A Cross-Attention BERT-Based Framework for Continuous Sign Language Recognition

Zhenxing Zhou, *Member, IEEE*, Vincent W.L. Tam, *Senior Member, IEEE*, and Edmund Y. Lam, *Fellow, IEEE*

***Abstract*—Continuous sign language recognition (CSLR) is a challenging task involving various signal processing techniques to infer the sequences of glosses performed by signers. Existing approaches in CSLR typically use multiple input modalities such as the raw video data and the extracted hand images to improve their recognition accuracy. However, the large modality differences make it difficult to define an integrative framework to effectively exchange and combine the knowledge obtained from different modalities such that they can complement each other for improving the framework's robustness against the gesture variations and background noises in CSLR. To address this issue, we propose a novel cross-attention deep learning framework named the CA-SignBERT. This framework utilizes multiple Bidirectional Encoder Representations from Transformers (BERT) models to analyze the information from different modalities. Among these BERT models, we introduce a special cross-attention mechanism to ensure an efficient inter-modality knowledge exchange. Besides, an innovative weight control module is proposed to dynamically hybridize their outputs. Experimental results reveal that the CA-SignBERT framework attains state-of-the-art performance in four benchmark CSLR datasets.**

***Index Terms*—Attention mechanism, sign language recognition, bidirectional encoder representations from transformers.**

I. Introduction

SIGN languages are the most important communication methods used by the deaf to express their ideas through hand gestures and body movements. Yet the special linguistic rules and the complicated gestures in sign languages seriously hinder the hearing people from using it. As a result, many communication barriers exist for the deaf. In order to address this issue, some real-world applications [1], [2], [3] of continuous sign language recognition (CSLR) have been developed to decipher the continuous sequences of the glosses in sign languages [4], [5], [6] for breaking down these barriers. However, the performance of these applications can be seriously affected by the gesture variations and background noises, which significantly limits their applicability. In addition, some glosses in sign languages can be similar in terms of their hand movements or shapes which makes them difficult to be differentiated by the existing CSLR approaches.

In fact, these issues can be alleviated by deriving additional modalities such as hand images and joint coordinates from the raw video input such that they can be used as the extra information channels during the recognition. These modalities could contain domain-specific merits that are complementary to the raw input data. For example, the hand images contain the most detailed information of the signers' gestures while the joint coordinate data is immune to the illumination conditions and other environmental issues. Integrating them with the original input modalities can thus not only improve the robustness of the recognition models against the interference caused by the non-standard gestures and environmental noises [7], but also provide more information channels to distinguish the similar glosses.

Therefore, many existing works have employed multiple input modalities for CSLR [8], [9], [10], [11]. Most of them directly combine the knowledge learned from different modalities with a fixed weight [11], [12], [13]. Yet this is insufficient for CSLR. As a matter of fact, to attain an improved recognition accuracy in CSLR with multiple input modalities, a dynamic combination weight is required since each modality is suitable to recognize a different group of glosses. For instance, some glosses such as the digits in sign languages mainly consist of finger movements. In this case, the hand images will be a suitable modality as it contains the most concrete information of the signers' hands. On the other hand, the joint coordinate data could be a better choice when targeting at the glosses with many body and arm movements. Accordingly, it is more reasonable to dynamically adjust their combination weights according to the recognition targets and their recognition accuracy.

Besides, a systematic and feasible communication scheme for exchanging the knowledge from different modalities is also vital to CSLR. Taking the attention mechanism [14] that has widely been employed in CSLR [11], [15] as an example, the attention matrix generated from a single input modality could be biased due to some modality-specific limitations when recognizing a special gloss. Clearly, this will ultimately influence the overall performance of the CSLR approaches. It is thus critical to exchange the attention information among different modalities in the recognition models to minimize the risk of being adversely affected by this error.

Motivated by the above discussions, we propose a cross-attention BERT-based framework named the CA-SignBERT for CSLR. The graphical overview of it is shown in Fig. 1. Basically, this framework consists of multiple information paths. Each path contains a deep learning based feature extractor

Manuscript received 10 July 2022; revised 11 August 2022; accepted 11 August 2022. Date of publication 17 August 2022; date of current version 1 September 2022. The work is supported in part by the Research Council of Hong Kong under Grant T44-707/16-N. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianxin Li. *(Corresponding author: Zhenxing Zhou.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee (HREC) of the University of Hong Kong (Application No. EA1604035).

