A Hong Kong Sign Language Corpus Collected from Sign-interpreted TV News

Zhe Niu[1,*] **Ronglai Zuo**[1,*] **Brian Mak**[1] **Fangyun Wei**[2]

[1]The Hong Kong University of Science and Technology [2]Microsoft Research Asia
{zniu,rzuo,mak}@cse.ust.hk fawe@microsoft.com

Abstract

This paper introduces TVB-HKSL-News, a new Hong Kong sign language (HKSL) dataset collected from a TV news program over a period of 7 months. The dataset is collected to enrich resources for HKSL and support research in large-vocabulary continuous sign language recognition (SLR) and translation (SLT). It consists of 16.07 hours of sign videos of two signers with a vocabulary of 6,515 glosses (for SLR) and 2,850 Chinese characters or 18K Chinese words (for SLT). One signer has 11.66 hours of sign videos and the other has 4.41 hours. One objective in building the dataset is to support the investigation of how well large-vocabulary continuous sign language recognition/translation can be done for a single signer given a (relatively) large amount of his/her training data, which could potentially lead to the development of new modeling methods. Besides, most parts of the data collection pipeline are automated with little human intervention; we believe that our collection method can be scaled up to collect more sign language data easily for SLT in the future for any sign languages if such sign-interpreted videos are available. We also run a SOTA SLR/SLT model on the dataset and get a baseline SLR word error rate of 34.08% and a baseline SLT BLEU-4 score of 23.58 for benchmarking future research on the dataset.

Keywords: Sign Language Dataset, Sign Language Recognition (SLR), Sign Language Translation (SLT).

1. Introduction

Sign language (SL) is the primary mode of communication for many individuals who are deaf or hard of hearing. Advances in large deep networks demand more and larger datasets to support the development and research of sign language recognition (SLR) and sign language translation (SLT). Consequently, several large-scale datasets (Koller et al., 2015; Camgöz et al., 2018; Huang et al., 2018; Zhou et al., 2021a; Albanie et al., 2021) have been created for many sign languages such as ASL, BSL, CSL, DGS, etc. Given that there are currently few publicly available resources for Hong Kong Sign Language (HKSL) research, we introduce, in this paper, a new large-vocabulary HKSL dataset, called TVB-HKSL-News, for the development of SLR and SLT for HKSL.

The TVB-HKSL-News dataset contains HKSL videos collected from the News Report with Sign Language program (as depicted in Figure 1) from the Television Broadcasts Limited (TVB) over a period of seven months. The dataset is intended to support research in large-vocabulary, signer-dependent continuous SLR and SLT. It includes 16.07 hours of sign videos from two signers, with a vocabulary of 6,515 glosses (for SLR) and 2,850/18K Chinese characters/words (for SLT). One primary objective of building the dataset is to investigate new modeling methods for how well large-vocabulary SLR/SLT can be done for a single signer, given a relatively large amount of his/her training data. This is lacking in current HKSL resources and is not common even in other sign languages (with the exception of BOBSL (Albanie et al., 2021)). With a larger amount of data available for a signer, researchers can explore more sophisticated modeling techniques and evaluate the impact of amount of data on SLR and SLT.

Figure 1: In the "TVB News Report with Sign Language" program, a news anchor speaks on the left side of the screen while an HKSL interpreter signs on the right. Chinese subtitles are displayed at the bottom of the screen.

Besides the TVB-HKSL-News dataset, we also introduce a data collection pipeline for SL data from TV programs, which was utilized during the collection of our dataset. This pipeline includes an automatic collection process for Sign Language Translation (SLT) data and a computer-assisted

* Equal contribution.

LREC-COLING 2024, pages 636–646
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Table 1: Comparison of some existing datasets for continuous SLR and SLT with TVB-HKSL-News. The superscript † represents character count rather than word count, which is more suitable for gauging translation performance in Chinese (Li et al., 2019). The approximate word counts for CSL-Daily and TVB-HKSL-News are 8K and 18K, respectively. Our word counts are estimated using Jieba (Sun, 2012).

Dataset/Work	Lang.	SLR Vocab.	SLT Vocab.	Duration (h)	Resolution	FPS	#Videos	#Signers	Source
GSL (Adaloglou et al., 2022)	GSL	310	-	9.59	848×480	30	10K	7	Lab
PHOENIX-14 (Koller et al., 2015)	DGS	1,081	-	12.54	210×260	25	7K	9	TV
PHOENIX-14T (Camgöz et al., 2018)	DGS	1,066	2,887	10.53	210×260	25	8K	9	TV
CSL100 (Huang et al., 2018)	CSL	178	-	100	1280×720	30	25K	50	Lab
CSL-Daily (Zhou et al., 2021a)	CSL	2,000	2,370†	23.27	1920×1080	30	21K	10	Lab
BOBSL (Albanie et al., 2021)	BSL	2,281	78K	1,467	444×444	25	1.2M	39	TV
SignBERT (Zhou et al., 2021b)	HKSL	55	-	-	1920×1080	30	2K	6	Lab
TVB-HKSL-News (ours)	HKSL	6,515	2,850†	16.07	248×360	25	7K	2	TV

annotation process for Sign Language Recognition (SLR) data. In the initial phase, we collect SLT data from the TVB program, and process them to obtain sign video segments and their corresponding text transcriptions. This stage leverages various computer vision techniques to perform detection of sign activities, extraction of subtitles, and alignment between sign and subtitle segments. We then engage professional sign language annotators and design an annotation software for them to label the sign glosses in each collected sign video segment for SLR research. The software presents sign video segments alongside the text transcriptions from the previous stage, providing a convenient tool for efficient sign language glossing.

