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

Hao Chen[1*]
hchen22@link.cuhk.edu.hk
Jiaze Wang[1*]
jzwang@link.cuhk.edu.hk
Jinpeng Li[1]
1155074899@link.cuhk.edu.hk
Ziyu Guo[1]
2101210573@pku.edu.cn
Donghao Zhou[1]
dhzhou@link.cuhk.edu.hk
Bian Wu[2]
wub@zhejianglab.com
Chenyong Guan[3]
cy@gudsen.com
Guangyong Chen[2†]
gychen@zhejianglab.com
Pheng-Ann Heng[1]
pheng@cse.cuhk.edu.hk

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

Abstract

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. To address this challenge, we present **SignVTCL**, a multi-modal continuous sign language recognition framework enhanced by visual-textual contrastive learning, which leverages the full potential of multi-modal data and the generalization ability of language model. First, SignVTCL consolidates multi-modal data to train a unified visual feature extractor, resulting in more robust visual representations. Subsequently, it employs a visual-textual alignment approach that integrates gloss-level and sentence-level alignments, establishing precise correspondences between visual and textual features to enhance SLR accuracy. Experimental results conducted

*Equal contribution
†Corresponding author

on three datasets, Phoenix-2014, Phoenix-2014T, and CSL-Daily, demonstrate that SignVTCL achieves state-of-the-art results compared with previous methods. Project page: https://jiazewang.com/projects/signvtcl.html.

1 Introduction

Videos provide a remarkably faithful representation of how humans consistently perceive the visual world. Consequently, the ability to comprehend videos is of paramount importance for artificial intelligence systems to acquire a profound understanding of the human world. This has positioned video understanding [25, 27] as the next frontier in the field of computer vision. Sign language recognition (SLR) is a exceptionally challenging task in video understanding, primarily because it requires the interpretation of intricate and precise semantic information in sign language videos. However, study of SLR is meaningful and impactful as it serves as a highly inclusive means of communication for the hearing-impaired community, effectively bridging the gap between deaf and hearing individuals. In this work, we focus on studying continuous sign language recognition (CSLR) which aims to transcribe co-articulated sign videos into sign sequences on a gloss-by-gloss basis. In the subsequent discussion, we adopt the abbreviation SLR to refer to CSLR.

Since SLR is a weakly supervised task that only annotates the sign video with a sequence of glosses, existing methods [6, 17, 19, 41, 45] face the challenge of determining the corresponding gloss for each video segments from a video. From this perspective, we propose that the accuracy of SLR can be improved by aligning the textual features of each gloss with its corresponding visual features. We employ a reliable pre-trained language model to convert glosses into textual features, necessitating the training of an additional visual feature extractor to obtain the visual features. Recently, many approaches [7, 10, 18] for SLR extract visual features solely from videos. However, these methods often face difficulties in effectively addressing the inherent complexities of sign language, such as the variations in signing styles among individuals and the intricate movements of dynamic body parts. Consequently, in this paper, we focus on utilizing multi-modal data. We extract various types of modalities from videos, including keypoints and optical flow. These sources provide diverse information that enriches the model's comprehension of sign language, leading to the development of more robust visual representations.

After obtaining the textual and visual features, we propose a visual-textual alignment approach to align these features at both the gloss and sentence levels. In gloss-level alignment, we initially identify the video segment corresponding to each gloss. Subsequently, we align the textual feature of the gloss with the visual feature of the video segment in a high-dimensional space. While sentence-level alignment aims to force the model to contain a holistic understanding of the semantic and contextual information within the sentence. Therefore, the visual-textual alignment approach can establish a potential correspondence between visual signs and textual context, further enhancing the SLR accuracy.

Finally, by integrating the aforementioned thoughts, we present SignVTCL, a multi-modal continuous sign language recognition framework enhanced by visual-textual contrastive learning. To validate the effectiveness of our proposed method, we conduct extensive experiments on Phoenix-2014 [21], Phoenix-2014T [2] and CSL-Daily [43]. The results demonstrate that our approach achieves state-of-the-art performance in SLR. Our main contributions can be summarized as follows:

- We effectively utilize video, keypoints, and optical flow modalities together to train a

unified visual backbone, resulting in more robust visual representations.

