IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.DOI

SignBERT: A BERT-based Deep Learning Framework for Continuous Sign Language Recognition

ZHENXING ZHOU (Student Member, IEEE), VINCENT W.L. TAM (Senior Member, IEEE), and EDMUND Y. LAM (Fellow, IEEE)
Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
Corresponding author: Zhenxing Zhou (e-mail: zxchow@connect.hku.hk).

ABSTRACT Continuous sign language recognition (CSLR) is a very challenging task in intelligent systems, since it requires to produce real-time responses while performing computationally intensive video analytics and language modeling. Previous studies mainly focus on adopting hidden Markov models or recurrent neural networks with a limited capability to model specific sign languages, and the accuracy can drop significantly when recognizing the signs performed by different signers with non-standard gestures or non-uniform speeds. In this work, we develop a deep learning framework named SignBERT, integrating the bidirectional encoder representations from transformers (BERT) with the residual neural network (ResNet), to model the underlying sign languages and extract spatial features for CSLR. We further propose a multimodal version of SignBERT, which combines the input of hand images with an intelligent feature alignment, to minimize the distance between the probability distributions of the recognition results generated by the BERT model and the hand images. Experimental results indicate that when compared to the performance of alternative approaches for CSLR, our method has better accuracy with significantly lower word error rate on three challenging continuous sign language datasets.

INDEX TERMS bidirectional encoder representations from transformers, continuous sign language recognition, deep learning, video analytics

I. INTRODUCTION

Sign language, using different hand and body gestures to convey information, is the most important means of communication among those who are hard of hearing. As reported by the World Federation of the Deaf, more than 70 million people are using different sign languages around the world [1]. However, most other people do not understand any sign language, which severely hinders deaf people from blending into the society. To facilitate communication between the deaf and normal people, researchers have paid attention to sign language recognition (SLR), which aims at recognizing the words or sentences from videos [2], [3], and sign language annotation, which focuses on temporally locating instances of signs among the sequences of continuous gestures [4], [5].

Generally speaking, there are two main branches of SLR: isolated sign language recognition, and continuous sign language recognition (CSLR). The former focuses on recognizing the individual word or phrase from the videos [2], which is similar to video classification [6]. In contrast, the latter aims at recognizing a sequence of glosses with the correct order from the videos, which is akin to video captioning [7]. The gloss here denotes the translation of a specific gesture in sign language. CSLR is more difficult than isolated SLR, because it not only contains more glosses but also is confounded by the co-articulation effect (the start of a new gloss is affected by the ending of the previous one) and non-uniformed speed [8]. In addition, in most CSLR datasets, the ground truth labels are provided without an annotated temporal boundary for each gloss in the videos. Thus, it is important to consider the temporal relations among different gestures in the videos to achieve a better recognition accuracy.

Although it is more complicated, CSLR has wider application scenarios as human beings require translation from the continuous streams of sign language during real-world communications [9], [10]. Therefore, the focus of our work is CSLR, where the videos of continuous sign language are translated into sentences in a spoken language. In what follows, unless specifically noted, SLR will be used to refer to CSLR.

The early works in SLR mostly focus on extracting hand-crafted features, including the coordinates of the hands [11], the joints of the signers [12] and the histogram of oriented gradients (HOG) [13]–[15]. Along with the success of convolutional neural network (CNN) in solving different computer vision tasks, such as action recognition [16], [17] and video recognition [18], [19], many researchers employ various 2D or 3D CNN models to extract spatial features from the videos [20]–[22]. After obtaining the spatial features, they make use of hidden Markov model (HMM) [23]–[25] or recurrent neural network (RNN) [26], [27] to consider the temporal information inside the videos and generate the corresponding translation results. Some researchers also attempt to separate the processing of temporal dependencies into two levels, namely the gloss level and the sentence level, through using two temporal models [28], [29].

These methods consider SLR as a video understanding task only, and focus on learning the low-level representations from the videos. However, SLR is also a language modeling task. Compared with HMM or RNN, the Transformer [30] and its variants, especially the bidirectional encoder representations from transformers (BERT) [31], have shown stronger capabilities in language modeling through achieving the outstanding performance in different natural language processing (NLP) benchmark datasets [31] and video-based tasks [32]. Moreover, as some of the poses in a sign language can be complicated, it is also unavoidable that the datasets are biased and contain some nonstandard signs [8], which may affect the recognition accuracy. It is therefore important for the deep learning framework in SLR to provide an effective and robust recognition for those incorrect signs. Yet, there is hardly any prior research study conducted on adopting the BERT structure and considering the issue of robustness in SLR.

To fill in this gap, we propose a robust BERT-based deep learning framework named SignBERT. In this framework, a key frame selection mechanism is firstly used to identify the critical frames in the videos and construct high-quality video clips. Then, the (3+2+1)D ResNet [2] is adopted as the feature extraction module to obtain the spatial features from the video clips. After that, the BERT structure is employed to consider the temporal relations among the spatial features as the first level sequential module, followed by a bidirectional long-short term memory (BLSTM) [33] as the second level sequential module. To improve the robustness of the SignBERT framework, a new masked pretraining method is proposed, in which a small part of the videos in the isolated dataset of the specific sign language are randomly masked for pretraining the BERT model. In addition, we develop a multimodal version SignBERT framework, which integrates the hand images as additional inputs with a feature alignment strategy to minimize the distance between the probability distributions projected from the BERT features and hand image features to increase the recognition accuracy. Moreover, a new iterative training strategy is introduced, in which the outputs from the SignBERT framework are partially masked and reused for re-training the feature extraction module and the BERT model iteratively.

The main contributions of this work are:

1) A robust BERT-based deep learning framework named SignBERT is developed, which achieves state-of-the-art performance on SLR. The BERT model in this framework is pretrained with the partially masked videos of isolated sign language to ensure its robustness when encountering atypical signs.
2) A multimodal version of the SignBERT framework is proposed, in which the hand images of the signers are used as additional inputs with a feature alignment training strategy for comparing the features generated by the BERT model and the hand images to increase the recognition accuracy.
3) A new iterative training strategy for SLR is designed to train the SignBERT framework efficiently, in which a small part of its outputs are masked and re-used for post-training the BERT model iteratively.
4) A new high-resolution Hong Kong continuous sign language dataset is collected, where there are 50 sentences performed by 6 signers with 8 repetitions. All the sentences are highly correlated with each other, so that a high proportion can be used for the unseen sentence test. This dataset will be made available to the public to facilitate further research in SLR.

The rest of this paper is organized as follows. Some related works in SLR will be discussed in Section II. In Section III, the proposed SignBERT framework will be introduced in detail. Some aspects of its implementation will be described in Section IV, including frame selection, pretraining and iterative training. In Section V, both the experimental settings and the detailed information of the public Chinese sign language dataset, RWTH-PHOENIX-Weather 2014 dataset, as well as the newly collected Hong Kong sign language dataset will be provided. In Section VI, an ablation study will be conducted to demonstrate the effectiveness of different components of the SignBERT framework. In Section VII, the performance comparisons among the proposed SignBERT framework and other state-of-the-art methods in different continuous sign language datasets will be presented. Lastly, concluding remarks and future directions will be considered in Section VIII.

II. RELATED WORK

Generally, most of the existing frameworks proposed for SLR contain two key components: a feature extraction module to extract visual features from the raw videos, and a sequential module to consider the temporal dependencies among the visual features. In addition, many training strategies are also invented for improving the recognition accuracy in SLR.

A. THE FEATURE EXTRACTION MODULE

The feature extraction module is one of the most fundamental and important parts in SLR. Before the popularity of deep neural network, researchers focused on extracting hand-crafted features from the frames and videos, including HOG [13]–[15], hand motion trajectories [34], [35] and body joint coordinates [12]. In recent decades, convolutional neural network (CNN), including both 2D and 3D CNN, has gradually become the most common choice for feature extraction [20]–[22], [36]. Various types of CNN structures have been experimented in SLR, including 3D ResNet [37] and 3D Inception [26]. To extract more meaningful features, some researchers also use modified 3D CNN, such as 2D+1D CNN [26], [38] and (3+2+1)D ResNet [2] in SLR. Except for extracting features from color images, other information streams are also utilized in SLR. For example, Aly et al. [39] extracted features from depth images, and Sarhan et al. [40] integrated optical flow as an additional input in their recognition model.