The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: zxchow@connect.hku.hk; vtam@eee.hku.hk; elam@eee.hku.hk).

Digital Object Identifier 10.1109/LSP.2022.3199665

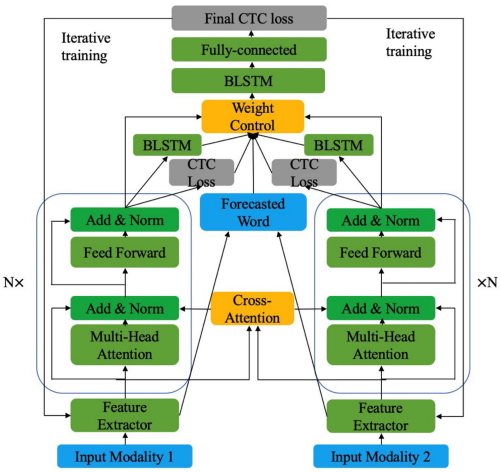

Fig. 1. The Structure of the CA-SignBERT Framework.

and a pretrained Bidirectional Encoder Representations from Transformers (BERT) model [16] for processing the input from one modality such as the raw video input or the extracted hand images. Between any pair of the information paths, we propose a new cross-attention mechanism to exchange the knowledge learned from different input modalities. Meanwhile, an innovative weight control module is introduced in the CA-SignBERT to flexibly adjust the fusion weights of the outputs from different information paths.

Furthermore, to provide additional training to the CA-SignBERT, we introduce a new masked iterative training method which uses the partially masked pseudo labels for re-training the framework iteratively.

In summary, the main contributions of this work are stated as follows:

1) A BERT-based deep learning framework named the CA-SignBERT is developed for CSLR in which multiple BERT models are utilized to process the inputs from different modalities;
2) A cross-attention mechanism is proposed for the CA-SignBERT to exchange the information obtained from different modalities such that they can complement each other for attaining a higher recognition accuracy.
3) A weight control module is designed to dynamically adjust the fusion weights of the outputs from different information paths such that more attention can be payed to the modalities with better performance.

The rest of this letter is organized as follows. In Section II, the proposed CA-SignBERT framework consisted of the cross-attention mechanism and the weight control module will be discussed in detail. Section III will present an ablation study and the experimental results on four public CSLR datasets. Lastly, Section IV will conclude this work.

II. The Proposed CA-SignBERT Framework

A. An Overview of the Framework

The objective of the CA-SignBERT framework is to effectively utilize the information from different input modalities for improving the recognition accuracy in CSLR. To achieve this goal, we integrate multiple information paths in the CA-SignBERT. Each information path is composed of a modality-specific feature extractor and a BERT model for analyzing the data of one input modality. Between any pair of the information paths, we introduce a novel cross-attention mechanism to exchange the knowledge learned from different modalities such that they can complement each other. Furthermore, in order to promote the strengths of each input modality while avoiding its weaknesses, we propose a new weight control module in the CA-SignBERT to dynamically hybridize the outputs from different information paths according to multiple evaluation factors including the preliminary recognition results from the feature extractors, the recognition performance of different modalities and the sequential information generated by the bidirectional long short-term memory (BLSTM) layers.

Fig. 1 demonstrates the structure of the CA-SignBERT framework with two information paths (It is possible to be extended to three or more). In each information path, there is a feature extractor and a BERT model. The feature extractor is used to extract representative features from the input data for which its structure can be modified according to different input modalities. In this letter, we adopt the (3+2+1)D ResNet [17] for the visual data inputs and a two-layer CNN model [18] for other non-visual inputs. The extracted features are then fed into the BERT model for sequential learning and language modeling. Between the BERT models in different information paths, there is a cross-attention mechanism to share the attention knowledge learned from each modality. After that, the outputs from both information paths are combined by the weight control module. Lastly, the fused knowledge will be processed by the final BLSTM layer followed by a fully-connected layer and a softmax layer to produce the prediction and calculate the CTC loss [19] for back-propagation.

B. The Cross-Attention Mechanism

Although the BERT model has a strong capability in language modeling, it can only exploit the intra-modality information but not the inter-modality which restricts its further development in CSLR. To solve this issue, we propose a novel cross-attention mechanism such that the BERT models can utilize the knowledge learned from other modalities.