Following the compilation of our dataset, we conduct experiments with various state-of-the-art models (Xie et al., 2018; Chen et al., 2022b; Zuo and Mak, 2022a) for both SLR and SLT. We establish a strong baseline WER of 34.08% and a baseline BLEU-4 score of 23.58, respectively. These results serve as benchmarks for future research on the TVB-HKSL-News dataset. Furthermore, we also investigate signer-dependent tasks, and explore the impact of training dataset size under this setting. We demonstrate how the amount of training data affects performance and our findings indicate that additional data from a different signer can also lead to small improvements. In summary, we present the following contributions:

- We publish a new HKSL dataset called TVB-HKSL-news to support large-vocabulary continuous SLR/SLT for HKSL.
- We present an automated pipeline for collecting substantial SL data from subtitled, sign-interpreted videos for SLT research, along with an annotation software to facilitate sign glossing for SLR studies.
- We establish an SLR/SLT benchmark on the new dataset using a state-of-the-art SLR/SLT model.

Instructions to access the dataset can be found on our dataset webpage[1].

2. Related Works

Numerous sign language (SL) datasets have been created for various tasks, including SL spotting (Viitaniemi et al., 2014), SLR (Adaloglou et al., 2022; Huang et al., 2018; Koller et al., 2015; Camgöz et al., 2018; Özdemir et al., 2020; Li et al., 2020; Albanie et al., 2021; Momeni et al., 2020), SLT (Adaloglou et al., 2022; Ko et al., 2018; Camgöz et al., 2018; Albanie et al., 2021), and SL production (Kapoor et al., 2021). These datasets cover a wide range of different sign languages, such as ASL (Athitsos et al., 2008; Joze and Koller, 2019; Li et al., 2020), BSL (Momeni et al., 2020; Albanie et al., 2021), DGS (Koller et al., 2015; Camgöz et al., 2018), CSL (Huang et al., 2018; Zhou et al., 2021a), GSL (Adaloglou et al., 2022), TSL (Özdemir et al., 2020; Sincan and Keles, 2020), and ISL (Kapoor et al., 2021). Although there are some HKSL resources for educational purposes, there are very few HKSL resources for the development of automatic SLR/SLT models, with the exception of the relatively small dataset mentioned in SignBERT (Zhou et al., 2021b) and HKU-Portable (Zhou et al., 2022). Both works collected HKSL datasets composed of 50 unique sentences. SignBERT collected data using vision-based technologies, while HKU-Portable employed the Inertial Measurement Unit (IMU) data from smart watches. These datasets, despite their important contributions, are limited in their small amount of unique sentences and restricted vocabulary. In contrast, our TVB-HKSL-News dataset was collected from TV programs that comprises more diverse sign sentences with a much larger vocabulary size for both SLR and SLT research.

Several datasets such as (Athitsos et al., 2008; Joze and Koller, 2019; Li et al., 2020; Albanie et al., 2020; Huang et al., 2019; Sincan and Keles, 2020) have been developed specifically for isolated SLR. While these datasets are beneficial

[1] Dataset webpage: `https://tvb-hksl-news.github.io/`

Table 2: Statistics of the TVB-HKSL-News train/development/test data split.

	Hours	# Samples	Glosses for SLR				Chinese Characters for SLT			
			Vocab	Running	# OOVs	# Singletons	Vocab	Running	# OOVs	# Singletons
Train	14.71	6,516	6,515	111,204	N/A	2,925	2,816	212,108	N/A	466
Dev	0.67	322	1,091	5,222	0	471	1,279	10,003	17	395
Test	0.69	322	1,130	5,391	0	518	1,276	10,199	19	399
Overall	16.07	7,160	6,515	121,817	N/A	2,820	2,850	232,310	N/A	462

for training models to recognize individual signs, they do not capture the complex co-articulations between signs that are inherent in continuous signing. Continuous sign language recognition and translation, which recognize the natural interconnection between signs at the sentence level, provide the opportunity to gain a more comprehensive understanding of sign language expression. There has been a recent focus on developing datasets for continuous SLR and SLT (Ko et al., 2018; Huang et al., 2018; Zhou et al., 2021a; Koller et al., 2015; Camgöz et al., 2018; Albanie et al., 2021) as shown in Table 1. Datasets, such as KETI (Ko et al., 2018), CSL100 (Huang et al., 2018), and CSL-Daily (Zhou et al., 2021a), were created by recording a fixed set of sign sentences in a laboratory setting. Given that the signers perform sentences verbatim in these datasets, the resulting sign language can appear less natural compared to everyday, free-form signing. The datasets BOBSL (Albanie et al., 2021), PHOENIX-14 (Koller et al., 2015), LSA-T (Bianco et al., 2022), and PHOENIX-14T (Camgöz et al., 2018), which were sourced from more spontaneous environments like TV news broadcasts, feature a more natural form of signing. However, this way of data collection requires a post-glossing step to obtain the sign labels for SLR, which can be time-consuming. Some previous works (Bianco et al., 2022) lack sign glosses entirely, while others (Albanie et al., 2021) use an automatic sign gloss spotting method that may introduce errors.
Our approach to data collection parallels that of BOBSL (Albanie et al., 2021), with both datasets being sourced from TV programs via automated or semi-automated pipelines. However, in contrast to BOBSL, our methodology does not leverage the audio track and instead primarily employs computer vision techniques throughout the pipeline. Although much smaller than BOBSL, our dataset has been carefully and professionally annotated at the gloss level, whereas the gloss annotations in BOBSL are generated through automatic spotting. Among the SLs listed in Table 1, our dataset has the largest gloss vocabulary, and is the second largest in the size of SLT vocabulary in terms of words.