B. THE SEQUENTIAL MODULE

The objective of the sequential module is to consider the temporal dependencies among the extracted visual representations. HMM is the most traditional choice for the sequential module in SLR [23]–[25]. Some research works also use dynamic time warping (DTW) [41] or support vector machine (SVM) [42], [43] for comparing the extracted visual representations. In recent years, due to the rapid development of RNN in natural language processing (NLP), many researchers attempt to apply RNN in SLR [26], [27], [44]. To model the underlying sign language more precisely, different types of encoder-decoder networks, such as the Transformer model, are also employed [45]–[49]. In addition, many new ideas are introduced in SLR. For instance, Grassmann Covariance Matrix (GCM) was proposed by Wang et al. [50] for sign description, which was further extended to hierarchical GCM by Wang et al. [51] for better performance.

C. THE TRAINING STRATEGY

Many training strategies are developed in SLR to improve the recognition accuracy. For example, Cui et al. [26] proposed a new iterative training strategy to fully optimize the feature extract module, in which they used the final labels generated by the sequential module as the pseudo-labels for video clips and re-trained feature extraction module with these pseudo labels, similar to the expectation-maximization (EM) algorithm. Pu et al. [52] introduced a data augmentation process for SLR, in which they created a lot of pseudo text-video pairs for training the model with multiple loss terms. Moreover, extracting key frames or clips from the videos [53]–[55] is also a popular training strategy for removing the noisy frames and improving the recognition accuracy. For instance, Guo et al. [56] adopted clip summarization method for selecting the key video clips, and Liu et al. [57] used the hand coordinates for sampling the key frames.

III. THE PROPOSED SIGNBERT FRAMEWORK

The structure of the SignBERT framework is illustrated in Figure 1. In this framework, the key frames are firstly selected according to their hand height, hand movements and frame blurriness levels for constructing high quality video clips. And then, (3+2+1)D ResNet [2] is selected as the basic feature extraction module, which has a similar structure with 3D ResNet but replaces the 3D operations in the stem block and the first ResNet block with the (2+1)D operations [58] as demonstrated in Figure 2. After that, different from previous works, the BERT model is adopted as the first sequential module followed by a BLSTM layer as the second sequential module. To improve the robustness of the SignBERT framework against the nonstandard signs, the videos from the corresponding isolated sign language dataset are partially masked for pretraining the feature extraction module and the BERT model.

In addition, we further propose a multimodal version of the SignBERT framework, in which the dominant hand image in each key frame is extracted and then processed by the conventional ResNet-BLSTM structure. To fully train the feature extraction module and the BERT structure, a feature alignment strategy is introduced for minimizing the distance between the probability distributions generated by the features from the BERT structure and the hand images. Moreover, a new iterative training strategy is devised for the SignBERT framework, in which a small part of the final sentences produced by the SignBERT framework are masked and used for re-training the feature extraction module and the BERT structure.

In the rest of this section, the proposed SignBERT framework will be discussed in detail.

A. THE MODEL FORMULATION

The model formulation of the SignBERT framework can be explained as follows. Given a sign language video, the proposed SignBERT framework firstly samples T frames according to the key frame selection mechanism (which will be introduced in the next section). Assuming $x^{T\times C\times H\times W}$ to be the selected frames of the video, in which C is the color channel size while H and W represent the heights and widths of the images, a sliding window with size L and stride S will then be applied to convert the selected frames into F video clips $X^{F\times L\times C\times H\times W}$, in which F can be calculated as

$$F = \frac{T-L}{S} + 1. \tag{1}$$

The aforementioned (3+2+1)D ResNet will then transform the video clips into a sequence of features

$$R^{F\times K} = f_{ResNet}(X^{F\times L\times C\times H\times W}), \tag{2}$$

where K denotes the feature dimension. This operation will treat each video clip independently and generate the features separately. The extracted features will be processed by the

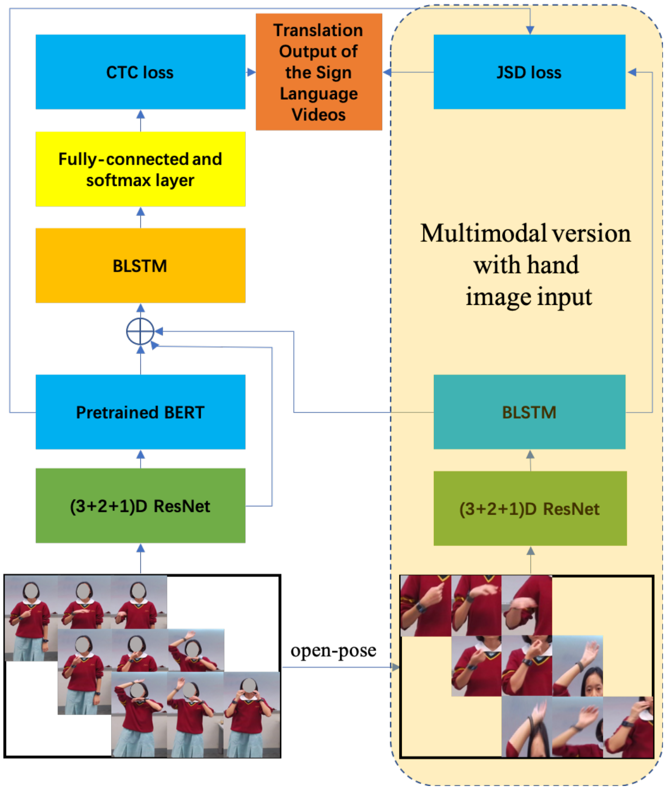

CTC loss
Translation Output of the Sign Language Videos
JSD loss
Fully-connected and softmax layer
BLSTM
Multimodal version with hand image input
Pretrained BERT
BLSTM
(3+2+1)D ResNet
(3+2+1)D ResNet
open-pose

FIGURE 1: The basic structure of the SignBERT framework

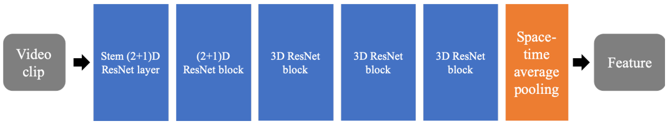


FIGURE 2: The basic structure of the (3+2+1)D ResNet

pretrained BERT model (the original word embedding layer in the BERT model is removed)

$$B^{F\times K} = f_{BERT}(R^{F\times K}). \tag{3}$$

The feature dimension will remain unchanged during this process. As suggested by [32] for maintaining low level features, the features from the ResNet model are concatenated with the features from the BERT structure as the final input features for the next BLSTM layer

$$S^{F\times 2D} = f_{BLSTM_1}(concat(B^{F\times K}, R^{F\times K})), \tag{4}$$

where D is the size of the BLSTM layer and $S^{F\times 2D}$ is the hidden state vector in the BLSTM layer. After that, the hidden state vector is projected from the feature space into the vocabulary space through a fully-connected layer, and the final probability distribution are computed through a softmax layer

$$p = \mathrm{softmax}(W \cdot S + b), \tag{5}$$

in which W and b denote the weight matrix and the bias vector in the fully-connected layer, respectively.