As demonstrated in Fig. 1, the cross-attention mechanism firstly stacks the input feature vectors (denoted as F_1 and F_2 respectively) of the first and second information paths as:

$$\mathbf{F}_c = \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix}, \tag{1}$$

This integrated feature vector (F_c) is then fed into a multi-head attention unit [14] where the query (Q_c), key (K_c) and value (V_c) are defined as

$$\mathbf{Q}_c = \mathbf{F}_c \mathbf{W}_c^Q = \begin{pmatrix} \mathbf{F}_1 \mathbf{W}_c^Q \\ \mathbf{F}_2 \mathbf{W}_c^Q \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{pmatrix}, \tag{2}$$

$$\mathbf{K}_c = \mathbf{F}_c \mathbf{W}_c^K = \begin{pmatrix} \mathbf{F}_1 \mathbf{W}_c^K \\ \mathbf{F}_2 \mathbf{W}_c^K \end{pmatrix} = \begin{pmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{pmatrix}, \tag{3}$$

$$\mathbf{V}_c = \mathbf{F}_c \mathbf{W}_c^V = \begin{pmatrix} \mathbf{F}_1 \mathbf{W}_c^V \\ \mathbf{F}_2 \mathbf{W}_c^V \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}. \tag{4}$$

in which the W_c^Q, W_c^K and W_c^V are the trainable weight vectors while (Q_1, Q_2), (K_1, K_2) and (V_1, V_2) denote the vectors of query, key and value from the first and second information paths, respectively. Based on the Q_c, K_c and V_c, we can compute the

"Scaled Dot-Product Attention" [14] as

$$\mathbf{A}_c = Attention(Q_c, K_c, V_c) = \text{softmax}\left(\frac{\mathbf{Q}_c\mathbf{K}_c^T}{\sqrt{d}}\right)\mathbf{V}_c. \quad (5)$$

To make it easier for explanation, we remove the softmax function and the scaling factor $\sqrt{d}$ in equation (5) in the following derivation, which in fact has no impact on the key idea of the proposed cross-attention mechanism. Thus, the derivation process can be expressed as

$$\begin{aligned}\mathbf{Q_c}\mathbf{K_c^T}\mathbf{V_c} &= \begin{pmatrix}\mathbf{Q}_1\\ \mathbf{Q}_2\end{pmatrix}(\mathbf{K_1^T}, \mathbf{K_2^T})\begin{pmatrix}\mathbf{V_1}\\ \mathbf{V_2}\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{Q_1}\mathbf{K_1^T}\mathbf{V_1} + \mathbf{Q_1}\mathbf{K_2^T}\mathbf{V_2}\\ \mathbf{Q_2}\mathbf{K_2^T}\mathbf{V_2} + \mathbf{Q_2}\mathbf{K_1^T}\mathbf{V_1}\end{pmatrix}. \quad (6)\end{aligned}$$

Then, we split the output A_c into A_1 for the first modality and A_2 for the second modality, in which

$$\mathbf{A}_1 = \mathbf{Q_1}\mathbf{K_1^T}\mathbf{V_1} + \mathbf{Q_1}\mathbf{K_2^T}\mathbf{V_2}, \quad (7)$$

$$\mathbf{A}_2 = \mathbf{Q_2}\mathbf{K_2^T}\mathbf{V_2} + \mathbf{Q_2}\mathbf{K_1^T}\mathbf{V_1}. \quad (8)$$

It is worth noting that the second and third terms in equation 7 and equation 8 actually represent the intra-modality knowledge and the inter-modality knowledge generated by the cross-attention mechanism, respectively. After that, both the A_1 and A_2 will be scaled by a coefficient α which is positive but less than 1 and then separately summed up with the original outputs produced by the multi-head attention operations in the first and second modalities for the subsequent feed-forward networks.

C. The Weight Control Module

In the CA-SignBERT framework, we introduce a new weight control module to adaptively adjust the combination weights of different input modalities according to three evaluation terms: the forecasted glosses, the sequential information and the intra-modality recognition loss.

1) The Forecasted Glosses: As presented in Fig. 1, the forecasted glosses control the combination weighted based on the preliminary recognition results from the feature extractors. Specifically, for a sequence of input data clips, we use a two-layer fully-connected network to forecast the meaning of each data clip based on the outputs from the feature extractors and generate the fusion weights accordingly. With this evaluation term, the weight control module can estimate the gloss of each data clip and arrange larger weights on the input modalities that are skilled in recognizing this gloss.