3. TVB-HKSL-News Dataset

The proposed TVB-HKSL-News dataset was sourced from the *TVB News Report with Sign Language program*[2]. The program covers a wide range of topics, including but not limited to politics, economy, sports, and weather, with each episode running approximately 20 minutes. Broadcast at a

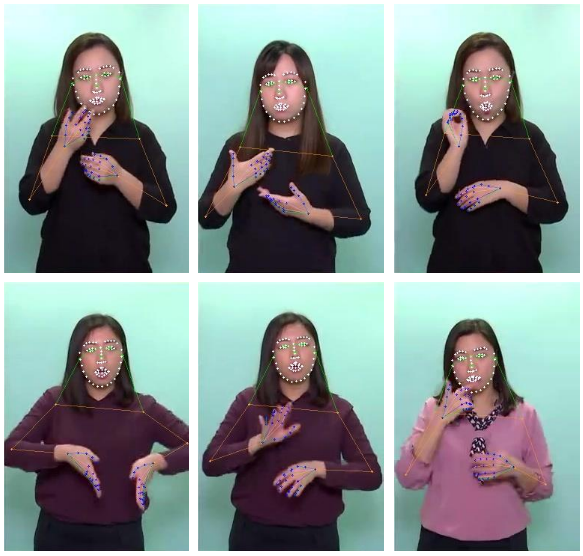

Figure 2: Keypoints extracted from Signer-1/Signer-2 in the first/second row, respectively.

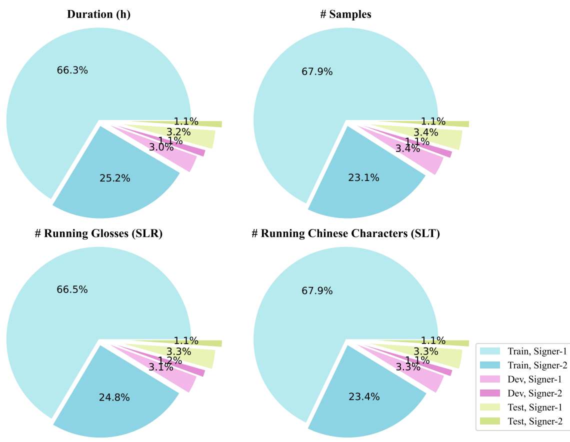

Figure 3: Pie charts showing statistics: number of samples, duration (hours), and running glosses/characters for SLR/SLT tasks across train/dev/test sets.

[2]The link to the TVB News Report with Sign Language: https://news.tvb.com/tc/programme/newsreportwithsignlanguage.

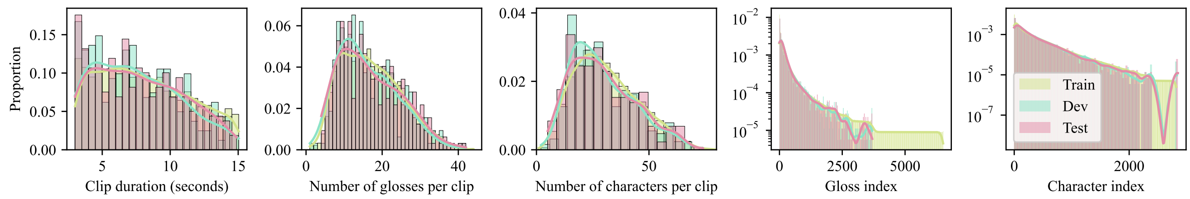

Figure 4: The first three figures show the distribution of clip lengths, number of glosses per clip for SLR, and characters per clip for SLT. The last two figures show the occurrences of particular sign glosses and characters among the samples, sorted by their overall occurrences. Different colors represent different dataset splits.

frame rate of 25 fps and a resolution of 1024×576, it features Chinese subtitles and an HKSL interpreter performing sign language interpretation in a fixed 248×360 area on the right side of the screen, alongside the news anchor (see Figure 1). To build our dataset, we acquired 206 episodes of the program, aired from January 16 to August 17, 2020, with the permission of Television Broadcasts Limited (TVB). These episodes featured simultaneous sign interpretation by two professional HKSL interpreters.

3.1. Dataset Specifications

Our dataset, compiled from the source data above, comprises the following four data types:

Sign Video Clips. The sign video clips are RGB video segments extracted from the designated region of the sign interpreter on the screen. It has a resolution of 248×360 and a frame rate of 25 fps, with the signer centered in the frame.

