2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

BTVSL: A Novel Sentence-Level Annotated Dataset for Bangla Sign Language Translation

Iftekhar E Mahbub Zeeon[1], Mir Mahathir Mohammad[2] and Muhammad Abdullah Adnan[3]
[1,3] Department of Computer Science, Bangladesh University of Engineering and Technology
[2] Kahlert School of Computing, University of Utah

Abstract— **Sign language, a vital communication tool for individuals with hearing impairments worldwide, is prominently utilized within the Bangla-speaking community. Bangla is recognized as the seventh most spoken language globally. However, research in Bangla Sign Language Recognition (SLR) - the process of translating symbols or words from images and videos - has been predominantly confined to controlled environments with limited samples and rudimentary symbol annotations, impeding its application in real-world scenarios. In contrast to previous studies, our research concentrates on SLR. It delves into the relatively unexplored territories of Bangla Sign Language Translation (SLT) and Sign Language Production (SLP), areas that have been largely overlooked due to dataset constraints. We introduce BTVSL, a comprehensive Bangla Sign Language dataset derived from the YouTube series 'BTV Desh o Jonopoder Khobor'. This dataset, featuring 60 hours of news content in sign language delivered by professionals, represents the largest sentence-level dataset available for Bangla SLT, encompassing a broad spectrum of expressions. Leveraging BTVSL, we evaluated four distinct SLT models, achieving an average BLEU score of 20.42. This result underscores the potential of BTVSL in enhancing the accuracy of sign language translation.**

I. INTRODUCTION

Sign Language (SL) is a comprehensive communication system that employs physical gestures, encompassing facial expressions, hand movements, and arm postures, to articulate thoughts and meanings. It is the primary communication medium for the deaf and mute community globally. As per the World Federation of the Deaf (WFD), approximately 70 million individuals rely predominantly on sign language for their daily communication. The structure and syntax of sign language exhibit significant variation across different regions, with multiple dialects often co-existing within a single country's borders. Analogous to spoken languages, sign languages possess a unique grammar native to the source language. A combination of facial expressions, hand movements, and full-body gestures shapes the semantics of signed expressions. For instance, the American Sign Language (ASL) allows the expression of the alphabet using a single hand, whereas both hands are utilized in German and British Sign Language.

Bangla, a language spoken by 337 million individuals, also has a substantial deaf community that employs the Bangla Sign Language daily. According to statistics, 34.6% of the population (49.2 million) experiences some degree of hearing loss, and profound hearing loss (a loss of 90 dB or more) is present in 1.2% of the population (1.7 million). The development of an automated translator, specifically designed to comprehend the language of people who are deaf or hard of hearing, could significantly facilitate their communication with the broader population of Bangla users.

In the domain of sign language research, three primary challenges have been identified: Sign Language Recognition (SLR), Sign Language Translation (SLT), and Sign Language Production (SLP). SLR is concerned with converting sign language into a symbol-level representation of spoken language. In contrast, SLT focuses on translating sign language into a sentence-level representation of spoken language. Conversely, SLP is essentially the inverse of SLT, translating written or spoken language into sign language. Various datasets have been established in the literature for sign language translation across different languages. The benchmark dataset, RWTH-PHOENIX-Weather 2014 [3], is predominantly used for all three problems. Additionally, datasets for Chinese Sign Language (CSL) are also frequently employed [19], [53]. However, it is noteworthy that existing datasets for Bangla sign language, such as the one mentioned in [22], primarily include symbol-level annotations for SLR. This highlights the need for more comprehensive datasets to facilitate research in SLT and SLP for Bangla sign language.

In this paper, we introduce **BTVSL**, the first comprehensive dataset for sentence-level Bangla Sign Language Translation (SLT). The dataset is compiled from a publicly accessible YouTube channel with text data. Subtitles are obtained via a third-party speech-to-text API. We have devised post-processing algorithms to enhance the organization of the dataset, thereby facilitating further training. To evaluate the performance of our dataset, we employ the S2T Stochastic Transformer model [49]. Standard Natural Language Processing (NLP) metrics, such as BLEU and ROUGE scores, are used, with a varying Alpha and a fixed beam size. After the initial evaluation, we extend our experimentation to four distinct models: TwoStream-SLT [8], CiCo-Sign Language Retrieval [10], Gloss Attention for Gloss-free Sign Language Translation [55] and Gloss-free Sign Language Translation [57]. This further exploration allows us to assess the versatility and applicability of our dataset across various models.