B. THE MULTIMODAL VERSION OF THE SIGNBERT WITH HAND IMAGES AND FEATURE ALIGNMENT

As hand gesture is a crucial part in sign language, we utilized the images of the dominant hand as the additional inputs for the multimodal version SignBERT framework. As presented in Figure 1, given T selected frames, open-pose [59] is firstly adopted to locate the region of the dominant hand for each frame. Given the hand images, the identical sliding window with size L and stride S is applied to construct F video clips. Then, another (3+2+1)D ResNet model is employed to each video clip independently for generating a sequence of hand image features

$$R_{hand}^{F\times K} = f_{ResNet}(x_{hand}^{F\times L\times C\times H\times W}). \tag{6}$$

Instead of using the BERT model, we employ an extra BLSTM layer as the sequential module for processing these hand image features

$$B_{hand}^{F\times K} = f_{BLSTM_2}(R_{hand}^{F\times K}). \tag{7}$$

After that, all the extracted features will be concatenated as a hybrid feature for the final BLSTM layer

$$S^{F\times 2D} = f_{BLSTM_1}(concat(B^{F\times K}, R^{F\times K}, B_{hand}^{F\times K})), \tag{8}$$

in which $R^{F\times K}$ and $B^{F\times K}$ can be obtained from Equation (2) and Equation (3), respectively.

The major reason for not adopting the BERT model for the hand image inputs is to limit the size of the multimodal SignBERT framework for easier training and efficient recognition. In addition, as some gestures in sign language are accomplished with shoulder movements and facial expressions, it is difficult to recognize all the sign language poses with only the hand images. We thus consider the hand images as supplementary inputs and adopt the BLSTM layers for considering the temporal relations among them, instead of the BERT model which requires a longer running time.

C. THE OBJECTIVE FUNCTION

As the ground truth labels of SLR are provided in the sentence-level, the connectionist temporal classification (CTC) [60] is adopted as the main objective function in this work, which can solve the alignment problem between the inputs and target sequences. In CTC, an extra "blank" token is introduced to represent the transitions between two recognizable signs in the video streams. Before computing the loss, CTC firstly conducts an alignment process V to the generated sequences for removing the blank and repeated tokens. For instance, both generated sequences $[\#, \#, a, b, b, \#]$ and $[a, a, \#, b, b]$ are mapped to the same target sequence $[a, b]$ after the alignment process V, in which $\#$ represents the blank token.

In the proposed SignBERT framework, let $P(k, t|\boldsymbol{x}, \boldsymbol{\theta})$ denote the generated probability of a specific gloss k (including blank) at timestamp t, in which $\boldsymbol{x}$ and $\boldsymbol{\theta}$ represent the input video frames and the parameters of the SignBERT framework, respectively. The probability of a specific generated sequence $\boldsymbol{S} = s_{t=1}^{T}$ can then be computed as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132668, IEEE Access

$$P(\boldsymbol{S}|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{t=1}^{T} P(s_t, t|\boldsymbol{x};\boldsymbol{\theta}). \tag{9}$$

One generated sequence $\boldsymbol{S}$ is considered as a correct prediction only if $V(\boldsymbol{S}) = \boldsymbol{Y}$, where $\boldsymbol{Y}$ is the ground truth sentence-level label. Thus, the probability of all correct predictions can be computed as

$$P(\boldsymbol{Y}|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{V(\boldsymbol{S})=\boldsymbol{Y}} P(\boldsymbol{S}|\boldsymbol{x};\boldsymbol{\theta}). \tag{10}$$

Given the probabilities of all correct predictions, the objective function of CTC can then be defined as

$$\mathcal{L}_{CTC}(\boldsymbol{\theta}) = -\log P(\boldsymbol{Y}|\boldsymbol{x};\boldsymbol{\theta}). \tag{11}$$

In addition, to fully train the feature extraction module and BERT model under the limited sizes of the SLR datasets, we further propose an extra feature alignment loss in the multimodal version SignBERT framework.

As the hand image model and the BERT model have the same input videos, their output features at each timestamp are expected to represent the same gloss in the vocabulary. Thus, the feature extraction module and the BERT structure can be trained through minimizing the distance between the probability distributions projected from these output features. Specifically, in the feature alignment loss, the multimodal SignBERT framework firstly adopts a fully-connected layer followed by softmax layers to convert both the features $B^{F\times K}$ and $B^{F\times K}_{hand}$ from Equation (3) and Equation (7) into the probability distributions of different glosses. After that, the distances between these two generated distributions P_{BERT} and P_{hand} are measured through Jensen-Shannon divergence (JSD) loss, which is a symmetrized and smoothed version of the Kullback-Leibler divergence (KLD) loss and can be defined as

$$\mathcal{L}_{JSD} = \frac{1}{2}D(P_{BERT}||M) + \frac{1}{2}D(P_{hand}||M), \tag{12}$$

where D denotes the Kullback-Leibler divergence

$$D(P||Q) = \sum p(x)\log\frac{p(x)}{q(x)}, \tag{13}$$

and M can be computed as

$$M = \frac{1}{2}(P_{hand} + P_{BERT}). \tag{14}$$

Given the JSD loss and the CTC loss, the final loss function can then be described as

$$\mathcal{L} = \mathcal{L}_{CTC} + \lambda\cdot\mathcal{L}_{JSD} + \mu||w||^2, \tag{15}$$

where λ is the weight factor for the JSD loss, w represents the weights in the SignBERT framework and $\mu||w||^2$ is the regularization term to avoid overfitting.

IV. THE IMPLEMENTATION DETAILS

A. THE KEY FRAME SELECTION MECHANISM

Different from previous works which selected the key frames according to only one or two evaluation terms [53]–[57], we propose a new key frame selection mechanism to improve the recognition accuracy. In this mechanism, multi-evaluation terms are considered to identify the critical frames from the raw videos, which include hand movement, hand height and frame blurriness level. The functionalities of these evaluation terms can be explained as follows.

For hand movement, it can be used to avoid selecting similar frames. Since the number of key frames is limited, it is important to ensure the selected key frames have no repetition. Thus, hand movement is adopted to measure the similarity of the two frames for avoiding duplication, which is demonstrated in Figure 4.

For hand height, as shown in Figure 5, we utilize it for removing the useless frames in the videos. As most of the signs are conducted by hands in front of human faces and chests while signers' hands are usually located near their legs at the beginning and the end of sign language videos, the height of the hand can therefore be used to avoid selecting the meaningless frames in which the signers have not yet started to perform the sign language.

For frame blurriness level, it is used to get rid of the frames with motion blurs. Some fast movements of signers can result in motion blur, which are difficult to be recognized by the deep learning models. Thus, we employ the frame blurriness level for picking out the frames with less motion blur. The comparison between the frames with and without motion blur is presented in Figure 6.

To be specific, given a video of sign language with T frames, we define the frame score S_i of $frame_i$ as follows

$$\begin{aligned} S_i = \alpha\cdot\max\Bigg(&\sqrt{(x_i^l - x_{i-1}^l)^2 + (y_i^l - y_{i-1}^l)^2},\\ &\sqrt{(x_i^r - x_{i-1}^r)^2 + (y_i^r - y_{i-1}^r)^2}\Bigg) \\ &-\beta\cdot\min(y_i^l, y_i^r) \\ &+\gamma\cdot Variation(conv2d(gray(frame_i), K_f)). \end{aligned} \tag{16}$$

in which x_i^l, y_i^l and x_i^r, y_i^r are the central coordinates of the left and right hand images recognized by open-pose [59], while α, β and γ are all positive coefficients. In addition, K_f is the Laplacian kernel which is defined as

$$K_f = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \tag{17}$$

In Equation (16), the first and second terms represent hand movement and the height of the hand. Hand movement can prevent the key frame selection mechanism from selecting the repeated frames, while the height of the hand is used for removing the beginning and end of the sign language videos where the heights of signers' hands are usually low. The third term measures the blurriness level of the grayed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132668, IEEE Access