Sign Keypoints. In accordance with recent research trends that emphasize human body keypoints modeling (Zhou et al., 2020; Zuo and Mak, 2022a; Chen et al., 2022b; Jiang et al., 2021; Hu et al., 2021; Zuo et al., 2023), we also provide estimated keypoints for each video frame. These keypoints, comprising 68 for the face, 42 for the hands, and 11 for the upper body (as depicted in Figure 2), are designed to be robust to variations in signer appearance, hand positions, and shapes.

Chinese Subtitles for SLT. We extract textual subtitles from the sign clips using Optical Character Recognition (OCR). These subtitles transcribe the content spoken by the news anchors and are presented in the form of standard spoken Chinese sentences, serving as a resource for SLT research.

Chinese Glosses for SLR. For SLR studies, we provide Chinese glosses, obtained from professional sign annotators, for each of the sign clips. These glosses are words that describe the signs verbatim and are sequenced in the order of HKSL, making them suitable for SLR tasks.

3.2. Dataset Statistics

The overall statistics of our proposed dataset are presented in Table 2. The dataset consists of 6,515 glosses for SLR and 2,850 Chinese characters for SLT. We divided the dataset into training, development (dev), and test sets according to a rough ratio of 90:5:5. Out-of-vocabulary (OOV) glosses are intentionally avoided in the dev and test sets by rejection sampling. All data splits cover both interpreters, labeled as *Signer-1* and *Signer-2*. Figure 3 shows the detailed data distribution among the three dataset splits: Signer-1 contributes 11.66 hours of videos, accounting for about 73% of all sign videos, while Signer-2 contributes 4.41 hours of videos. The dataset exhibits a wide range of video durations, glosses per segment, and Chinese characters per segment. As shown in Figure 4, the duration of the video segments ranges from 3–15 seconds, with a mode of 7.88 seconds; the number of glosses per segment ranges from 1–44, with a mode of 16 glosses; the number of Chinese characters per segment ranges from 1–120, with a mode of 29 characters.

4. Dataset Collection Methodology

In this section, we outline our data collection methodology for both SLT and SLR. We begin by describing the SLT collection pipeline. Afterwards, we detail the glossing method we adopted to obtain annotations for the SLR task.

4.1. Data Collection for SLT

Obtaining Sign Video Clips. Since the sign language interpreter does not continuously sign throughout the TV news program, we need to perform signer activity detection. To achieve this, we train a video clip binary classifier for each sign language interpreter to detect his/her active and inactive periods. The classifier is built based on a ResNet18+TCN (He et al., 2016; Lea et al., 2016) model. It takes five video frames as input and produces the probability that the signer is actively signing in the center (i.e., the third) frame. To train the classifier, we use the data from 60 episodes with one minute per episode, which are manually annotated with activity labels (i.e., 0 and 1). We use the trained classifier to generate these labels for the remaining videos. The labels help us group consecutive active frames together to create active

Figure 5: Example of subtitle background removal. Lines 1 and 2 show an example of sample input and target image used for training the U-Net model. The trained model processes Line 3 (input) to produce Line 4 (output) with background removed.

sign clips. Finally, we filter and retain only clips with durations ranging from 3 to 15 seconds.

Obtaining Text from Subtitles. Text transcriptions for SLT are obtained from the subtitles that are embedded in a fixed 768×48 region at the bottom of the TV program screen. After cropping this fixed region from the video, we apply the following procedure to extract the text:

1. *Subtitle Background Removal.* We first train a background removal model based on the U-Net (Ronneberger et al., 2015). The model clears the background of the subtitle with black, as shown in Figure 5. To train this model, we first create a synthetic dataset consisting of pairs of source and target images. A source image comprises a non-textual TV program background overlaid with random text, augmented with white noise and Gaussian blur. A target image is a pure black background overlaid with the same text at the same position. The non-textual TV program backgrounds are obtained by randomly cropping a region above the subtitle on the screen, and random text is generated from the zhwiki corpus[3]. The well-trained model is applied to each frame of the episode specifically within the subtitle region. This process effectively removes the background, yielding clean subtitle videos.

2. *Subtitle Segmentation.* Subtitle video clips are extracted by identifying rapid transitions in cleaned subtitle videos. These rapid transitions include the appearance, disappearance, and switching of subtitles. To identify such rapid transitions, we apply a temporal Laplacian filter to the intensity values of each pixel across consecutive cleaned frames. We then take the average of all the filtered pixel values in each frame and use it as a measure of the subtitle transition. Peaks in this measure signify rapid transitions in subtitles. We select frames with large transitions via thresholding and use them to segment the video into shorter subtitle clips. Ideally, each of these clips should contain a unique subtitle (including the empty subtitle). Within each subtitle clip, we average all frames. Clips that yield a pure black image, indicating that the subtitle is absent, are subsequently removed.

3. *Optical character recognition (OCR).* After obtaining the averaged frame of each subtitle video clip, we use the Google OCR API[4] to recognize the text in the averaged frame.