The key contributions of this paper are encapsulated as follows:

- We introduce **BTVSL**, a pioneering dataset designed specifically for the Bangla sign language translation task.
- We comprehensively describe the methodology employed in creating our sign language translation (SLT)

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

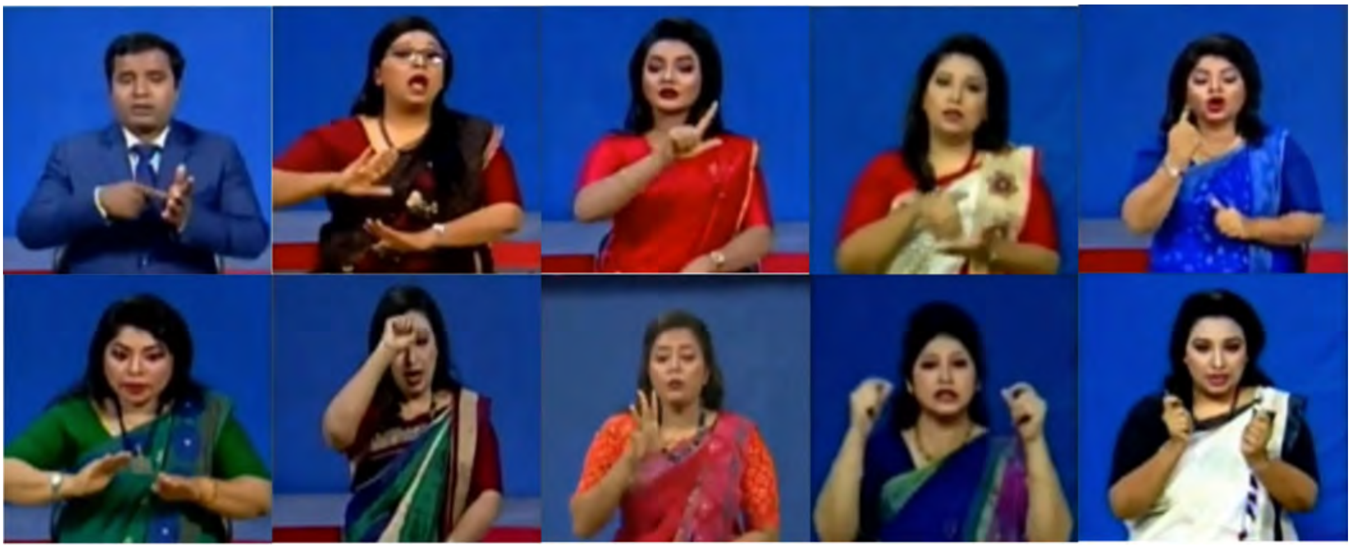


Fig. 1. Signers of BTVSL

dataset, including utilizing third-party models for data extraction.

- We evaluate our dataset's performance for SLT tasks, employing four well-established deep-learning models.
- We explore various parameters and preprocessing techniques to enhance the performance of our dataset.

The structure of this paper is as follows: In Section II, we provide a comprehensive review of the existing literature, focusing on the datasets and methodologies employed in Sign Language Translation (SLT) and Sign Language Processing (SLP). We also delineate our unique contributions in the context of these works. Section III is dedicated to formulating the SLT problem and describing our novel approach, which results in the production of **BTVSL**. In Section IV, we explain our experimental setup in-depth and discuss the evaluation metrics used, including ROUGE and BLEU scores. Finally, we conclude the paper in Section V, outlining our plans to expand and enhance our dataset.

II. RELATED WORKS

A. Datasets for SLT

Many sign language datasets, encompassing a diverse range of languages, have been documented in the sign language literature. These corpora typically comprise sentences, corresponding videos, and gloss information. Sign language experts generate most of these datasets in controlled, studio-like environments. For instance, datasets were constructed by Wilbur et al. [52] and Dreuw et al. [14] in such settings, with sign language professionals interpreting scripts. However, these works scarcely reflect real-world scenarios, as investigated by Yin et al. [56]. Furthermore, the expansion of these datasets is challenging due to substantial costs. In contrast, Camgoz et al. [3], [7], and Albanie et al. [1] have created datasets derived from television programs.