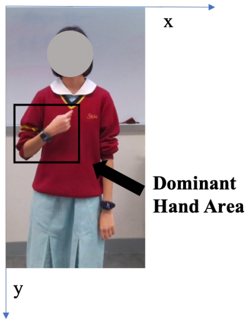


FIGURE 3: The coordinate system for frame selection

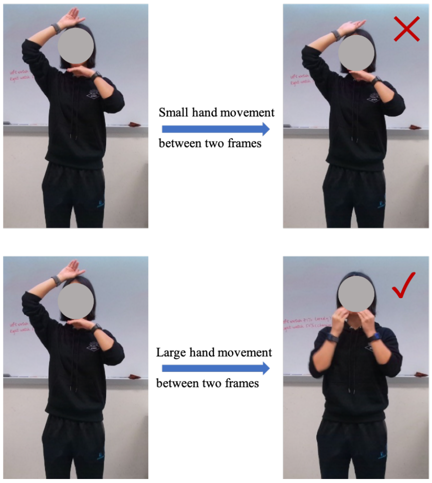


FIGURE 4: The frame with large hand movement is selected by the key frame selection mechanism

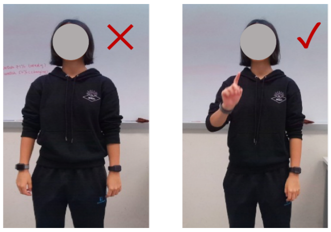


FIGURE 5: The frame with higher hand height is selected by the key frame selection mechanism

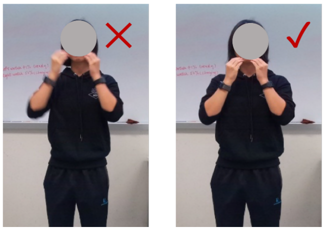


FIGURE 6: The frame with less motion blur is selected by the key frame selection mechanism

$frame_i$ which can be explained as follows. The variance of a well-focus image after edge detection with a Laplacian kernel is usually large, as it contains a lot of clear edges. On the contrary, a blurry image barely has any edge or the edges are smoothed. Thus, the variance after conducting edge detection with a Laplacian kernel is small. This variance can therefore be considered as the frame blurriness level.

B. THE PRETRAINING OF THE BERT MODEL WITH MASKED VIDEO INPUT

Pretraining is one of the crucial steps for deep learning models as it not only improves the model performance but also reduces the need for large datasets. In the proposed SignBERT framework, we pretrain the feature extraction module (ResNet) in the Kinetics-400 dataset and adopt the $BERT_{base}$ model pretrained on the corresponding language as the initial BERT model, which has 12 layers of Transformer blocks, 768 hidden units and 12 self-attention heads.

In addition, as the importance of pretraining the models with different isolated sign language datasets for CSLR has been demonstrated by numerous existing works [3], [28], [46], [56], [61], we also use the corresponding isolated sign language datasets to pretraining the SignBERT framework, including the isolated Chinese sign language dataset [62], the isolated RWTH-PHOENIX-Weather dataset [63] and the Hong Kong isolated sign language dataset [2].

Among these three isolated datasets, the isolated Hong Kong sign language dataset and the isolated Chinese sign language dataset both contain a larger vocabulary than their corresponding continuous sign language datasets, while the isolated RWTH-PHOENIX-Weather dataset contains a smaller vocabulary than the continuous RWTH-PHOENIX-Weather 2014 dataset.

However, unlike other existing works, to ensure the robustness of the SignBERT framework to the possible nonstandard signs of signers, we propose a new pretraining approach for the SignBERT framework in this work: Pretraining the BERT model with masked video input. During the masked pretraining, as demonstrated in Figure 7 (in which $E(i)$ represents the features extracted by the ResNet model while $T(i)$ means the outputs from the BERT model at time i), the last BLSTM layer in the SignBERT framework is firstly removed and a fully-connected layer with a softmax layer is adopted after the BERT model. Secondly, the key frame selection mechanism is utilized to construct the video clips. Instead of using all the video clips as the pretraining input, one video clip in the sequence of video clips is randomly selected and masked. Following with the original implementation of the BERT model [31], if the i^{th} video clip is selected, this video clip is replaced by (1) the [MASK] video clip 80% of the time, in which the [MASK] video clip contains only the background images (2) a random video clip 10% of the time (3) the unchanged i^{th} video clip 10% of the time. After the masking, the masked sequences of video clips are processed by the same feature extraction module and the BERT structure in the SignBERT framework. Then, the i^{th} final hidden state of the BERT structure is projected from the feature space into the vocabulary space by a fully-connected layer for recognizing the meaning of the whole video with cross-entropy loss.

The benefits of pretraining the BERT model with masked video input are threefold: firstly, after pretraining with masked input, the SignBERT framework can still provide a robust recognition result even if a small part of the signs in the videos are incorrect. Secondly, as it does not know which video clip will be masked, the BERT model is forced to keep a distribution of contextual representation for each input video clip. Last but not least, it is feasible to mask different video clips iteratively to make the best use of the isolated sign language datasets for pretraining.

C. THE ITERATIVE TRAINING WITH MASKED SENTENCES

Given the limited size of the datasets, purely training the whole SignBERT framework end-to-end with the loss function may lead to insufficient optimization of the low level ResNet model and the BERT model due to the chain rule in back-propagation. To fully explore the power of the SignBERT framework, a new iterative training strategy with masked sentences for the SignBERT framework is introduced. Different from the implementation of [3], [26], we consider the integration of (3+2+1)D ResNet and the BERT structure in the SignBERT framework as a ResNet-BERT feature extractor and apply the iterative training to optimize its parameters.

During the iterative training, the whole SignBERT framework is firstly trained end-to-end through minimizing the loss term $\mathcal{L}$ from Equation (15). After the training, the sequences of words $\boldsymbol{S}$ generated by the last BLSTM layer in the SignBERT framework are considered as the pseudo labels of the videos at different timestamps. Similar to the procedure in Figure 7, given the sequences of pseudo labels, one input video clip is randomly selected for masking and the SignBERT framework (without the final BLSTM layer) is used to predict the pseudo label at the timestamp of the masked video clip. Note that, during the mask operation, only the video clips with pseudo labels of real words ("blank" label is not included) will be masked. For example, supposed that the sequence of pseudo labels is $[\#, \#, a, \#, \#, b, b, c, c, c]$ and the 6^{th} label is selected for masking, which refers to b, the 6^{th} video clip will then be masked with the same masking method introduced in the last subsection. After the masking, the whole sequence of video clips, including both the masked 6^{th} video clip and other unmasked video clips, will be processed by the (3+2+1)D ResNet and the BERT structure in the SignBERT framework. After that, as illustrated in Figure 7, the 6^{th} output from the ResNet-BERT structure will be used to predict the pseudo label b with cross-entropy loss.

After training the ResNet-BERT structure with the pseudo labels, the whole SignBERT model can then be re-trained through the loss function $\mathcal{L}$ from Equation (15) with the updated ResNet-BERT structure and provide improved pseudo-labels for next iteration. This iterative training procedure can be run continuously until the performance of the SignBERT framework cannot be further improved.

TABLE 1: The Dataset Used for the Signer Independent Test

statistics	HKSL dataset		CSL dataset	
	Train	Test	Train	Test
sentences	50	50	100	100
signers	5	1	40	10
repetitions	8	8	5	5
videos	2,000	400	20,000	5,000
vocabulary	55	55	178	178

TABLE 2: The Dataset Used for the Multi-Signer Test

statistics	RWTH-PHOENIX-Weather 2014		
	Train	Dev	Test
signers	9	9	9
frames	963,664	75,186	89,472
sentences	5,672	540	629
vocabulary	1,081	467	500

V. THE EXPERIMENTAL SETTINGS

To evaluate the performance of the proposed SignBERT framework, extensive experiments are conducted in this paper on three different datasets: a publicly available RWTH-PHOENIX-Weather 2014 Dataset [64], a publicly available Chinese sign language (CSL) dataset [65] and a new Hong Kong sign language (HKSL) dataset, which is newly introduced in this paper and will be available to the public for research purpose.