4. *Text Re-grouping.* In practice, we find that the above simple thresholding method can result in oversegmentation, where the same subtitle appears in multiple consecutive video clips. To address this, we iteratively merge adjacent subtitles as long as their edit distance (Levenshtein et al., 1966) is below a threshold. This process can form groups consisting of multiple subtitles. Within each group, we choose the most representative subtitle by selecting the one that has the minimum average edit distance to all the other subtitles in the group. In the end, we obtain the subtitle text along with the precise starting and ending timestamps by considering the frame index of the first and last subtitle frames in the group, respectively.

Aligning Sign and Subtitle Clips. To align the sign clips and subtitle clips, we adopt a Dynamic Time Warping (DTW)-based algorithm. This algorithm treats all the sign clips and subtitle clips within one episode as two time series, and attempts to align them based on the distance between their temporal midpoints. Notably, in our algorithm, we allow multiple subtitle clips to align with a single sign clip, as subtitle clips are typically shorter than sign clips.

Keypoint Extraction from Sign Videos. To extract the facial, hand and body keypoints from sign videos, we use HRNet (Sun et al., 2019), which is pre-trained on COCO-WholeBody (Jin et al., 2020), following the approach in (Chen et al., 2022b; Hu et al., 2021; Zuo et al., 2023, 2024). The model predicts 133 keypoints, from which we discard 6 lower body keypoints and 6 foot keypoints, as they are irrelevant to sign language tasks and our sign frames only show the upper body of the signers. This process leads to a final count of 121 keypoints per frame, including those for the upper body, hands, and face.

4.2. Glossing for SLR

Sign languages have their own rules for word order, sentence structure, and grammar that are different

[3]The zhwiki corpus: `https://dumps.wikimedia.org/zhwiki/`

[4]Google OCR API: `https://cloud.google.com/vision/docs/ocr`.

from those of spoken languages. Therefore, subtitles for spoken languages cannot be directly used as glosses for sign language recognition, where monotonic gloss sequences are required as training targets. To obtain accurate gloss annotations of the signed clips collected, we hired HKSL signers from SLCO Community Resources[5] to annotate them. The annotators are fluent in both HKSL and Cantonese (Chinese). To facilitate the annotation process, we have developed a software tool for the annotators. The software presents sign language video clips side by side with the corresponding subtitle text, which is derived from the previous pipeline. Annotators can efficiently gloss the sign language videos using the tool with the subtitle text as a reference. During the annotation process, the annotator glosses the signs in the video verbatim, from left to right. In addition to the common signs that can be described using single Chinese words, there are signs that require special attention. The glossing of these special cases is dealt with using the following rules:

1. *Compound signs* are annotated in the form of *X+Y*. They are formed by combining two or more signs to create a phrase of distinct meaning. During post-processing, we split these signs into *X Y*.

2. *Ill-performed signs* are signs that are not signed correctly but can still be recognized by the annotators. They are annotated as *X(?)*. During post-processing, we assume they are the truly *X*.

3. *Many-to-one glossing* occurs when multiple signs are glossed with the same term, represented as *X(1)*, *X(2)*, and so on. We choose the common glossing, *X*, for all these variations.

4. *One-to-many glossing* arises when a single sign can be transcribed into multiple glosses due to the availability of various suitable gloss words. This is annotated as a homosign group: *X(=Y=Z=...)*. During training, we select the gloss with the highest compound count for each group (e.g., A+B+C has the compound count of three), or the lowest lexicographical order if there is a tie. In the testing phase, we proceed by globally merging all homosign groups that share at least one common element. The selection of the representative gloss for the newly merged group follows the same strategy used during training. Before computing the score, we replace glosses appearing in any homosign group with its representative gloss in both the hypothesis and reference sentences for each sample.

These post-processing rules simplify raw annotations and result in cleaner and more consistent glossing of signs, while preserving the necessary information for the SLR task[6].

5. Experiments

Here, we present our experiments on SLR and SLT. We first introduce the baseline models and the evaluation metrics that are used to assess their performance. We then report our experimental results on the full dataset and provide a qualitative analysis of the outcomes. Finally, we evaluate the impact of the amount of training data under a signer-dependent setting for a single signer. We hope that the experimental results will serve as benchmarks for future research on this dataset.

5.1. Baselines and Evaluation Metrics

Video-based Baselines. Most existing deep-learning-based SLR models adopt 2D- or 3D-CNNs (Simonyan and Zisserman, 2015; He et al., 2016; Carreira and Zisserman, 2017; Xie et al., 2018) for visual feature extraction. A sequential module, which is composed of, for example, temporal CNNs (Lea et al., 2016), Transformer (Vaswani et al., 2017), or a mixture of them (Yu et al., 2018), is optionally appended to the CNNs for temporal modeling. In this paper, we choose S3D (Xie et al., 2018) for the implementation of 3D-CNNs, which is adopted as the backbone network of the video/keypoint encoder in TwoStream-SLR (Chen et al., 2022b) (the current best model on several widely adopted SLR benchmarks (Koller et al., 2015; Camgöz et al., 2018; Zhou et al., 2021a)) due to its strong spatial-temporal modeling capability and high efficiency. Besides, we also implement VLT (Zuo and Mak, 2022a) as our sequential module with a mixed architecture, which is the backbone network of C^{2}SLR (Zuo and Mak, 2022a); it uses VGG11 (Simonyan and Zisserman, 2015) to extract frame-wise visual features and a local transformer (LT) (Zuo and Mak, 2022b; Yu et al., 2018) to model temporal dependencies.