The benchmark datasets for Sign Language Recognition (SLR) tasks have been the German datasets: RWTH-PHOENIX-Weather 2014: Continuous Sign Language Recognition and Parallel Corpus of Sign Language Video, Gloss and Translation [3], [25]. Despite this, a multitude of models have been constructed on a variety of other sign languages. These include Indian Sign Language (ISL) [39], [33], [50], Chinese Sign Language (CSL) [19], [53], Korean Sign Language [24], Turkish Sign Language [16], [23], Persian Sign Language [40], and American Sign Language (ASL) [9], [44]. Furthermore, datasets comprising Bangla sign language data with symbol-level annotations have been created by Rafi et al. [38], Islam et al. [22], and Hasib et al.[17].

We introduce the Bangla Text to Video Sign Language (BTVSL) dataset, which provides sentence-level annotations for Bangla Sign Language. This dataset has been compiled from a YouTube playlist. Furthermore, to promote open and collaborative research, we have made all the annotations publicly accessible, which can be found here.

In sign language translation and recognition, many non-deep learning algorithms have been employed, including but not limited to Support Vector Machines (SVM), Hidden Markov Models (HMM), and various statistical analysis techniques. Deep learning models have become increasingly prevalent for these tasks as the field has evolved. These encompass Convolutional Neural Networks (CNN), sequence models, and Generative Adversarial Networks (GAN). Historically, sequence models such as LSTM [47], Bi-LSTM [21], and GRU [11] networks have been leveraged for Sign Language Recognition (SLR) and Sign Language Production (SLP). However, in recent times, transformer models have emerged as a popular choice, attributed to their enhanced performance in these tasks.

Sign language recognition is primarily concerned with accurately identifying the semantic content conveyed by a sign language interpreter. A multitude of models, including those proposed by [51], [12], [40], [39], [9], [19], [36], [44],

[50], [23], have been used to convert sequences of images or videos depicting sign language into translations at the word level. Similarly, the works of [37], [2], [20], [26], [3] have strived to generate translations at the sentence level from sequences of images. A significant number of studies, such as those by [4], [33], [16], [53], [24], [23], have adopted an intermediary approach. They have extracted key points from the body of sign language interpreters from sequences of images, and these extracted representations have been translated into translations at the word or sentence level. This approach has shown promising results in the ongoing research in the field of sign language recognition.

Convolutional Neural Networks (CNN) have emerged as the preferred choice in image sequence processing. A multitude of studies [51], [39], [36], [44], [50], [23] have employed CNN models for Sign Language Recognition (SLR) tasks. Despite the effectiveness of CNNs, there has been a surge in developing more potent CNN models. This has been achieved by amalgamating CNN models with traditional sequential models. Examples of such models include Long Short-Term Memory (LSTM) [2], [20], [40], [9], [33], Bidirectional LSTM (Bi-LSTM) [37], [12], [19], Gated Recurrent Units (GRU) [24], Recurrent Neural Networks (RNN) [3], and Hidden Markov Models [26]. The advent of transformer models in Natural Language Processing (NLP) inspired researchers to apply these models to Sign Language Recognition (SLR) tasks. Studies by [4], [16] have successfully interpreted body-pose skeletons into textual representations using transformer networks.

Currently, most of the sign language translation methods have been using an architecture that uses intermediate gloss annotation of the language to produce the text from sign language. However, acquiring gloss annotations for any language is challenging and expensive. In this scenario, different models have emerged recently that do not use the gloss annotation to do the translation process. Gloss Free SLT based on Visual-Language Pretraining [57], GASLT [55], Gloss-Free-End-to-End sign language translation framework [31] various sign language translation methods are being done to mitigate this overhead of generating gloss annotations for the language.

B. Other Related Works

Sign Language Production (SLP) is a complex task that involves the generation of sign language from textual data. This complexity arises from the need to capture the contextual semantics of words and generate corresponding sequences of body-pose skeletons. Transformer and Generative Adversarial Network (GAN) models are often employed for this task due to their respective abilities to capture word context and generate effective body poses. Transformer models have successfully created sequences of body-pose skeletons [41], [43], [42]. GAN models, in conjunction with other models such as Convolutional Neural Networks (CNN) [45], Recurrent Neural Networks (RNN) [46], and transformer networks [42], have been applied to SLP tasks. The RWTH-PHOENIX-Weather-2014 dataset [25], a German dataset, has been predominantly used for training SLP models. All deep learning models referenced in this SLP subsection have utilized this benchmark dataset for training and validation.