- 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.
- The proposed SignVTCL achieves state-of-the-art results across multiple datasets, including Phoenix-2014, Phoenix-2014T and CSL-Daily.

2 Related Works

Sign Language Recognition. Deep learning-based methods have significantly transformed the domain of SLR, succinctly characterized by two critical phases: visual feature extraction and recognition. Predominantly, 3D CNNs [5, 6, 24, 30, 33, 34, 44] have gained widespread adoption for feature extraction. Additionally, certain approaches [4, 11, 18, 28, 40, 44] opt to commence with a 2D CNN to extract frame-wise features before subsequently incorporating hybrid architectures composed of 1D CNNs and Long Short-Term Memory (LSTM) [14] networks to capture temporal dependencies. Upon deriving features, classifiers can compute posterior probabilities to facilitate the recognition process. In addition to utilizing videos for SLR, keypoints [8, 9] and optical flow [11, 32] can be employed as auxiliary modal information to enhance the performance of SLR. Keypoints offer specific details about manual and non-manual elements. Optical flow, on the other hand, provides information about the movement of human body parts between consecutive frames of a video. In this study, we introduce a unified architecture that harnesses the information from three modalities (video, keypoints, optical flow) simultaneously to extract visual representations. By integrating these modalities, we aim to enhance the capability of SLR by capturing a more comprehensive understanding of sign language.

Visual-Textual Contrastive Learning. Recently, there has been a growing trend in developing visual-textual approaches for visual problems [20, 31, 35]. Inspired by this concept and recognizing the inherent video-text pair structure in SLR data, we have discovered that learning language-indicated visual representations from sign language videos is an effective approach to enhancing SLR performance. Few recent approaches [8, 12, 36, 41, 42] have considered visual-textual contrastive learning in the field of sign language. Zheng *et al.* [41] proposed a visual-textual transformation-based SLR framework that leverages the unique properties of autoencoders to implicitly align the visual and textual modalities. Zhou *et al.* [42] integrated contrastive language-image pre-training with masked self-supervised learning to create pre-tasks that bridge the semantic gap between visual and textual representations. Both approaches utilize contrastive learning at the macro level, encompassing the entire videos and complete sentences within a mini-batch, necessitating an additional pre-training phase. However, we advance such methods by delving into the semantic alignment between distinct video segments and their corresponding glosses in the sentence, while eliminating the need for pre-training.

3 Methods

As shown in Figure 1, SignVTCL consists of a multi-modal visual backbone (three-branch network and head networks), a text encoder, two feature adapters (V2T adapter and T2V adapter), and a CTC decoder. The task of SLR is to translate the input video $x^v = \{x_t^v\}_{t=1}^T \in \mathcal{R}^{T\times H\times W\times 3}$ into a series of glosses $y = \{y_i\}_{i=1}^N$ to express a sentence, with N denoting the

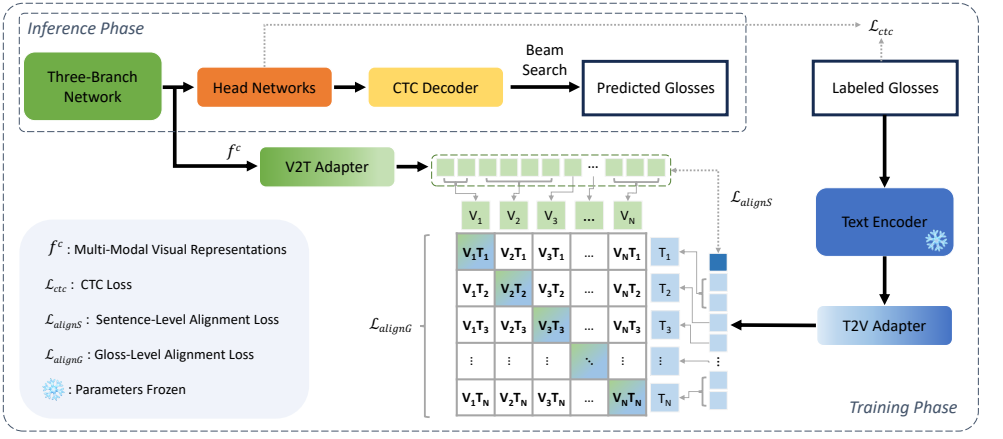

Figure 1: **The Pipeline of SignVTCL.** The three-branch network aims to extract visual representations from three different modalities. These features are then passed through head networks for predicting frame-wise gloss probabilities. Simultaneously, we input labeled glosses into a frozen pretrained 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.

length of the glosses sequence. In this section, we begin by introducing the multi-modal visual backbone. Then, we explain the utilization of a text encoder and two adapters to align visual and textual representations within a shared multi-modal semantic space. Finally, we delve into the training and inference processes of SignVTCL.