2) The Sequential Information: In this evaluation term, we employ multiple BLSTM layers to analyze the sequential relation among the data clips within each modality. As shown in Fig. 1, the output from each BERT model will be fed into an extra BLSTM layer for generating a value for each input data clip. The ratio of the values which have the same timestamp but from different input modalities will then be used as the combination weight for their corresponding data clips. Through this evaluation term, the weight control module can provide personal weight factors for the data clips at different timestamps based on their temporal dependency.

3) The Intra-Modality Recognition Loss: Different from other evaluation terms which use neural networks to generate a specific weight coefficient for each data clip, this evaluation term adjusts the combination weights iteratively based on the modality performance. As presented in Fig. 1, in each training iteration, we utilize the output from each information path to conduct CSLR independently and measure its CTC loss. After that, the inverse ratio of the CTC loss in each modality will be used as its combination weight in next iteration. The initial weight of each modality in the first iteration is set to be 1. In this way, the weight control module can then penalize the input modalities with poor performance and focus more on the one with high recognition accuracy.

D. Iterative Training

The significance of iterative training in CSLR has been verified by numerous works [11], [20], [21] as it can provide extra supervisions for the recognition models under the limited amount of data. Yet existing iterative training methods can only offer additional training to the feature extractors but not the BERT models [5], [20]. To address this issue, we employ a new masked iterative training (MAIT) for the CA-SignBERT.

In the MAIT, the framework is firstly trained end-to-end by minimizing the final CTC loss such that a sequence of glosses can be obtained from the last BLSTM layer. We consider these glosses as the pseudo labels of the input data clips and randomly mask one of them through replacing it with background noise. After that, the whole sequence of data clips (including the one being masked) is fed into different information paths separately such that the feature extractors and the BERT models can be trained by predicting the pseudo label of the masked data clip with cross-entropy loss.

The benefits of the MAIT are threefold. First, the MAIT can provide extra supervisions to train not only the feature extractors but also the BERT models. Second, as it does not know which data clip will be masked, the CA-SignBERT framework is forced to keep a distribution of contextual representation for each data clip and thus robust to the possible non-standard signs. Last but not least, we can mask different video clips iteratively such that more training data can be provided for the CA-SignBERT framework.

III. Experiments

In this section, we will firstly introduce some experimental setups and the four CSLR datasets on which we conduct extensive experiments. After that, we will analyze the performance of different groups of input modalities through an ablation study. Lastly, we will compare the performance of the CA-SignBERT framework with those of other existing approaches.

A. Datasets Information

In this letter, we conduct the experiments on four benchmark CSLR datasets which include the Chinese sign language (CSL) dataset [22], the RWTH-Phoenix-Weather-2014 (RWTH-2014) dataset [23], the Greek sign language (GSL) dataset [24] and the newly collected Hong Kong sign language (HKSL) dataset. The information of the first three datasets can be found in [22], [23] and [24], respectively. The HKSL dataset is newly introduced by us for facilitating the research in CSLR. It contains 50 Hong Kong sign language sentences performed by 6 signers with 8 repetitions. In each sentence, three types of input modalities are collected including RGB videos, depth videos and smart watch data.

TABLE I
THE PERFORMANCE OF DIFFERENT MODALITIES GROUPS

Modality Groups	WER(%)
RGB Only	5.94
RGB and Joint Coordinates	5.17
RGB and Depth	4.68
RGB and Hand Images	4.02

TABLE II
THE EXPERIMENTAL RESULTS ON THE CSL DATASET

Methods	SIT WER(%)	UST WER(%)
IAN [20]	6.10	32.70
SF-Net [28]	3.80	/
FCN [29]	3.00	/
SignBERT [11]	2.26	24.90
CMA [30]	/	24.50
VAC [31]	1.60	/
CA-SignBERT	**1.14**	**19.80**

TABLE III
THE EXPERIMENTAL RESULTS ON THE RWTH-2014 DATASET

Methods	Dev WER (%)	Test WER (%)
Self-Attention [32]	31.7	31.3
SBD-RL [33]	28.6	28.6
Re-Sign [34]	23.8	24.4
CrossModal [8]	23.9	24.0
VAC [31]	21.2	22.3
CMA [30]	21.3	21.9
SignBERT [11]	21.2	21.4
CA-SignBERT	**18.3**	**18.6**