III. Our Method

A. Problem Statement

The primary objective of our BTVSL dataset is to facilitate Sign Language Translation (SLT) for the Bangla language. Even though sign language translation has grown in different languages over the years, the Bangla language has been left out till now. That is why the main aim of our BTVSL dataset is to provide a comprehensive collection of sign language gestures and expressions so that researchers and developers can create more accurate and effective Sign Language Translation systems.

For SLT, we may represent a sentence as Y = $(y_1, y_2, ...y_T)$ where y_t represents a token and T is the number of tokens for Y. For any sentence Y, there is a corresponding sequence of frames X = $(x_1, x_2, ...x_U)$ where U is the total count of time steps for frames X and x_i represents the matrix representation of an image for a specific time step. The objective of SLT is to learn the probabilities $p(Y|X)$. In general, our goal of crafting our dataset **BTVSL** is to train an SLT model that would take Bangla sign language sequences as input and generate Bangla sentences as text output

B. The BTVSL Dataset

We introduce **BTVSL**, the inaugural sentence-level annotated dataset designed specifically for Bangla Sign Language Translation. The detailed statistics of BTVSL can be found in Table I.

TABLE I
DATASET STATISTICS

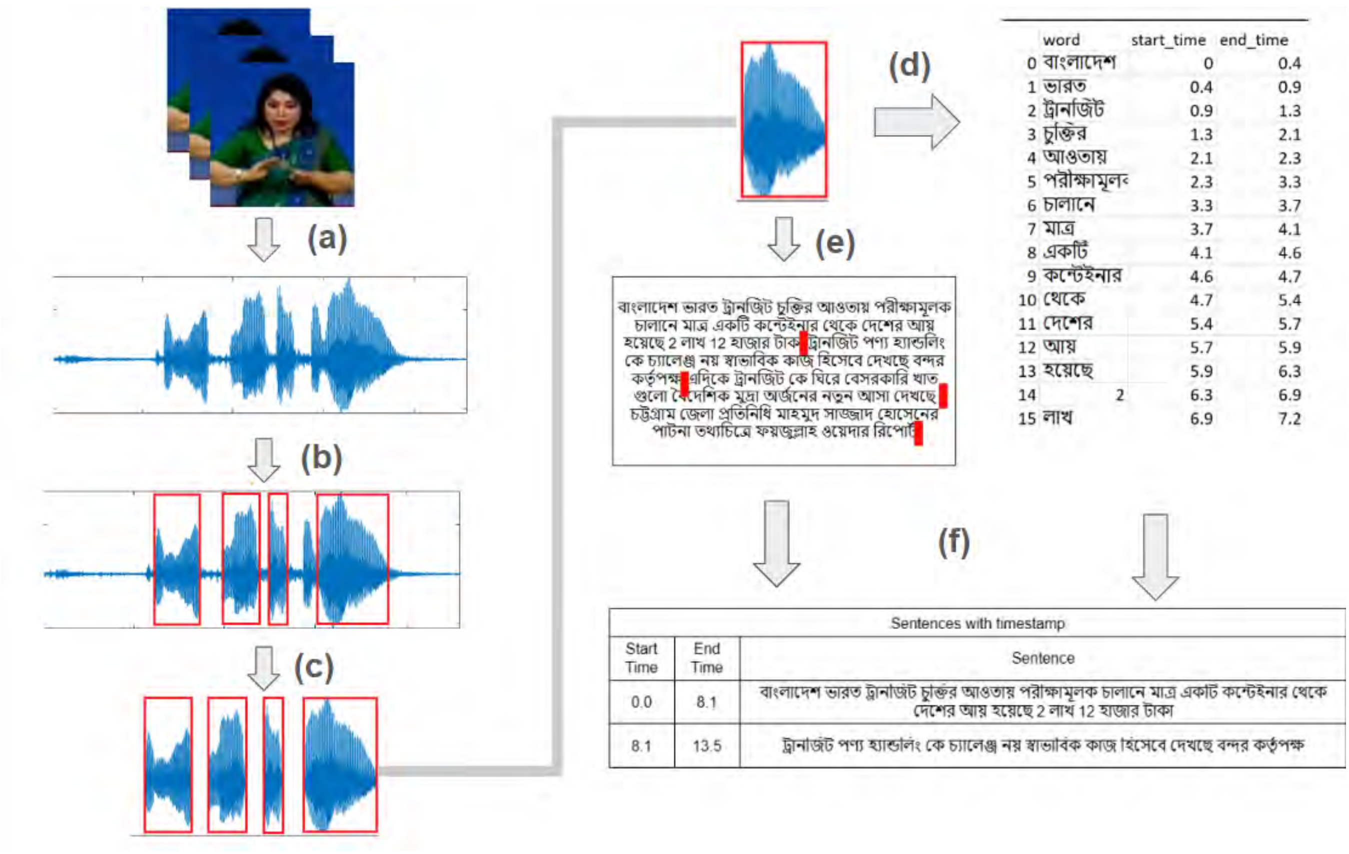
Total Video Length	60 Hours
Number of Signers	22
Total Frame Count	1.8 Million (approx)
Total Sentences	24085
Total Vocabulary	48623
Total Words	340172