3.1 Multi-Modal Visual Backbone

Three-Branch Network. To enhance the quality of visual features, we employed a strategy that involves extracting keypoints and optical flow from videos. These additional modalities are incorporated alongside the video itself to jointly train a three-branch network, forming a multi-modal training setup. The sequences of keypoints are denoted as $x^k = \{x_t^k\}_{t=1}^T \in \mathcal{R}^{T\times H\times W\times K}$, K is the keypoints number. Additionally, optical flow is processed and stored as a sequence of images, defined as $x^o = \{x_t^o\}_{t=1}^T \in \mathcal{R}^{T\times H\times W\times 3}$. The methods used to extract keypoints and optical flow from the video are detailed in the supplementary materials.

The architecture of three-branch network is shown in Figure 2 (a). Each branch comprises the first four blocks of S3D [58] and detailed in Figure 2 (b). Between each block, a multi-modal fusion module is incorporated to effectively merge information from different modalities. The input of this module is the concatenation of hidden features from three modalities, while the output is the fused feature, which maintains the same size as the input size of the hidden features from three modalities. The fused feature will be respectively added to the three input modality features to complete the fusion process. We studied different fusion methods, including MLP-based, attention-based, and convolution-based methods, and our experiments demonstrate that the MLP-based method yields the best results. Details of these fusion methods are placed in the supplementary materials.

Three modalities data will be processed by their corresponding branch into frame-wise

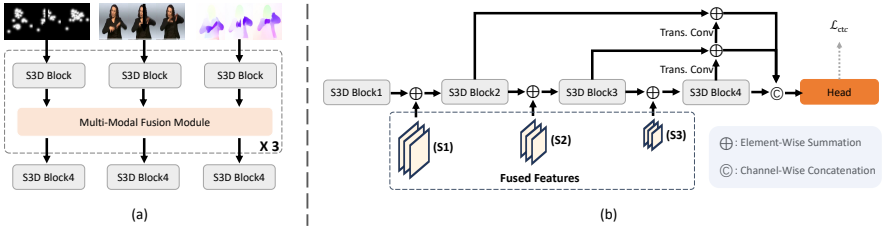

Figure 2: **Illustration of Three-Branch Network.** (a) The architecture of three-branch network. (b) Details of each branch. Herein, the fused feature are the outputs of three multi-modal fusion modules and 'S1' $\sim$ 'S3' represent three fusion stages. The 'Head' is the component of 'Head Networks' and is depicted here for ease of understanding.

features $f^{\{v,k,o\}} = \left\{f_t^{\{v,k,o\}}\right\}_{t=1}^{T/4} \in \mathcal{R}^{T/4\times d_v}$, which will be individually fed into separate head in our head networks, where d_v is the visual hidden dimension.

Head Networks. We adopt separate temporal heads for video, keypoint, and optical flow branches. To fully harness the potential of our three-branch architecture, we additionally combine the outputs of the video, keypoint, and optical flow branches. The combined representation $f^c = \{f_t^c\}_{t=1}^{T/4} \in \mathcal{R}^{T/4\times(3\times d_v)}$ will be then fed into a joint temporal head, which shares the same architecture as the individual temporal head of the video, keypoint, and optical flow. Each head is composed of a temporal linear layer, two temporal convolutional layers employing a kernel size of 3 and a stride of 1, and a linear layer as the classifier. We forward $f^{\{v,k,o,c\}}$ into four separate temporal heads, which will output frame-wise gloss probabilities $p^{\{v,k,o,c\}} = \left\{p_t^{\{v,k,o,c\}}\right\}_{t=1}^{T/4} \in \mathcal{R}^{T/4\times C}$, C is the size of the gloss vocabulary. Finally, $p^{\{v,k,o,c\}}$ will be used to compute CTC losses [15].