A. THE INVOLVED DATASETS

In the Hong Kong sign language dataset, we collect the videos for 50 sign language sentences with 6 signers and 8 repetitions for each sentence. Thus there are totally $2,400$ videos in this dataset. All the videos are collected through the Kinect Azure [66] with both color and depth information. For color information, each video has a resolution of 1920×1080 and a frame rate of 30Hz. For depth information, each video has a resolution of 640×576 and a frame rate of 30Hz. The information of the sentences in this dataset are all related to eating and restaurant. All the words in this dataset must have appeared in at least three different sentences to ensure the feasibility for conducting unseen sentence test. In addition, we will keep updating the proposed HKSL dataset so that more sentences can be included.

In addition to the Hong Kong sign language dataset, two public datasets are also considered, including the RWTH-PHOENIX-Weather 2014 dataset [64] and the Chinese sign language dataset [65]. The former contains the videos of around $7,000$ sign language sentences with 9 signers. On the other hand, there are 100 sentences in the latter while each sentence is performed by 50 signers with 5 repetitions. Thus totally $25,000$ videos are included.

We conduct three types of experiments on these datasets:

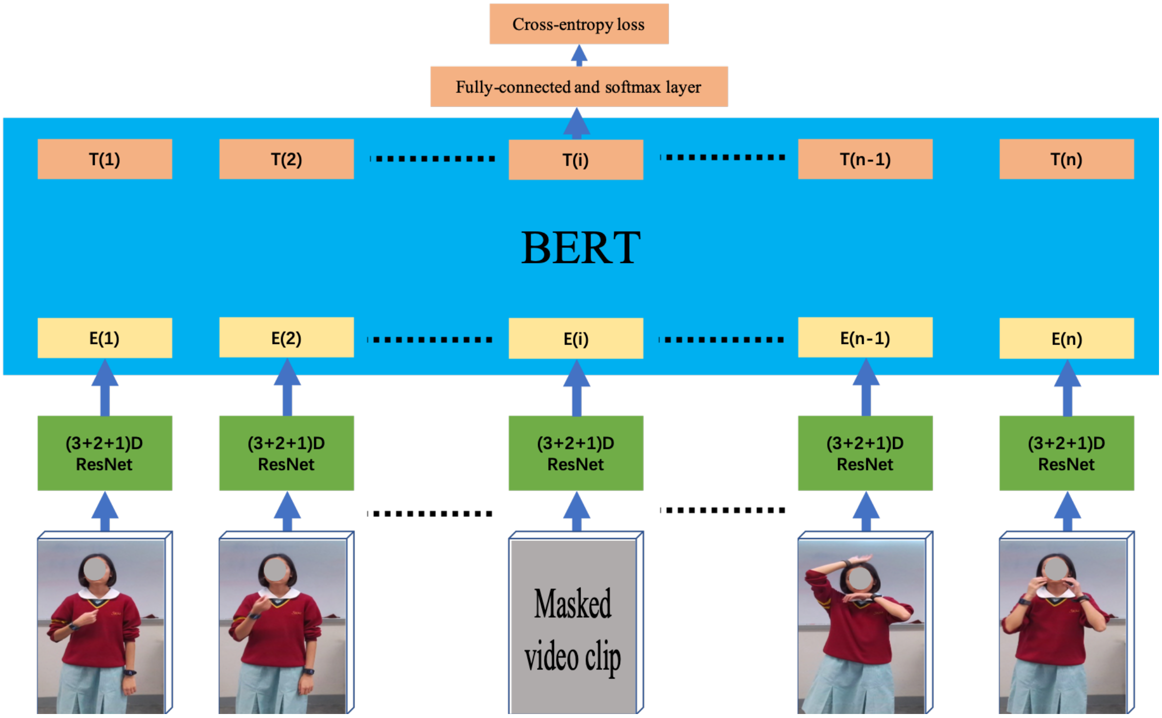


FIGURE 7: Training the BERT model with masked input

TABLE 3: The Dataset Used for the Unseen Sentence Test

statistics	HKSL dataset		CSL dataset	
	Train	Test	Train	Test
sentences	35	15	94	6
signers	6	6	50	50
repetitions	8	8	5	5
videos	1,680	720	23,500	15,000
vocabulary	55	55	178	20

(a) Signer Independent Test: The signers in the testing set have never occurred in the training set. (b) Multi-Signer Test: The signers in the testing set are the same as the signers in the training set. (c) Unseen Sentence Test: the sentences in the testing set have never occurred but each of their words has appeared in other sentences in the training set. For example, if there are two sentences, including “I like apple” and “I have an orange”, in the training set, then, the possible unseen sentence could be “I like orange”. Compared with other existing datasets, one of the major contributions of the proposed HKSL dataset is that the sentences are highly related with each other for supporting the unseen sentence test. Therefore, a high proportion (around 30%) of sentences can be used as the testing set for the unseen sentence test in the proposed HKSL dataset. Some statistics of these three datasets and three experiments are presented in Table 1, Table 2 and Table 3.

B. THE DETAILED EXPERIMENTAL SETTINGS

In the proposed SignBERT framework, $T = 64$ key frames are selected by the key frame selection mechanism for each video and the window size L is set to 4 with a stride $S = 2$ for constructing video clips with the selected frames. In addition, the channel size of the last block in the (3+2+1)D ResNet is set to 768 to match the hidden size of the $BERT_{base}$ model. During the training, the Adam optimizer is utilized to train the framework with a learning rate of 1×10^{-3} and a weight decay of 5×10^{-5}. The hidden size of the final BLSTM layer is designed to be 256 and the λ in the loss function is set to 0.1.

In this work, to evaluate the model accuracy precisely, word error rate (WER) is adopted as the main criterion. Specifically, WER measures the minimal requirements of substitution, deletion and insertion at the word level for converting the predicted sequences into the reference sentences, which can be computed as

$$WER = \frac{S + D + I}{N}, \tag{18}$$

in which S is the count of substitutions, D is the count of deletions, I is the count of insertions and N is the total number of words in the reference sentence. The lower WER represents a better recognition performance.

VI. THE ABLATION STUDIES

In this section, we conduct three experiments to show the effectiveness of different parts in the proposed SignBERT framework, which include the pretraining of the BERT model with masked inputs, the multimodal version of the SignBERT

TABLE 4: The Experimental Results on the Pretraining of the BERT

Method	WER(%)
Baseline	10.02
No pretraining SignBERT	5.81
Normally-pretrained SignBERT	4.75
Mask-pretrained SignBERT	**2.96**

TABLE 5: The Experimental Results on the Multimodal SignBERT Framework

Method	WER(%)
Multimodal baseline	9.07
SignBERT	2.96
Multimodal SignBERT without feature alignment	2.53
Multimodal SignBERT with feature alignment	**2.15**

framework with feature alignment, and the iterative training of the SignBERT framework with masked sentences.

A. THE EXPERIMENTAL RESULTS ON THE PRETRAINING OF THE BERT MODEL WITH MASKED INPUTS

To explore the importance of the BERT pretraining with masked input, an experiment is conducted on the signer independent test of the CSL dataset for the SignBERT framework (without hand image input) in different versions, including no pretraining, normally-pretrained and mask-pretrained. In the no pretraining version, the initial weight of the BERT structure comes from a text pretrained checkpoint provided by the authors of [31] while the initial weight of the (3+2+1)D ResNet is only pretrained on the Kinetics-400 dataset. In the normally-pretrained version, the SignBERT framework is pretrained on the isolated sign language datasets without any masking and the final hidden state of the BERT model at the first timestamp is used to predict the corresponding word with cross-entropy loss. On the contrary, the mask-pretrained version of SignBERT is produced through the mask-pretraining method introduced in Section IV. To provide a fair and systematic comparison, we also construct a baseline method which is a CNN-BLSTM model created through removing the BERT model in the SignBERT framework and connecting the ResNet with the final BLSTM layer directly. The experimental results of these models are listed in Table 4.

