME 2021 and won 2021 ICME Outstanding Service Award. He is currently an Associate Editor and a Lead Guest Editor of IEEE Transactions on Multimeida.

Hang Fang is a master student under the joint cultivation of Anhui University and the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. His research interests include sign language recognition, action quality assessment, multimedia understanding and applications.

Houqiang Li (S'12, F'21) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively, where he is currently a Professor with the Department of Electronic Engineering and Information Science.

His research interests include image/video coding, image/video analysis, computer vision, reinforcement learning, etc.. He has authored and co-authored over 200 papers in journals and conferences. He is the winner of National Science Funds (NSFC) for Distinguished Young Scientists, the Distinguished Professor of Changjiang Scholars Program of China, and the Leading Scientist of Ten Thousand Talent Program of China. He is the associate editor (AE) of IEEE TMM, and served as the AE of IEEE TCSVT from 2010 to 2013. He served as the General Co-Chair of ICME 2021 and the TPC Co-Chair of VCIP 2010. He received the second class award of China National Award for Technological Invention in 2019, the second class award of China National Award for Natural Sciences in 2015, and the first class prize of Science and Technology Award of Anhui Province in 2012. He received the award for the Excellent Ph.D Supervisor of Chinese Academy of Sciences (CAS) for four times from 2013 to 2016. He was the recipient of the Best Paper Award for VCIP 2012, the recipient of the Best Paper Award for ICIMCS 2012, and the recipient of the Best Paper Award for ACM MUM in 2011.