3.2 Visual-Textual Alignment

In this section, we introduce our visual-textual alignment approach to align visual and textual feature embeddings in a joint multi-modal semantic space. This approach establishes a potential correspondence between visual signs and textual context.

Text Encoder. To ensure the efficient encoding of textual data, the selection of a reliable text encoder holds paramount importance. Therefore, we opt for the encoder derived from mBART [26], which has undergone pre-training on CC25 [26], an extensive multilingual corpus incorporating 25 languages. Given a sentence with N glosses, it is first input to a tokenizer that converts the raw text to N_1 tokens. Then, these tokens are fed into the text encoder to obtain high-dimensional textual features $f^t = \{f_n^t\}_{n=1}^{N_1} \in \mathcal{R}^{N_1\times d_t}$, where d_t is the textual hidden dimension.

Feature Adapters. Inspired by [13, 39], we employ two lightweight adapters, a video-to-text (V2T) adapter and a text-to-video (T2V) adapter, to establish the connections between visual and textual features. The V2T adapter comprises two MLPs, with each MLP consisting of two hidden layers. One is dedicated to gloss-level alignment, while the other handles sentence-level alignment. The input of these two MLPs is the same visual feature f^c, and outputs are $f^{\{c_1,c_2\}} = \left\{f_t^{\{c_1,c_2\}}\right\}_{t=1}^{T/4} \in \mathcal{R}^{T/4\times d_j}$, d_j is the joint hidden dimension. The T2V adapter is also implemented as an MLP with two hidden layers. We forward f^t to this adapter

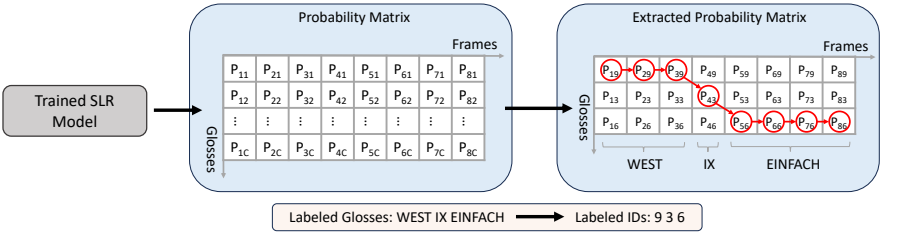

Figure 3: **Finding the frames corresponding to each gloss.** Assuming a vocabulary size of C glosses, each gloss can be represented by an 'ID' starting from 1. In the probability matrix, each column adds up to 1. In the extracted probability matrix, the guiding principle for determining the path with 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. This constraint is imposed to maintain the correspondence between the order of video frames and labeled glosses.

to obtain $f^{t_1} = \left\{f_n^{t_1}\right\}_{n=1}^{N_1} \in \mathcal{R}^{N_1 \times d_j}$. Thanks to these two adapters, we can align visual and textual representations within a unified multi-modal semantic space.

Gloss-Level Alignment. In gloss-level alignment, we need to align the textual features and visual features corresponding to each gloss. For textual features, token-wise features can be merged based on the labeled glosses, since we can know the specific tokens to which each gloss is converted. We use the local average pooling to obtain $f^{t_{11}} = \left\{f_n^{t_{11}}\right\}_{n=1}^{N} \in \mathcal{R}^{N \times d_j}$ from f^{t_1}. When it comes to visual features, a key challenge lies in assigning frame-wise features to the correct gloss. Drawing inspiration from the work of [37], which utilizes the dynamic time warping (DTW) algorithm [1] for dictionary construction, we employ the DTW algorithm for aligning frame-wise features to glosses in hidden space. For details, first, a trained SLR model is utilized to generate a frame-wise probability matrix. This matrix represents the probability of each frame belonging to each gloss. Then, we extract the rows corresponding to the labeled glosses one by one to construct an extracted probability matrix. Subsequently, a path is traced starting from the first position of the extracted probability matrix, aiming to identify the path with the highest multiplication of probability, as shown in Figure 3. By following this path, the frame-wise visual features associated with each gloss can be determined. Utilizing this prior knowledge, we can employ local average pooling to obtain $f^{c_{11}} = \left\{f_n^{c_{11}}\right\}_{n=1}^{N} \in \mathcal{R}^{N \times d_j}$ from f^{c_1}. Finally we can calculate the pair metrics as:

$$\begin{aligned} V2T_{pair} &= f^{c_{11}} \times (f^{t_{11}})^T, \\ T2V_{pair} &= f^{t_{11}} \times (f^{c_{11}})^T, \end{aligned} \tag{1}$$

where $\{V2T_{pair}, T2V_{pair}\} \in \mathcal{R}^{N \times N}$. Taking $V2T_{pair}$ as an example, $V2T_{pair}^{i,j}$ represents the similarity of the i-th frame-wise visual feature and the j-th token-wise textual feature. The visual and textual features from the same gloss can be a positive sample, and the features from different glosses are negative samples.

Sentence-Level Alignment. The objective of sentence-level alignment is to integrate and align visual features with textual features at the sentence level. In this process, a global textual feature $f^{t_{12}} \in \mathcal{R}^{1 \times d_j}$ that obtain from f^{t_1} at the position of <EOS> token is used to encapsulate the information of the entire sentence. On the other hand, we employ a global pooling layer on f^{c_2} to obtain the global visual feature $f^{c_{21}} \in \mathcal{R}^{1 \times d_j}$. Aligning these two

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 [24]	-	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 [8]	-	23.7	-	23.9	23.3	25.1	33.2	33.5
Joint-SLRT [8]	-	-	-	-	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 [42]	7.7/3.4	21.1	7.4/2.6	20.7	19.6	21.0	-	-
MMTLB [6]	-	-	-	-	21.9	22.5	-	-
TLP [14]	6.3/2.8	19.7	6.1/2.9	20.8	19.4	21.2	-	-
C^2SLR [35]	-	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**

Table 1: **Comparison with Previous Methods on the Phoenix-2014, Phoenix-2014T and CSL-Daily Datasets.** 'DEL' and 'INS' are average deletion and insertion rates, respectively.

features at the sentence level enables a holistic understanding of the semantic and contextual information conveyed by both visual and textual modalities within the sentence.

3.3 Training and Inference

Training Losses. First, we apply CTC losses [15] on the output of four temporal heads in head networks $\mathcal{L}_{ctc}^{v}$, $\mathcal{L}_{ctc}^{k}$, $\mathcal{L}_{ctc}^{o}$ and $\mathcal{L}_{ctc}^{j}$ (for joint head, detailed in Section 3.1). We sum them together to obtain $\mathcal{L}_{ctc}$.

Then, we compute the contrastive alignment losses, which consist of two parts, the gloss-level alignment loss $\mathcal{L}_{alignG}$ and the sentence-level alignment loss $\mathcal{L}_{alignS}$. Given $V2T_{pair}$ and $T2V_{pair}$, the ground truth of them can be denoted as $G \in \mathcal{R}^{N\times N}$, $G^{i,j}$ is equal to 1 when the i-th frame-wise visual feature and the j-th token-wise textual feature are from the same gloss, and otherwise 0. More details about ground truth G are placed in the supplementary materials. Therefore, the gloss-level alignment loss can be calculated as:

$$L_{alignG} = \frac{1}{2}\Big(\mathrm{CE}\big(\mathrm{SoftMax}(V2T_{pair}) \times G_c, G\big) + \mathrm{CE}\big(\mathrm{SoftMax}(T2V_{pair}) \times G_c, G\big)\Big), \quad (2)$$

where CE refers to the CrossEntroy loss, $G_c \in \mathcal{R}^{N\times 1}$ is a counter for counting the number of 1 in each line of G. The sentence-level alignment loss $\mathcal{L}_{alignS}$ is implemented by the KL divergence loss [23].

Finally, we weighted sum the mentioned losses to obtain the total training loss, which can be represented as:

$$\mathcal{L}_{total} = \lambda_{ctc}\mathcal{L}_{ctc} + \lambda_g\mathcal{L}_{alignG} + \lambda_s\mathcal{L}_{alignS}. \quad (3)$$

Inference. In the inference stage, we employ a parameter-free CTC decoder to derive the final gloss predictions using the beam search algorithm [15] with a beam width of 5. Specifically, the input is the average of frame-wise gloss probabilities $p^{\{v,k,o,c\}}$ that output from four separate temporal heads.