As seen in Table 4, the traditional CNN-BLSTM structure performs worse than the SignBERT framework significantly. This is mainly due to the insufficient abilities of the sequential module in language modeling, which contains only one BLSTM layer. Without the BERT structure, no language prior information is provided for the model, which makes it difficult to infer the output sentences accurately with only the videos input. Meanwhile, it can also be observed that the WER of the SignBERT framework drops from 5.81% to 4.75% after being normally-pretrained, which indicates the importance of pretraining, as it can transfer the knowledge from other datasets. Furthermore, compared with the normally-pretrained SignBERT, the mask-pretrained SignBERT achieves a lower WER of 2.96%. This demonstrates the effectiveness of mask-pretraining as it not only improves the robustness of the SignBERT framework but also minimizes the undue influence of the nonstandard signs from the signers.

B. THE EXPERIMENTAL RESULTS ON THE MULTIMODAL VERSION OF THE SIGNBERT FRAMEWORK WITH FEATURE ALIGNMENT

To demonstrate the effectiveness of the multimodal SignBERT framework, we conduct an experiment for comparing its performance with the original SignBERT framework. The signer independent test of the CSL dataset is used as the testing method and the WER is considered as the evaluation criterion. All the BERT models are mask-pretrained as presented in Section IV. For a fair comparison, the baseline method described in the last subsection is further upgraded to the multimodal baseline by integrating the hand images as its additional inputs, and concatenating the features of the hand images generated by the (3+2+1)D ResNet with the original features before the final BLSTM layer. In addition, to show the effectiveness of the feature alignment, we also compare the performance of the multimodal SignBERT with and without feature alignment (the multimodal SignBERT without feature alignment only uses CTC in its loss function).

As presented in Table 5, although the WER is decreased after introducing the hand images in the baseline method, the multimodal baseline still performs worse than the SignBERT framework. Meanwhile, compared with the original SignBERT framework, the multimodal SignBERT without feature alignment achieves a lower WER of 2.53%. This result is expected as hand gesture is the most important part in the sign language. The hand image inputs can provide direct information of the crucial part in SLR for the multimodal SignBERT. In addition, it can be observed that, by integrating feature alignment, the multimodal SignBERT achieves a better performance of 2.15% WER. This could be explained as the process of feature alignment can not only help to train the low-level layers in the multimodal SignBERT framework more efficiently but also force the multimodal SignBERT model to pay more attention to the hand image areas. Therefore, in what follows, the multimodal SignBERT framework will be integrated with feature alignment by default unless specifically noted.

C. THE EXPERIMENTAL RESULTS ON THE ITERATIVE TRAINING WITH MASKED SENTENCES

As mentioned in Section IV, the proposed SignBERT framework can be further optimized through iterative training with masked sentences. Figure 8 shows the performance of the SignBERT framework and multimodal SignBERT with feature alignment in different iterations for the signer independent test of the CSL dataset. Likewise, both the SignBERT

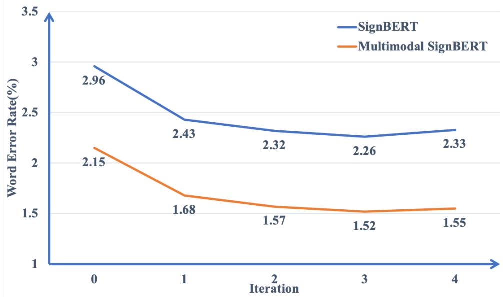


FIGURE 8: The Experimental Results of the Iterative Training

and the multimodal SignBERT are pretrained in the masked isolated CSL dataset. The word error rate drops along with the iterative training process until the third iteration. Specifically, the WER of the SignBERT framework decreases from 2.96% to 2.26% while the WER of multimodal SignBERT framework drops from 2.15% to 1.52%. Besides, the first iteration shows the most positive impact on the model performance while no further improvement can be found after the third iteration.

VII. THE DETAILED PERFORMANCE ANALYSIS

In this section, the proposed SignBERT framework and other state-of-the-art methods that listed in Table 6 are compared systematically based on their performance in the three different continuous sign language datasets.

A. THE PERFORMANCE COMPARISON ON THE CSL DATASET

Table 7 presents the performance of the proposed SignBERT framework and other existing methods in the signer independent test of the CSL dataset. The SignBERT frameworks (both original version and multimodal version with feature alignment) are compared with the following methods: S2VT [67], LS-HAN [68], SubUNet [69], HRF-S [56], HRF-S-att [56], IAN [3], SF-NET [28] and FCN [29].

We also compare the SignBERT framework with other methods in the unseen sentence test of the CSL dataset and the experimental results are shown in Table 8. The performances of the following methods are included: S2VT [67], HRF-S [56], HRF-S-att [56], DGM [71], IAN [3], SBD-RL [72] and CMA [52].

It can be observed from Table 7 and Table 8 that the proposed multimodal SignBERT achieves the best performance compared with other state-of-the-art methods in these two tests. Specifically, in signer independent test of the CSL dataset, the SignBERT framework outperforms other methods by 0.74% in WER while the multimodal SignBERT achieves the lowest WER of only 1.52%. For unseen sentence test of the CSL dataset, the SignBERT model reaches a similar performance as CMA while the WER is further reduced to 23.30% after introducing the hand image inputs and feature alignment into the SignBERT framework. Note that, compared with CMA, the proposed SignBERT framework is an end-to-end approach without the need of additional input data while CMA requires a lot of human efforts to create additional pseudo data for training.

B. THE PERFORMANCE COMPARISON ON THE RWTH-PHOENIX-WEATHER 2014 MULTI-SIGNER DATASET

Table 9 compares the performance of the SignBERT framework with other existing approaches in the public RWTH-PHOENIX-Weather 2014 Multi-Signer dataset [64]. The following approaches are included for comparison: SubUNet [69], LS-HAN [68], IAN [3], DenseTCN [70], SF-Net [28], SBD-RL [72], FCN [29], Re-Sign [73] and CMA [52]. In this experiment, the BERT model was firstly pretrained by German text. Then, the isolated RWTH-PHOENIX-Weather dataset was utilized for mask-pretraining the SignBERT framework as introduced in Section IV. Compared with other existing approaches, lower WER (21.2% in the validation set and 21.4% in the test set) is attained by the proposed SignBERT framework while the multimodal SignBERT framework attains the lowest WER of 20.1% in the validation set and 20.2% in the test set.

C. THE PERFORMANCE COMPARISON ON THE HKSL DATASET

The Hong Kong sign language dataset is a newly collected dataset and there is no existing experimental result on this dataset. To solve this issue, we implement three methods which achieve good performance in the CSL dataset, including DenseTCN [70], SF-Net [28] and FCN [29] (CMA is not included as it requires extra human effort for creating pseudo data), and apply them to the proposed HKSL dataset for comparison. Similarly, experiments are conducted on both the signer independent test and the unseen sentence test in the proposed HKSL dataset. In these two tests, both the SignBERT and the multimodal SignBERT frameworks have been pretrained and iteratively trained as presented in Section IV.

As exhibited in Table 10, the proposed SignBERT framework and the multimodal SignBERT outperform other state-of-the-art methods in both experiments. For the signer independent test, the proposed SignBERT model outperforms the state-of-the-art method by 1.06% in WER while the multimodal SignBERT reaches a lowest WER of 4.84%. For the unseen sentence test, the proposed SignBERT framework reduces the WER from 20.37% of the FCN to 15.92% while the multimodal SignBERT reaches the best performance with a WER of 12.35%.

