# SignVTCL: Multi-Modal Continuous Sign Language Recognition Enhanced by Visual-Textual Contrastive Learning

Hao Chen[1*], Jiaze Wang[1*], Jinpeng Li[1], Ziyu Guo[1], Donghao Zhou[1], Bian Wu[2], Chenyong Guan[3], Guangyong Chen[2†], Pheng-Ann Heng[1]

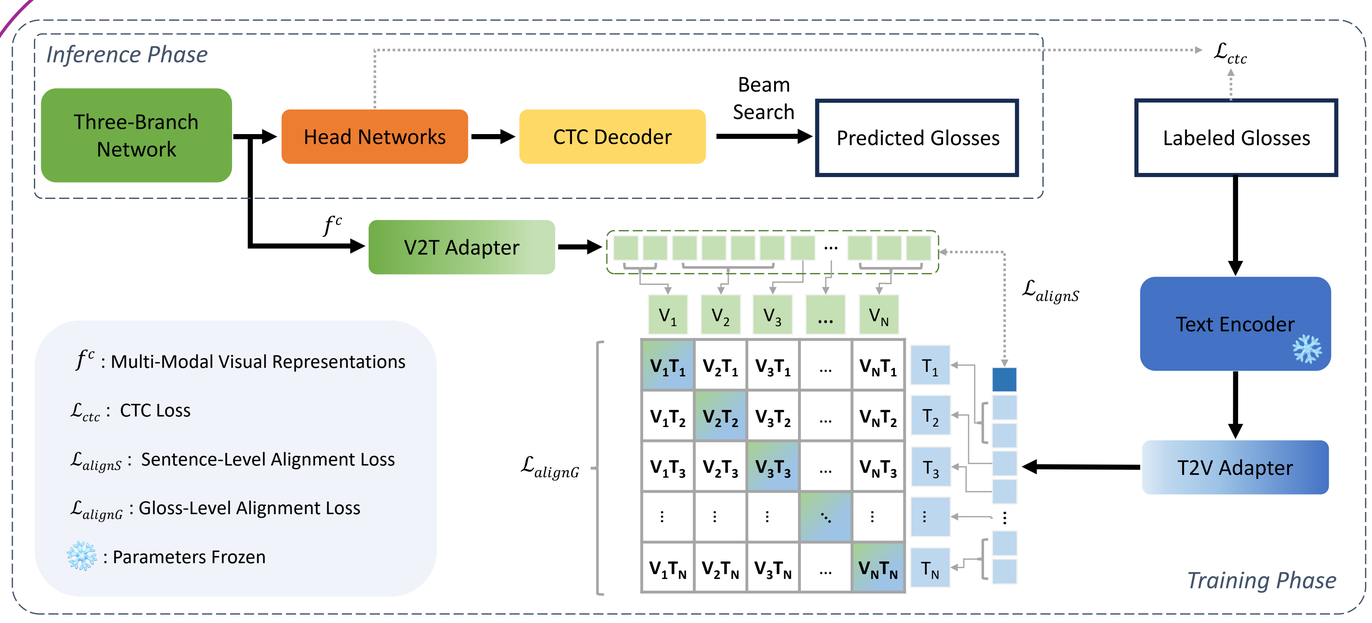
[1] The Chinese University of Hong Kong [2] Zhejiang Lab [3] Gudsen Technology Co. Ltd

## Introduction

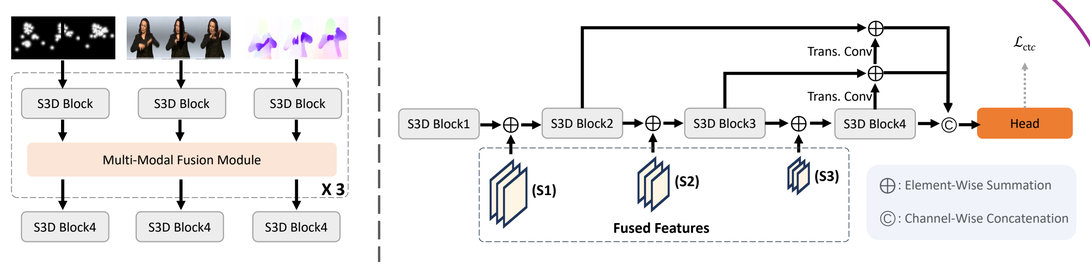
Sign language recognition (SLR) plays a vital role in facilitating communication for the hearing-impaired community. A significant challenge in SLR arises from its weakly supervised nature, where each entire video is annotated with a sequence of glosses. This makes it particularly difficult to accurately identify the corresponding gloss for specific video segments. Therefore, we proposed **SignVTCL** to address this challenge:

1. We effectively utilize video, keypoints, and optical flow modalities together to train a unified visual backbone, resulting in more robust visual representations. 2. We present a visual-textual alignment approach to align textual and visual features from the gloss level and sentence level, thereby enhancing the accuracy of SLR. 3. The proposed SignVTCL achieves state-of-the-art results across multiple datasets, including Phoenix-2014, Phoenix-2014T and CSL-Daily.