4 Experiments

4.1 Experimental Setup

Datasets. Phoenix-2014 [21] and Phoenix-2014T [2] are two German sign language datasets widely used in the field of SLR. The Phoenix-2014 dataset consists of 5672 training, 540 development, and 629 testing samples, with a vocabulary size of 1295 glosses. On the other hand, Phoenix-2014T is an extension of Phoenix-2014, containing 7096 training, 519 development, and 642 testing samples and has a vocabulary size of 1085 glosses. CSL-Daily [43] is a large-scale Chinese sign language dataset. It comprises 18401 training, 1077 development, and 1176 testing videos. This dataset features a vocabulary size of 2000 glosses.
Evaluation Metrics. The Word Error Rate (WER) stands as the predominant metric for assessing the performance of SLR. It quantifies the essential insertions (#ins), substitutions (#sub), and deletions (#del) required to align predicted sentences with their corresponding reference sentences (#reference). The lower WER, the better accuracy.

$$\text{WER} = \frac{\#ins + \#sub + \#del}{\#reference}. \tag{4}$$

4.2 Comparison with State-of-the-art Methods

As shown in Table 1, we conducted a comparative analysis between the proposed SignVTCL and existing state-of-the-art methods. Our approach surpassed all others, achieving top performance across three datasets. The WER scores achieved by SignVTCL on the development set of Phoenix-2014 and Phoenix-2014T are 17.1% and 16.7%, respectively. While on the test set, the optimal WER scores achieved by SignVTCL outperform the previous best method by 1.4% on and 1.6%, respectively. Also, we can see SignVTCL achieves a very low percentage of deletion and insertion compared with other methods. We also present a comparative analysis between our SignVTCL and the previous state-of-the-art methods on the CSL-Daily dataset. Our model exhibits a notable reduction in WER by 1.2% on the development set and 1.3% on the test set when compared to the previous best method, showcasing its superior performance.

4.3 Ablation Studies

Study on Each Modality Data. In Group 1 of Table 2, we evaluated the performance of SLR by individually utilizing the video, keypoint, and optical flow branches. Using a single branch means relying solely on data from one modality. Notably, video yielded the most favorable outcomes, while the optical flow demonstrated comparatively less effectiveness.
Study on Fusion Stages and Methods. We also investigate the efficacy of the three fusion stages in Group 2 of Table 2. It become evident that each fusion stage contributed to a reduction in SLR error rates. When all of them employed in tandem, a notable improvement was observed, resulting in WER of 18.3% on the development set and 18.6% on the test set. Furthermore, we conducted a study of various fusion methods. Our findings revealed that the MLP-based fusion method outperformed both the attention-based and convolution-based methods. Finally, we also explored the effects of fusion between any two modalities, with the results displayed in the last three rows of Group 2. These results not only demonstrate the effectiveness of our multi-modal fusion stages and method but also provide a foundation for

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

Table 2: **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.

Group 3 to verify the benefits of applying our proposed visual-textual alignment approach across any two modalities.

Study on Visual-Textual Alignment. In Group 3 of Table 2, we initially examine the effectiveness of gloss-level and sentence-level alignments for SignVTCL, with respective results presented in the first and second rows of this group. When these alignments are combined for training, SignVTCL yield optimized WER of 17.1% on the development set and 17.4% on the test set. We also investigate the impact of visual-textual alignment on each individual modality data as well as on combinations of any two modalities. The observed reduction in WER confirms that visual-textual alignment significantly enhances SLR performance.

4.4 Visualizations

To visualize the similarity matrix and predicted glosses, we selected examples from three datasets in Figure 4. Based on the predicted glosses compared to their reference glosses, it is apparent that the reference and inferred glosses exhibit a high degree of similarity, leading to relatively perfect performance. Notably, the highlighted regions in the alignment matrices predominantly concentrate in proximity to the diagonal. This concentration serves as a clear indication that our gloss-level alignment plays a pivotal role in efficiently aligning visual and textual features, thereby enhancing SLR performance. More visualization examples can be seen in the supplementary materials.