D. ANALYSIS OF THE EXPERIMENTAL RESULTS

As presented in Table 7, Table 8, Table 9 and Table 10, the proposed multimodal SignBERT framework attains a lowest WER compared with other existing methods in three different continuous sign language datasets, which can be attributed to the following advantages:

TABLE 6: State-of-the-Art Methods for Comparison

Method	Description
S2VT [67]	SV2T is a sequence to sequence framework for video to text
LS-HAN [68]	LS-HAN employs latent space and hierarchical attention network in CSLR
SubUNET [69]	SubUNET decomposes the problem of CSLR into a series of specialized expert systems
HRF-S [56]	HRF-S performs a clip summarization and adopts hierarchical recurrent deep fusion for CSLR
HRF-S-att [56]	HRF-S-att upgrades the HRF-S with attention mechanism
BAE [48]	BAE applies a boundary-adaptive encoder-decoder framework for SLR
IAN [3]	IAN is an iterative alignment network with 3D ResNet and BLSTM
DenseTCN [70]	DenseTCN employs temporal convolution network for CSLR
SF-Net [28]	SF-Net is a structured feature network with one feature extractor and different level BLSTM layers
FCN [29]	FCN constructs a fully convolutional network by replacing the LSTM layers with convolutional networks
DGM [71]	DGM uses multi-classifiers modules to identify the words and n-grams in a sentence
SBD-RL [72]	SBD-RL conducts SLR through semantic boundary detection with reinforcement learning
Re-Sign [73]	Re-sign employs re-aligned sequence modeling with deep recurrent CNN-HMMs for CSLR
CMA [52]	CAM adopts cross modality augmentation with a lot of pseudo video-text pairs to train the model

TABLE 7: The Experimental Results on the Signer Independent Test of the CSL Dataset

Method	WER(%)
S2VT [67]	25.50
LS-HAN [68]	17.30
SubUNet [69]	11.0
HRF-S [56]	10.70
HRF-S-att [56]	10.20
BAE [48]	7.40
IAN [3]	6.10
SF-Net [28]	3.80
FCN [29]	3.00
SignBERT	2.26
Multimodal SignBERT	**1.52**

TABLE 8: The Experimental Results on the Unseen Sentence Test of the CSL Dataset

Method	WER(%)
S2VT [67]	67.00
HRF-S [56]	66.20
HRF-S-att [56]	64.10
DGM [71]	50.90
IAN [3]	32.70
SBD-RL [72]	26.80
CMA [52]	24.50
SignBERT	24.90
Multimodal SignBERT	**23.30**

TABLE 9: The Experimental Results on the RWTH-PHOENIX-Weather 2014 Dataset

Method	Dev WER (%)	Test WER (%)
SubUNet [69]	40.8	40.7
LS-HAN [68]	-	38.3
IAN [3]	37.1	36.7
DenseTCN [70]	35.9	36.5
SF-NET [28]	35.6	34.9
SBD-RL [72]	28.6	28.6
FCN [29]	23.7	23.9
Re-sign [73]	23.8	24.4
CMA [52]	21.3	21.9
SignBERT	21.2	21.4
Multimodal SignBERT	**20.1**	**20.2**

1) The SignBERT framework adopts the BERT model for processing the temporal dependencies among the extracted features. Compared with the RNN structure that are widely employed, the BERT model has a stronger ability in language modeling.
2) An effective key frame selection mechanism is adopted in the SignBERT framework, which can filter out some useless frames and select the most meaningful frames for the SignBERT framework.
3) The SignBERT framework is pretrained with masked videos and iteratively trained with masked sentences, which not only exploits its ability of in language understanding but also improves its robustness to the videos with incorrect signs.
4) The multimodal SignBERT framework considers the images of the dominant hand as its additional input with a feature alignment approach to fully train the feature extraction module and increase the recognition accuracy.

In addition, according to Table 7, Table 8 and Table 10, the SignBERT framework obtains a larger improvement than other existing methods in the unseen sentence test compared with the signer independent test. This is mainly because the unseen sentence test has a higher requirement in language understanding while the BERT model inside the SignBERT framework improves its capability in understanding and recognizing unseen sentences.

However, the SignBERT framework still contains some disadvantages which should be noted and improved in the future:

1) Compared with the traditional structure with only CNN and RNN, the proposed SignBERT framework is slower in processing the same video, which could be an issue for real-world application.
2) The key frame selection mechanism in the SignBERT framework needs to calculate the scores for all frames before selecting the key frames, which is difficult to implement when online sign language translation is required.

TABLE 10: The Experimental Results on the HKSL Dataset

Method	Signer Independent Test WER(%)	Unseen Sentence Test WER(%)
DenseTCN [70]	7.91	24.55
SF-Net [28]	7.69	22.08
FCN [29]	7.16	20.37
SignBERT	6.10	15.92
Multimodal SignBERT	**4.84**	**12.35**

VIII. CONCLUDING REMARKS

Continuous sign language recognition (CSLR) is a challenging research direction involving both video analytics and language modeling while subject to stringent real-time constraints. Previous studies in CSLR mostly employ hidden Markov model or recurrent neural network for processing the temporal information with a limited capability for language modeling, and often fail to utilize the important information from other datasets to achieve a high recognition accuracy when encountering the atypical signs performed by different signers. To deal with this critical issue, we propose a pioneering BERT-based and robust deep learning framework, namely SignBERT, for CSLR. In this framework, high quality video clips are constructed from the videos of sign language with an intelligent key frame selection mechanism, and the (3+2+1)D ResNet is utilized for extracting visual features from the video clips. Afterwards, the BERT model is applied for language modeling, which is pretrained with the partially masked videos from the datasets of isolated sign language to strengthen its resilience to non-standard signs. To further improve the recognition accuracy, we also develop a multimodal version of the SignBERT framework to integrate the hand images as an additional input supported by a feature alignment strategy to compare the probability distribution generated from the BERT model and the hand images. Furthermore, to fully exploit the capability of the SignBERT framework under the available datasets, a new iterative training strategy is devised, in which one randomly selected word of the final recognition results from the SignBERT framework is masked for re-training the feature extraction module and the BERT model iteratively. To evaluate the performance of the SignBERT framework, a radically new dataset of Hong Kong Sign Language containing $2,400$ videos of 50 sentences is collected. Experimental results on the newly collected Hong Kong Sign Language dataset, the publicly available dataset of Chinese Sign Language and the RWTH-PHOENIX-Weather 2014 dataset reveal that the proposed SignBERT framework achieves state-of-the-art performance in CSLR.

More importantly, various possible research directions are opened up in this work. First, it is valuable to study the performance of other intelligent algorithms originally designed for natural language processing to cater for CSLR, such as the Google T5 [74]. Second, only the hand images are included as the additional input in this work for the multimodal version SignBERT. It would be attractive if more sources of information such as the optical flow and depth information of the relevant images can be considered. Last but not least, it is worthwhile to develop an application for CSLR, such as [75], to break the communication barriers between the deaf and normal people.

ACKNOWLEDGEMENT

The authors wish to thank the Christian and Missionary Alliance Sun Kei Secondary School for helping us in collecting the dataset of the Hong Kong Sign Language.

REFERENCES

[1] "World federation of the deaf," http://wfdeaf.org/our-work/, accessed: 2021-04-06.

[2] Z. Zhou, K.-S. Lui, V. W. Tam, and E. Y. Lam, "Applying (3+2+1)D residual neural network with frame selection for Hong Kong sign language recognition," in *25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4296–4302.

[3] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4160–4169.

[4] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Read and attend: Temporal localisation in sign language videos," in *CVPR*, 2021.

[5] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman, "Watch, read and lookup: Learning to spot signs from multiple supervisors," in *Asian Conference on Computer Vision*, 2020.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[7] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4213–4222.

[8] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudrealt, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. Morris, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 08 2019, pp. 16–31.

[9] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," in *Proceedings of the IEEE International Conference on Computer Vision*, 12 1995, pp. 265 – 270.

[10] J. Forster, C. Oberdörfer, O. Koller, and H. Ney, "Modality combination techniques for continuous sign language recognition," in *Pattern Recognition and Image Analysis*, 2013, pp. 89–99.

[11] 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*, vol. 141, pp. 108–125, 2015.

[12] C. Monnier, S. German, and A. Ost, "A multi-scale boosted detector for efficient and robust gesture recognition," in *ECCV Workshops*, 2014, pp. 491–502.

[13] L. Prasuhn, Y. Oyamada, Y. Mochizuki, and H. Ishikawa, "A hog-based hand gesture recognition system on a mobile device," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 3973–3977.

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

[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*, vol. 141, pp. 108–125, 2015.