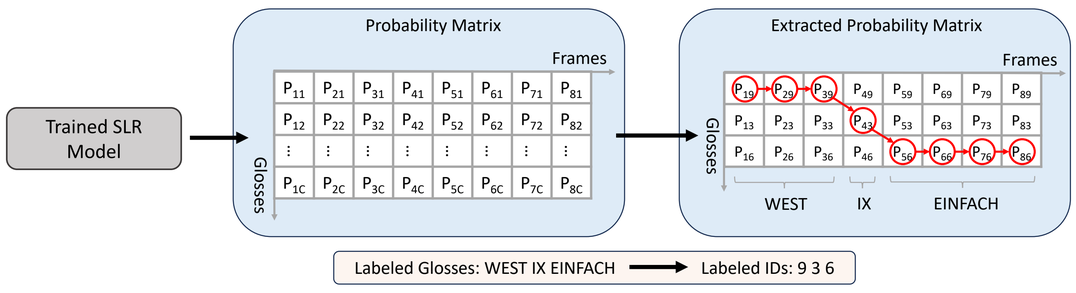
## Method




(a)




(b)




(c)

a) The Pipeline of SignVTCL. The three-branch network aims to extract visual representations from three different modalities. These features are then passed through head networks for predicting frame-wise gloss probabilities. Simultaneously, we input labeled glosses into a frozen pretrained text encoder to obtain textual representations. Then, the visual and textual representations are aligned within a joint multi-modal semantic space using two adapters: the V2T adapter and the T2V adapter. During the inference phase, a CTC decoder is employed to generate glosses based on the predicted gloss probabilities.

a) The architecture of three-branch network and details of each branch. c) Finding the frames corresponding to each gloss. The guiding principle for determining the path with the highest probability is as follows: commencing from the top-left corner of the matrix, movement is restricted to either downward or towards the lower-right position.