5 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,

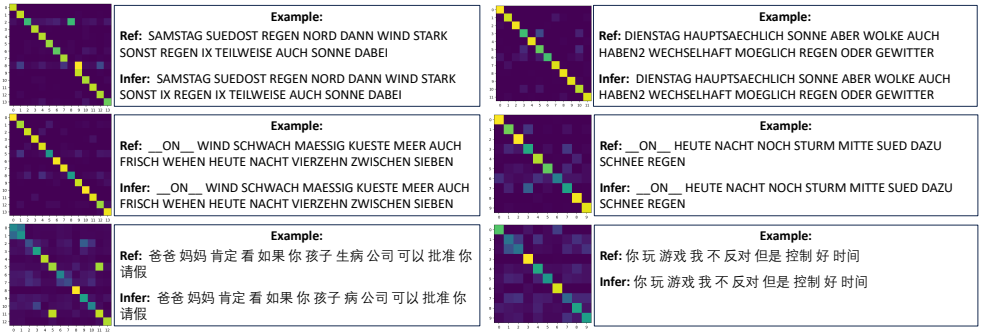

Figure 4: **Visualization of Gloss-Level Alignment Similarity Matrix and Predicted Glosses.** Each value in the similarity matrix corresponds to the similarity observed between visual and textual representations from different glosses.

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. We believe that SignVTCL will inspire future research to explore multi-modality and contrastive learning in video understanding tasks.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFE0200700), Hong Kong Innovation and Technology Fund (Project No. MHP/086/21), the National Natural Science Foundation of China (Project No. 6237073934, 3234101132), and Natural Science Foundation of Guangdong Province (2022A1515011579).

References

[1] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pages 359–370, 1994.

[2] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018.

[3] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020.

[4] Hao Chen, Jiaze Wang, Kun Shao, Furui Liu, Jianye Hao, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. Traj-mae: Masked autoencoders for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8362, 2023.

[5] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5130, 2022.

[6] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022.

[7] Ka Leong Cheng, Zhaoyang Yang, Qifeng Chen, and Yu-Wing Tai. Fully convolutional networks for continuous sign language recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 697–714. Springer, 2020.

[8] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026, 2023.

[9] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022.

[10] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3056–3065, 2017.

[11] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21 (7):1880–1891, 2019.

[12] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Kang Xia, Lei Xie, and Sanglu Lu. Contrastive learning for sign language recognition and translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 763–772, 2023.

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023.

[14] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[16] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11303–11312, 2021.

[17] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Temporal lift pooling for continuous sign language recognition. In *European Conference on Computer Vision*, pages 511–527. Springer, 2022.

[18] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2529–2539, 2023.

[19] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 854–862, 2023.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[21] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[22] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard 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*, 42(9):2306–2320, 2019.

[23] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[24] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33: 12034–12045, 2020.

[25] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[26] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 726–742, 2020.

[27] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[28] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11542–11551, 2021.

[29] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 172–186. Springer, 2020.

[30] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4165–4174, 2019.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[33] Jiaze Wang, Xiaojiang Peng, and Yu Qiao. Cascade multi-head attention networks for action recognition. *Computer Vision and Image Understanding*, 192:102898, 2020.

[34] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021.

[35] Jiaze Wang, Yi Wang, Ziyu Guo, Renrui Zhang, Donghao Zhou, Guangyong Chen, Anfeng Liu, and Pheng-Ann Heng. Tripletmix: Triplet data augmentation for 3d understanding. *arXiv preprint arXiv:2405.18523*, 2024.

[36] Yi Wang, Jiaze Wang, Jinpeng Li, Zixu Zhao, Guangyong Chen, Anfeng Liu, and Pheng Ann Heng. Pointpatchmix: Point cloud mixing with patch scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5686–5694, 2024.

[37] Fangyun Wei and Yutong Chen. Improving continuous sign language recognition with cross-lingual signs. *arXiv preprint arXiv:2308.10809*, 2023.

[38] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[39] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022.

[40] Zixu Zhao, Jiaze Wang, Max Horn, Yizhuo Ding, Tong He, Zechen Bai, Dominik Zietlow, Carl-Johann Simon-Gabriel, Bing Shuai, Zhuowen Tu, et al. Object-centric multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16601–16611, 2023.

[41] Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z Li. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23141–23150, 2023.

[42] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. *arXiv preprint arXiv:2307.14768*, 2023.

[43] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325, 2021.

[44] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779, 2021.

[45] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5131–5140, 2022.