Baselines that Utilize Keypoints. Substantial visual redundancy in RGB videos may lead video-based SLR models to overlook the key information for SL understanding (Chen et al., 2022b). Besides, the performance of video-based models may also suffer from large variations in video backgrounds and signer appearances. To obtain better representations, some SLR works (Chen et al., 2022b; Zuo and Mak, 2022a; Zhou et al., 2020; Hu et al.,

[5]SLCO Community Resources: https://www.slco.org.hk.

[6]The original raw annotations are available so that users may investigate their own post-processing rules.

Table 3: Baseline results for SLR.

Method	Modality Video	Modality Keypoints	WER (%) Dev	WER (%) Test
S3D (Xie et al., 2018; Chen et al., 2022b)		✓	45.73	44.56
S3D (Xie et al., 2018; Chen et al., 2022b)	✓		39.59	38.63
VLT (Zuo and Mak, 2022a)	✓		35.89	36.18
C^2SLR (Zuo and Mak, 2022a)	✓	✓	35.43	35.78
TwoStream-SLR (Chen et al., 2022b)	✓	✓	**34.52**	**34.08**

Table 4: Baseline results for SLT. (R: ROUGE, B: BLEU.)

Method	Dev R	Dev B-1	Dev B-2	Dev B-3	Dev B-4	Test R	Test B-1	Test B-2	Test B-3	Test B-4
S3D (video) (Xie et al., 2018; Chen et al., 2022b)	18.64	21.98	15.17	11.18	8.79	21.61	25.39	18.59	14.57	12.10
S3D (keypoints) (Xie et al., 2018; Chen et al., 2022b)	15.65	18.18	11.58	8.09	6.22	16.42	19.93	13.72	10.41	8.48
TwoStream-SLT (Chen et al., 2022b)	**38.12**	**43.22**	**33.44**	**26.04**	**21.00**	**39.80**	**44.68**	**35.27**	**28.29**	**23.58**

2021; Jiang et al., 2021) also make use of keypoints as complementary information in their models. In this paper, we first implement C^2SLR (Zuo and Mak, 2022a) which utilizes keypoint heatmaps of signers' faces and hands as a guidance for a spatial attention module to enforce the visual module to focus on informative regions of a signing frame. Second, we implement TwoStream-SLR (Chen et al., 2022b), which also represents keypoints as a sequence of heatmaps and shares an identical architecture (S3D) between the video and keypoint streams. Finally, TwoStream-SLT (Chen et al., 2022b), which simply appends an MLP and a translation network to each head of TwoStream-SLR is used for SLT.[7] We implement all baselines by following their original works. All models are trained on 8 NVIDIA Tesla V100 GPUs with a batch size of 8.

Evaluation Metrics. Following the usual practice (Chen et al., 2022b; Zuo and Mak, 2022a; Zhou et al., 2020; Camgöz et al., 2018; Koller et al., 2015), we adopt word (gloss) error rate (WER) for SLR, and BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) for SLT. Lower WER means better SLR performance while higher BLEU and ROUGE-L indicate better SLT performance.

5.2. SLR and SLT Baseline Results

SLR. The baseline results for SLR are shown in Table 3. Using S3D with keypoints as the baseline, video inputs can lead to much better performance than keypoint heatmap inputs (44.56% $\rightarrow$ 38.63% on the test set). The results are reasonable since RGB videos contain much more information than sparse keypoints. For video-based baselines, VLT outperforms S3D by 2.45% on the test set. The result suggests that a sequential module, e.g., a local transformer, with both local and global context modeling capability is preferred. For baselines that utilize keypoints, C^2SLR uses pre-extracted keypoint heatmaps to enforce the spatial attention module to focus on informative sign regions. The mechanism is also effective on the proposed HKSL dataset with an improvement of 0.46% on the dev set over the best video-based baseline. The lowest WERs (34.52%/34.08% on the dev/test set) come from the TwoStream-SLR model, which better exploits the benefits of keypoints by developing a dual-S3D architecture that takes both videos and keypoint heatmaps as inputs. It significantly improves the performance of the S3D baseline by 5.07%/4.55% on the dev/test set, respectively.

SLT. Table 4 provides several SLT baseline results. The TwoStream-SLT model achieves a BLEU-4 score of 23.58 on the test set. It outperforms the single-stream models by a large margin, and thus validates the effectiveness of the joint modeling of both videos and keypoints. The SLT performance on our new dataset is comparable to that on CSL-Daily (Chen et al., 2022a) (25.79 on the test set). The written texts in both datasets are Chinese, but the character vocabulary size of our dataset is about 20% larger, and our word vocabulary size is more than double of CSL-Daily's, and yet there are only 2 signers in our training set whereas there are 10 signers in CSL-Daily.

Qualitative Results. SLR and SLT results on some test data using the strongest baseline, TwoStream-SLR/SLT, are shown in Table 5. It is clear that the two-stream model can yield more accurate gloss/text predictions than both the single-stream models, which only take RGB videos or keypoint heatmaps as inputs. The results also suggest that the two modalities can complement each other.