TABLE IV
THE EXPERIMENTAL RESULTS ON THE GSL DATASET

Methods	GSL-SI		GSL-SD	
	Validation	Test	Validation	Test
SubUNets [35]	21.65	20.62	52.79	54.31
I3D [36]	6.63	6.10	49.89	49.99
CrossModal [8]	3.56	3.52	38.21	41.98
SLRGAN [37]	2.87	2.98	36.91	37.11
CA-SignBERT	**2.18**	**2.20**	**30.57**	**31.15**

TABLE V
THE EXPERIMENTAL RESULTS ON THE HKSL DATASET

Methods	SIT WER(%)	UST WER(%)
DenseTCN [38]	7.91	24.55
SF-Net [28]	7.69	22.08
FCN [29]	7.16	20.37
SignBERT [11]	6.10	15.92
CA-SignBERT (video and hand image)	**4.02**	**10.73**
CA-SignBERT (video and smart watch)	**2.63**	**7.19**

B. Experimental Setups

In the CA-SignBERT, we firstly adopt the frame selection mechanism introduced in [11] for selecting the key frames. These selected frames are then converted into video clips for the upcoming feature extraction through a sliding window of size 8 and stride 4. For the smart watch data in the HKSL dataset, we use the same starting and ending timestamps as the video clips for creating the smart watch data clips. To evaluate the performance of the framework precisely, we employ the word error rate (WER) as the main criterion [25].

C. The Ablation Study

In this ablation study, we compare the performance of four different groups of input modalities in the HKSL dataset, including the "RGB Only," "RGB and Depth," "RGB and Joint Coordinates" and "RGB and Hand Images". The RGB channel and the depth information are provided by the dataset while the joint coordinates and hand images are extracted from the RBG channel through OpenPose [26].

As shown in Table I, the "RGB and Hand images" attains the lowest WER of 4.02% among all the modality combinations. This is mainly because the hand images can provide the direct information of the most critical part in CSLR. Therefore, in the following experiments, we will use the extracted hand images as the second input modality by default. It is worth noting that the "RGB Only" performs the worst with a WER of 5.94%. This clearly demonstrates the importance of integrating additional modalities for CSLR.

D. A Comparison With State-of-The-Art Methods

In this subsection, we compare the performance of the CA-SignBERT framework with those of other existing approaches on the aforementioned CSLR datasets. All the experimental results are shown from Table II ?> to Table V.

Specifically, the CA-SignBERT outperforms the second-best approach by 0.46% in the signer independent test (SIT) [11] and 4.70% in the unseen sentence test (UST) [11] of the CSL dataset. Meanwhile, it also attains the lowest WER in both the validation set (18.3%) and test set (18.6%) of the RWTH-2014 dataset. In addition, it can be observed from Table IV that the CA-SignBERT surpasses other methods with a lowest WER of 2.20% in the GSL-SI test set and 31.15% in the GSL-SD test set. Lastly, as shown in Table V, the CA-SignBERT transcends other existing approaches by at least 2% in the SIT and 5% in the UST of the HKSL dataset.

In summary, it can be observed from Table II to Table V that the CA-SignBERT framework achieves better performance than all the existing methods on these benchmark CSLR datasets. This result strongly validates the effectiveness of the CA-SignBERT framework which integrates the knowledge from different input modalities with an innovate cross-attention mechanism and an adaptive weight control module.

IV. CONCLUDING REMARKS

In this work, we introduce a pioneering CA-SignBERT framework for CSLR with multiple input modalities. This framework contains an innovative cross-attention mechanism to effectively exchange the knowledge extracted from different modalities and a new weight control module to dynamically adjust their combination weights. Extensive experiments conducted on four benchmark CSLR datasets show that significantly lower word error rates are attained by the CA-SignBERT when compared to the results of other state-of-the-art approaches in CSLR.

More importantly, this work sheds lights on numerous directions for future investigation. First, it is valuable to explore the performance of adopting other intelligent algorithms such as the graph convolutional network [38] and the attributed network clustering [39] in the CA-SignBERT for CSLR. Moreover, it will be meaningful if future investigations may apply the techniques in the CA-SignBERT to other applications with multiple input sources, such as sentiment analysis [40], social network modelling [41] and knowledge tracing [42].

REFERENCES

[1] 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 *Proc. 22nd Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2020, pp. 1–4.

[2] Z. Wang et al., "Hear sign language: A real-time end-to-end sign language recognition system," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2398–2410, Jul. 2022.