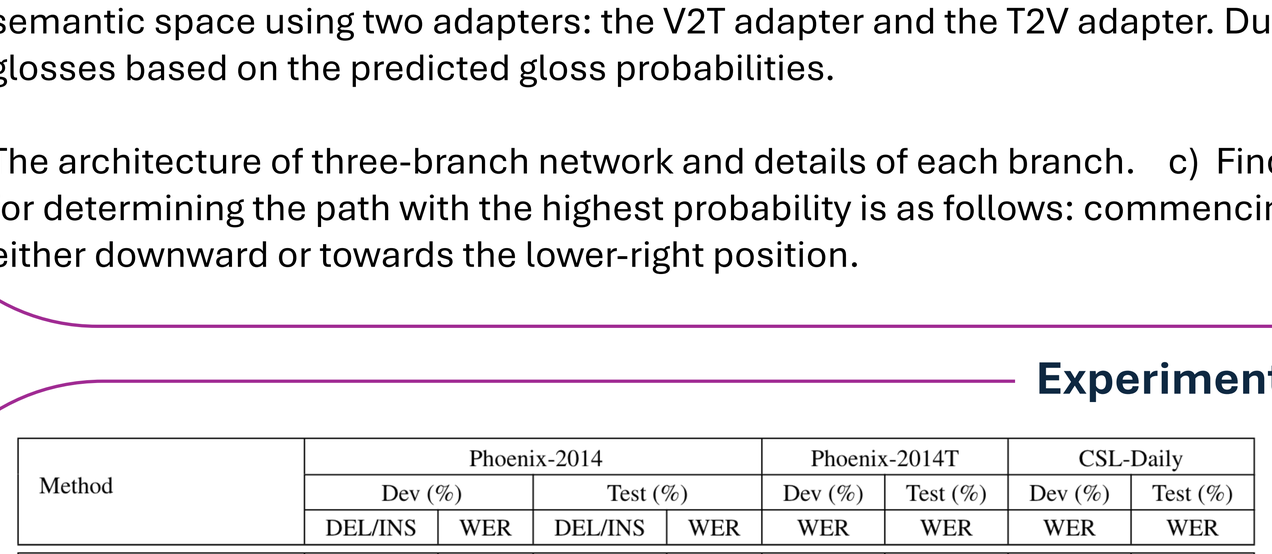
## Experiments

| Method | Phoenix-2014 | | | | Phoenix-2014T | | CSL-Daily | |
|---|---|---|---|---|---|---|---|---|
| | Dev (%) | | Test (%) | | Dev (%) | Test (%) | Dev (%) | Test (%) |
| | DEL/INS | WER | DEL/INS | WER | WER | WER | WER | WER |
| SubUNets [10] | 14.6/4.0 | 40.8 | 14.3/4.0 | 40.7 | - | - | 41.4 | 41.0 |
| CNN-LSTM-HMMs [22] | - | 26.0 | - | 26.0 | 24.1 | 26.1 | - | - |
| DNF [11] | 7.3/3.3 | 23.1 | 6.7/3.3 | 22.9 | - | - | 32.8 | 32.4 |
| SFL [29] | 7.9/6.5 | 26.2 | 7.5/6.3 | 26.8 | - | - | - | - |
| FCN [7] | - | 23.7 | - | 23.9 | 23.3 | 25.1 | 33.2 | 33.5 |
| Joint-SLRT [3] | - | - | - | - | 24.6 | 24.5 | 33.1 | 32.0 |
| VAC [28] | 7.9/2.5 | 21.2 | 8.4/2.6 | 22.3 | - | - | - | - |
| SignBT [43] | - | - | - | - | 22.7 | 23.9 | 33.2 | 33.2 |
| SMKD [16] | 6.8/2.5 | 20.8 | 6.3/2.3 | 21.0 | 20.8 | 22.4 | - | - |
| STMC-R [44] | 7.7/3.4 | 21.1 | 7.4/2.6 | 20.7 | 19.6 | 21.0 | - | - |
| MMTLB [5] | - | - | - | - | 21.9 | 22.5 | - | - |
| TLP [17] | 6.3/2.8 | 19.7 | 6.1/2.9 | 20.8 | 19.4 | 21.2 | - | - |
| $C^2$SLR [45] | - | 20.5 | - | 20.4 | 20.2 | 20.4 | - | - |
| TwoStream-SLR [6] | - | 18.4 | - | 18.8 | 17.7 | 19.3 | 25.4 | 25.3 |
| CorrNet [18] | **5.6**/2.8 | 18.8 | **5.7**/2.3 | 19.4 | 18.9 | 20.5 | 30.6 | 30.1 |
| CVT-SLR [41] | 6.4/2.6 | 19.8 | 6.1/2.3 | 20.1 | 19.4 | 20.3 | - | - |
| SEN [19] | 5.8/2.6 | 19.5 | 7.3/4.0 | 21.0 | 19.3 | 20.7 | 31.1 | 30.7 |
| **SignVTCL (Ours)** | 5.8/**2.4** | **17.1** | **5.7**/2.2 | **17.4** | **16.7** | **17.7** | **24.2** | **24.0** |

(a)

| Group | V | K | O | S1 | S2 | S3 | MLP | Attn | Conv | AlignG | AlignS | Dev (%) | Test (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ✓ | | | | | | | | | | | 20.1 | 20.5 |
| | | ✓ | | | | | | | | | | 27.0 | 26.2 |
| | | | ✓ | | | | | | | | | 38.9 | 38.1 |
| 2 | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | | | 19.0 | 19.2 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | | | | | 18.7 | 19.0 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | 18.3 | 18.6 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | | | | 18.6 | 19.0 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | 18.4 | 18.8 |
| | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | | | | 19.0 | 19.3 |
| | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | 19.6 | 19.9 |
| | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | 25.5 | 25.2 |
| 3 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | 17.5 | 17.8 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ | 17.7 | 17.9 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | **17.1** | **17.4** |
| | ✓ | | | | | | | | | ✓ | ✓ | 19.2 | 19.4 |
| | | ✓ | | | | | | | | ✓ | ✓ | 25.1 | 24.7 |
| | | | ✓ | | | | | | | ✓ | ✓ | 36.0 | 35.8 |
| | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | 17.9 | 18.4 |
| | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | 18.3 | 18.8 |
| | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | 24.5 | 24.1 |

(b)




(c)

a) Comparison with Previous Methods on the Phoenix-2014, Phoenix-2014T and CSL-Daily Datasets.
b) Ablation study of each component of SignVTCL on the Phoenix-2014. 'V', 'K', 'O' are video, keypoint and optical flow branches, respectively. 'S1' ~ 'S3' represent three fusion stages in three-branch network. 'MLP', 'Attn' and 'Conv' are different fusion methods. 'AlignG' and 'AlignS' denote gloss-level and sentence-level alignments.
c) Visualization of Gloss-Level Alignment Similarity Matrix and Predicted Glosses.

## Conclusion

In this paper, we propose SignVTCL, a multi-modal continuous sign language recognition framework enhanced by visual-textual contrastive learning. By combining video, keypoints, and optical flow modalities, SignVTCL can capture the intricate hand gestures and dynamic body movements of signers with greater accuracy, yielding more robust visual features that enhance the model's comprehension of sign language. Additionally, we introduce a visual-textual alignment approach that aligns visual and textual features at both the gloss and sentence level, ensuring a meaningful and precise correspondence between visual signs and textual context, which enhances the performance of sign language recognition. Comprehensive experiments on multiple datasets confirm the effectiveness of SignVTCL in achieving state-of-the-art performance.