5.3. Evaluation on Single-signer SLR/SLT

One key characteristic of our HKSL dataset is that it has a large amount of data for a single-

[7] We only focus on the Sign2Text setting, which directly translates sign videos into spoken languages. It has also been proven that it outperforms the two-stage Sign2Gloss2Text setting (Chen et al., 2022a,b).

Table 5: Qualitative results on TVB-HKSL-News. For SLR, we use different colors to represent substitutions, deletions, and insertions, respectively.

SLR		WER (%)
Ground Truth	昨天 温度 二 十 有 濕 百分比 七 六 (Yesterday Temperature Two Ten Has Humidity Percentage Seven Six)	-
Pred. (S3D Video)	以前 温度 小 * 有 濕 百分比 七 六 (Previously Temperature Low * Humidity Percentage Seven Six)	33.33
Pred. (S3D Keypoint)	** 温度 * 十 * 濕 百分比 七 九 六 (** Temperature * Ten * Humidity Percentage Seven Nine Six)	44.44
Pred. (TwoStream-SLR)	** 温度 二 十 * 濕 百分比 七 六 (** Temperature Two Ten * Humidity Percentage Seven Six)	22.22
SLT		BLEU-4
Ground Truth	而輸入個案有一宗是一名印度海員 (One of the imported cases was an Indian seafarer.)	-
Pred. (S3D Video)	至於輸入個案有11宗包括印度海員 (Regarding the imported cases, there are eleven cases including an Indian seafarer.)	39.38
Pred. (S3D Keypoint)	有~~輸入個案~~滯留印度的港人~~海員~~知道賭場重開 (Some ~~imported cases~~ Hong Kong people ~~seafarer~~ stranded in India know that casinos are reopening.)	0.00
Pred. (TwoStream-SLT)	至於輸入個案有一名印度海員 (Regarding the imported cases, there is one Indian seafarer.)	63.86

signer, Signer-1, the major signer in the dataset with 11.66 hours of sign videos that consists of 5,350 samples. This enables studies on how well SLR/SLT performance can be achieved under the signer-dependent setting. We study the effect of the amount of training data on SLR/SLT performance using the TwoStream-SLR/SLT model and only Signer-1's training data (with the exception of the "133% experiment"). Performance is evaluated only on Signer-1's test data. The results are shown in Figure 6 together with the corresponding OOV ratios (grey lines). Notice that since this evaluation uses Signer-1's data only, the results shown in Figure 6 are different from results shown in Table 3 and 4 (as the latter use data from all signers).

First, we notice that the maximum OOV ratio is just 4.27% and 1.35% for SLR and SLT, respectively, when using only 25% training data. Thus, the effect of OOV is negligible on the SLR/SLT performance. Second, we find that, as expected, the model performance on both SLR and SLT improves when more training data are available. Third, the addition of more training data from another signer (shown as 133% training data in Figure 6) can still improve the performance. We believe the use of keypoint information in the TwoStream-SLR/SLT model makes the SLR/SLT performance more robust to signer appearances. Finally, SLR performance converges with the use of 75% of the training data (which is around 8 hours of sign videos) but SLT performance improves linearly with the amount of training data. For instance, increasing the amount of Signer-1's data from 75% to 100% can only lead to a WER reduction of 1.22% (34.30% $\rightarrow$ 33.08%, or 3.56% relative) on the test set. On the other hand, the corresponding BLEU-4 improvement is 2.50 (20.29 $\rightarrow$ 22.79, or 12.32% relative). The results suggest that around 8 hours of training data may

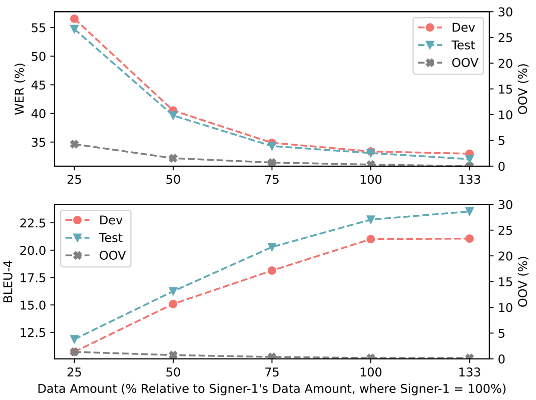

Figure 6: SLR and SLT performance of TwoStream-SLR/SLT with different amount of Signer-1's training data. 133% means that we use all the training data of both Signer-1 and Signer-2 (whose amount of data is about 33% of Signer-1's) but evaluate the trained model only on Signer-1's dev/test data.

be sufficient for signer-dependent SLR modeling in our setting, but SLT still benefits from much more data than SLR. This can serve as a guidance for future research on signer-dependent SLR/SLT.

6. Conclusion

This work introduces the TVB-HKSL-News dataset, which is a valuable resource for advancing research in continuous SLR and SLT. The dataset also provides a large amount of data for one of the signers with manual gloss annotations, enabling researchers to investigate new modeling methods and evaluate the impact of the amount of training data needed for the SLR and SLT of a single signer. Furthermore, the automated data collection pipeline introduced in the paper makes it easier to scale up the collection of a large amount of data for SLT in the future if sign-interpreted videos with subtitles are available. Moreover, the presented results of some state-of-the-art SLR/SLT models on the dataset serve as the benchmarks for future research on the dataset.