[16] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2017.

[17] Z. Tu, W. Xie, J. Dauwels, B. Li, and J. Yuan, "Semantic cues enhanced multimodality multistream cnn for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1423–1437, 2019.

[18] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 1002–1014, 2018.

[19] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2017.

[20] O. Koller, N. 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*, vol. 42, no. 9, pp. 2306–2320, 2020.

[21] E. K. Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. T. K. Kumar, and D. A. Kumar, "Training cnns for 3-d sign language recognition with color texture coded joint angular displacement maps," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 645–649, 2018.

[22] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3d-cnns for large-vocabulary sign language recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2822–2832, 2019.

[23] 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*, vol. 38, no. 8, pp. 1583–1597, 2016.

[24] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid cnn-hmm for continuous sign language recognition," in *BMVC*, 2016, pp. 1–12.

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

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

[27] L. Pigou, A. Oord, S. Dieleman, M. Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *Int. J. Comput. Vision*, vol. 126, no. 2–4, p. 430–439, 2018.

[28] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "Sf-net: Structured feature network for continuous sign language recognition," *ArXiv*, 2019.

[29] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," pp. 697–714, 2020.

[30] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, 2017.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[32] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7463–7472.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[34] G. Evangelidis, G. Singh, and R. Horaud, "Continuous gesture recognition from articulated poses," in *ECCV Workshops*, vol. 8925, 03 2015, pp. 595–607.

[35] 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*, vol. 141, pp. 108–125, 2015.

[36] M. A. Bencherif, M. Algabri, M. A. Mekhtiche, M. Faisal, M. Alsulaiman, H. Mathkour, M. Al-Hammadi, and H. Ghaleb, "Arabic sign language recognition system using 2d hands and body skeleton data," *IEEE Access*, vol. 9, pp. 59 612–59 627, 2021.

[37] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with blstm-3d residual networks," *IEEE Access*, vol. 7, pp. 38 044–38 054, 2019.

[38] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1610–1618.

[39] W. Aly, S. Aly, and S. Almotairi, "User-independent american sign language alphabet recognition based on depth image and pcanet features," *IEEE Access*, vol. 7, pp. 123 138–123 150, 2019.

[40] N. Sarhan and S. Frintrop, "Transfer learning for videos: From action recognition to sign language recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1811–1815.

[41] L.-C. Wang, R. Wang, D.-H. Kong, and B.-C. Yin, "Similarity assessment model for chinese sign language videos," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 751–761, 2014.

[42] T. Pfister, J. Charles, and A. Zisserman, "Large-scale learning of sign language by watching tv," 01 2013, pp. 20.1–20.11.

[43] F. Shah, M. S. Shah, W. Akram, A. Manzoor, R. O. Mahmoud, and D. S. Abdelminaam, "Sign language recognition using multiple kernel learning: A case study of pakistan sign language," *IEEE Access*, vol. 9, pp. 67 548–67 558, 2021.

[44] B. Xu, S. Huang, and Z. Ye, "Application of tensor train decomposition in s2vt model for sign language recognition," *IEEE Access*, vol. 9, pp. 35 646–35 653, 2021.

[45] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical lstm for sign language translation," in *AAAI*, 2018.

[46] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018.

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

[48] S. Huang and Z. Ye, "Boundary-adaptive encoder with attention method for chinese sign language recognition," *IEEE Access*, vol. 9, pp. 70 948–70 960, 2021.

[49] P. M. Ferreira, D. Pernes, A. Rebelo, and J. S. Cardoso, "Desire: Deep signer-invariant representations for sign language recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 9, pp. 5830–5845, 2021.

[50] H. jie Wang, X. juan Chai, X. Hong, G. Zhao, and X. Chen, "Isolated sign language recognition with grassmann covariance matrices," *ACM Trans. Access. Comput.*, vol. 8, pp. 14:1–14:21, 2016.

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

[52] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[53] W. Pan, X. Zhang, and Z. Ye, "Attention-based sign language recognition network utilizing keyframe sampling and skeletal features," *IEEE Access*, vol. 8, pp. 215 592–215 602, 2020.

[54] S. Huang, C. Mao, J. Tao, and Z. Ye, "A novel chinese sign language recognition method based on keyframe-centered clips," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 442–446, 2018.

[55] Y. Yan, Z. Li, Q. Tao, C. Liu, and R. Zhang, "Research on dynamic sign language algorithm based on sign language trajectory and key frame extraction," in *IEEE 2nd International Conference on Electronics Technology (ICET)*, 2019, pp. 509–514.

[56] 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*, vol. 29, pp. 1575–1590, 2020.

[57] Z. Liu, X. Qi, and L. Pang, "Self-boosted gesture interactive system with st-net," p. 145–153, 2018.

[58] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[59] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[60] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," vol. 2006, 01 2006, pp. 369–376.

[61] H. Zhou, W. Zhou, and H. Li, "Dynamic pseudo label decoding for continuous sign language recognition," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 1282–1287.

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

[63] "Rwth-phoenix-weather dataset," https://www-i6.informatik.rwth-aachen.de/ forster/database-rwth-phoenix.php, accessed: 2021-04-06.

[64] 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*, vol. 141, pp. 108–125, 2015.

[65] "Chinese sign language recognition dataset," http://home.ustc.edu.cn/ pjh/openresources/cslr-dataset-2015/index.html, accessed: 2021-04-06.

[66] "Azure kinect dk," https://azure.microsoft.com/en-us/services/kinect-dk/, accessed: 2021-04-06.

[67] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence – video to text," *IEEE International Conference on Computer Vision (ICCV)*, pp. 4534–4542, 2015.

[68] J. Huang, W. gang Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018.

[69] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," pp. 3075–3084, 2017.

[70] D. Guo, S. Wang, Q. Tian, and M. Wang, "Dense temporal convolution network for sign language translation," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 7 2019, pp. 744–750.

[71] C. Wei, W. Zhou, J. Pu, and H. Li, "Deep grammatical multi-classifier for continuous sign language recognition," pp. 435–442, 2019.

[72] C. Wei, J. Zhao, W. Zhou, and H. Li, "Semantic boundary detection with reinforcement learning for continuous sign language recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1138–1149, 2021.

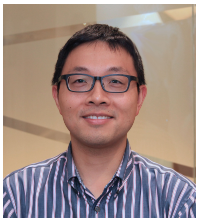
[73] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3416–3424.

[74] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 140:1–140:67, 2020.

[75] Z. Zhou, Y. Neo, K.-S. Lui, V. W. Tam, E. Y. Lam, and N. Wong, "A portable Hong Kong sign language translation platform with deep learning and jetson nano," in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 89:1–89:4.

ZHENXING ZHOU (Student Member, IEEE) received his B.S. degree from the Harbin Institute of Technology in 2017 and his Master degree from the University of Hong Kong in 2018. He is currently a Ph.D. candidate in the Department of Electrical and Electronic Engineering, the University of Hong Kong. His research directions mainly include deep learning, computer vision, pose estimation, big data analytics and artificial intelligence. Besides, he is now serving as the activity co-chair for the IEEE (HK) Computational Intelligence Chapter.

VINCENT W.L. TAM (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Melbourne, in 1998. He is currently a Principal Lecturer and also Honorary Associate Professor with the Department of Electrical and Electronic Engineering, The University of Hong Kong. He has over 150 internationally refereed publications, including 10 book chapters. His main research interests include big data analytics, computational intelligence, cloud computing, machine learning, and information visualization. Besides, he served as the Chairman of the IEEE (HK) Computational Intelligence Chapter (2014–2017).



EDMUND Y. LAM (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University. From 2010 to 2011, he was a Visiting Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. He is currently a Professor in electrical and electronic engineering with The University of Hong Kong. He also serves as the Computer Engineering Program Director of The University of Hong Kong. His main research interest includes computational imaging. He is also a fellow of the OSA, SPIE, IS&T, and HKIE. He was a recipient of the IBM Faculty Award.

• • •