[3] M. R. Abid, E. M. Petriu, and E. Amjadian, "Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 3, pp. 596–605, Mar. 2015.

[4] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," *IEEE Access*, vol. 9, pp. 126917–126951, 2021.

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

[6] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2306–2320, Sep. 2020.

[7] L. Myllyaho, J. K. Nurminen, and T. Mikkonen, "Node co-activations as a means of error detection–towards fault-tolerant neural networks," *Array*, vol. 15, 2022, Art. no. 100201. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000509

[8] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, "Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space," *IEEE Access*, vol. 8, pp. 91170–91180, 2020.

[9] A. A. Hosain, P. S. Santhalingam, P. Pathak, J. Košecká, and H. Rangwala, "Sign language recognition analysis using multimodal data," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics*, 2019, pp. 203–210.

[10] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, 2017.

[11] Z. Zhou, V. W. L. Tam, and E. Y. Lam, "SignBERT: A. BERT-based deep learning framework for continuous sign language recognition," *IEEE Access*, vol. 9, pp. 161669–161682, 2021.

[12] P. M. Ferreira, J. S. Cardoso, and A. Rebelo, "On the role of multimodal learning in the recognition of sign language," *Multimedia Tools Appl.*, vol. 78, pp. 10035–10056, 2019.

[13] C. Zhang, Y. Tian, and M. Huenerfauth, "Multi-modality american sign language recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 2881–2885.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, vol. 1, pp. 6000–6010.

[15] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2822–2832, Sep. 2019.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, Vol. 1, pp. 4171–4186.

[17] 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 *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 4296–4302.

[18] Z. Zhou, V. W. L. Tam, and E. Y. Lam, "A portable sign language collection and translation platform with smart watches using a BLSTM-based multi-feature framework," *Micromachines*, vol. 13, no. 2, 2022, Art. no. 333.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[20] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4160–4169.

[21] K. Koishybay, M. Mukushev, and A. Sandygulova, "Continuous sign language recognition with iterative spatiotemporal fine-tuning," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 10211–10218.

[22] J. Pu, "Chinese sign language recognition dataset," Accessed: Jun. 30, 2022. [Online]. Available: http://home.ustc.edu.cn/pjh/openresources/cslr-dataset-2015/index.html

[23] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Comput. Vis. Image Understanding*, vol. 141, pp. 108–125, 2015.

[24] N. Adaloglou et al., "A comprehensive study on sign language recognition methods," 2020, *arXiv:2007.12530*.

[25] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Commun.*, vol. 38, no. 1, pp. 19–28, 2002.

[26] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[27] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "Sf-net: Structured feature network for continuous sign language recognition," 2019, *arXiv:abs/1908.01341*.

[28] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 697–714.

[29] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1497–1505.

[30] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11522–11531.

[31] M. Zhou, M. K. Ng, Z. Cai, and K. C. Cheung, "Self-attention-based fully-inception networks for continuous sign language recognition," in *Proc. Eur. Conf. Artif. Intell.*, 2020, pp. 2832–2839.

[32] C. Wei, J. Zhao, W. Zhou, and H. Li, "Semantic boundary detection with reinforcement learning for continuous sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1138–1149, Mar. 2021.

[33] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3416–3424.

[34] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3075–3084.

[35] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[36] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, vol. 21, no. 7, 2021, Art. no. 2437.

[37] D. Guo, S. Wang, Q. Tian, and M. Wang, "Dense temporal convolution network for sign language translation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 744–750.

[38] Y. Yang, Z. Guan, J. Li, W. Zhao, J. Cui, and Q. Wang, "Interpretable and efficient heterogeneous graph convolutional network," *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2021.3101356.

[39] S. Yang et al., "Variational Co-embedding learning for attributed network clustering," 2021, *arXiv:abs/2104.07295*.

[40] H. Yin, X. Song, S. Yang, and J. Li, "Sentiment analysis and topic modeling for COVID-19 vaccine discussions," *World Wide Web*, vol. 25, pp. 1067–1083, 2022.

[41] N. A. H. Haldar et al., "Top-k socio-spatial co-engaged location selection for social users," *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2022.3151095.

[42] X. Song, J. Li, Q. Lei, W. Zhao, Y. Chen, and A. S. Mian, "Bi-clkt: Bi-graph contrastive learning based knowledge tracing," *Knowl. Based Syst.*, vol. 241, 2022, Art. no. 108274.