7. Ethical Consideration

During the creation and the use of the TVB-HKSL-News sign language dataset in our project, we have obtained informed consent from the sign language interpreters involved in the dataset. They are aware of their data being used for academic research purposes. Additionally, we have implemented a procedure to make the dataset accessible for academic research. The interested parties are required to sign an agreement with Television Broadcasts Limited (TVB), ensuring that the dataset will be used only for academic research.

8. Dataset Availability Statement

The complete dataset will be hosted on a dedicated server. Authorized users can download it using a specific URL and password, which can be requested at `https://tvb-hksl-news.github.io/`. To gain authorization, users must first sign an agreement that will be available on our dataset website, ensuring the dataset's appropriate use. The copyrights for all materials in the TVB-HKSL-News dataset belongs to Television Broadcasts Limited (TVB). We have obtained permission from TVB for its distribution strictly for **academic research** and the data **should not be used for any other commercial or non-commercial purposes**.

9. Bibliographical References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th. Papadopoulos, Vassia Zacharopoulou, George J. Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2022. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE TMM*, 24:1750–1762.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, volume 12356, pages 35–53. Springer.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BBC-Oxford British sign language dataset. *CoRR*, abs/2111.03635.

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan P. Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The American sign language lexicon video dataset. In *CVPR*, pages 1–8.

Pedro Dal Bianco, Gastón Ríos, Franco Ronchetti, Facundo Manuel Quiroga, Oscar Stanchi, Waldo Hasperué, and Alejandro Rosete. 2022. LSA-T: the first continuous argentinian sign language dataset for sign language translation. In *Advances in Artificial Intelligence - IBERAMIA 2022 - 17th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 23-25, 2022, Proceedings*, volume 13788 of *Lecture Notes in Computer Science*, pages 293–304. Springer.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *CVPR*, pages 7784–7793.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Xiujuan Chai, Hanjie Wang, and Xilin Chen. 2014. The devisign large vocabulary of Chinese sign language database and baseline evaluations. In *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)*. Institute of Computing Technology.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multimodality transfer learning baseline for sign language translation. In *CVPR*, pages 5120–5130.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022b. Two-stream network for sign language recognition and translation. In *NeurIPS*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. SignBERT: Pre-training of hand-model-aware representation for sign language recognition. In *ICCV*, pages 11087–11096.

Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2019. Attention-based 3D-CNNs for large-vocabulary sign language recognition. *IEEE TCSVT.*, 29(9):2822–2832.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *AAAI*, pages 2257–2264.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *CVPRW*, pages 3413–3423.

Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese lexical analysis with deep bi-gru-crf network. *arXiv preprint arXiv:1807.01882*.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Whole-body human pose estimation in the wild. In *ECCV*, pages 196–214.

Hamid Reza Vaezi Joze and Oscar Koller. 2019. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *BMVC*, page 100.

Parul Kapoor, Rudrabha Mukhopadhyay, Sindhu B. Hegde, Vinay P. Namboodiri, and C. V. Jawahar. 2021. Towards automatic speech to sign language generation. In *Interspeech*, pages 3700–3704.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choong Sang Cho. 2018. Neural sign language translation based on human keypoint estimation. *CoRR*, abs/1811.11436.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *CVIU*, 141:108–125.

Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *ECCVW*, Lecture Notes in Computer Science, pages 47–54.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, pages 1448–1458.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of Chinese representations? In *ACL*, pages 3242–3252.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2020. Watch, read and lookup: Learning to spot signs from multiple supervisors. In *ACCV*, volume 12627, pages 291–308.

Ogulcan Özdemir, Ahmet Alp Kindiroglu, Necati Cihan Camgöz, and Lale Akarun. 2020. BosphorusSign22k sign language recognition dataset. *CoRR*, abs/2004.01283.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351, pages 234–241.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. AUTSL: A large scale multi-modal Turkish Sign Language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Junyi Sun. 2012. Jieba chinese word segmentation tool.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703.

TVB News. 2023. News report with sign language. `https://news.tvb.com/tc/programme/newsreportwithsignlanguage`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. 2014. S-pot - a benchmark in spotting signs within continuous signing. In *LREC*, pages 1892–1897.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021a. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, pages 1316–1325.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI*, pages 13009–13016.

Zhenxing Zhou, Vincent WL Tam, and Edmund Y Lam. 2021b. SignBERT: a BERT-based deep learning framework for continuous sign language recognition. *IEEE Access*, 9:161669–161682.

Zhenxing Zhou, Vincent WL Tam, and Edmund Y Lam. 2022. A portable sign language collection and translation platform with smart watches using a BLSTM-based multi-feature framework. *Micromachines*, 13(2):333.

Ronglai Zuo and Brian Mak. 2022a. C2SLR: Consistency-enhanced continuous sign language recognition. In *CVPR*, pages 5131–5140.

Ronglai Zuo and Brian Mak. 2022b. Local context-aware self-attention for continuous sign language recognition. In *Interspeech*, pages 4810–4814.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *CVPR*.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2024. Towards online sign language recognition and translation. *arXiv preprint arXiv:2401.05336*.