The source of the videos for BTVSL is the YouTube Channel *Bangladesh Television*[1], which offers local Bangladeshi news with sign language interpretations. We manually identified and selected regions within these videos that contained a single signer against a fixed background. The frame rate was subsequently reduced to 10 FPS. A total of 134 videos were selected for inclusion in our dataset.

Each video in our dataset features a sign language interpreter working from a script. We used a voice activity detection API and a speech-to-text API to transcribe the spoken content into text. Subsequent post-processing was employed to refine and clean the transcribed texts.

Table II shows a comparison of our datasets with the widely used RWTH-PHOENIX-Weather 2014 and Chinese Sign Language (CSL) dataset.

[1] https://www.youtube.com/@BangladeshTelevision-BTV



	word	start_time	end_time
0	বাংলাদেশ	0	0.4
1	ভারত	0.4	0.9
2	ট্রানজিট	0.9	1.3
3	চুক্তির	1.3	2.1
4	আওতায়	2.1	2.3
5	পরীক্ষামূলক	2.3	3.3
6	চালানে	3.3	3.7
7	মাত্র	3.7	4.1
8	একটি	4.1	4.6
9	কন্টেইনার	4.6	4.7
10	থেকে	4.7	5.4
11	দেশের	5.4	5.7
12	আয়	5.7	5.9
13	হয়েছে	5.9	6.3
14	2	6.3	6.9
15	লাখ	6.9	7.2

বাংলাদেশ ভারত ট্রানজিট চুক্তির আওতায় পরীক্ষামূলক
চালানে মাত্র একটি কন্টেইনার থেকে দেশের আয়
হয়েছে 2 লাখ 12 হাজার টাকা ট্রানজিট পণ্য হ্যান্ডলিং
কে চ্যালেঞ্জ নয় স্বাভাবিক কাজ হিসেবে দেখছে বন্দর
কর্তৃপক্ষ এদিকে ট্রানজিট কে ঘিরে বেসরকারি খাত
গুলো বৈদেশিক মুদ্রা অর্জনের নতুন আসা দেখছে
চট্টগ্রাম জেলা প্রতিনিধি মাহমুদ সাজ্জাদ হোসেনের
পাটনা তথ্যচিত্রে ফয়জুল্লাহ ওয়েদার রিপোর্ট

Sentences with timestamp		
Start Time	End Time	Sentence
0.0	8.1	বাংলাদেশ ভারত ট্রানজিট চুক্তির আওতায় পরীক্ষামূলক চালানে মাত্র একটি কন্টেইনার থেকে দেশের আয় হয়েছে 2 লাখ 12 হাজার টাকা
8.1	13.5	ট্রানজিট পণ্য হ্যান্ডলিং কে চ্যালেঞ্জ নয় স্বাভাবিক কাজ হিসেবে দেখছে বন্দর কর্তৃপক্ষ

Fig. 2. Pipeline for data processing. a) Audio extraction from video, b) Voice activity detection using Hugging Face, c) Segment news reports in a news video, d) Detect starting and ending time for all Bangla words in a news report using Google API, e) Speech to Bangla text transcription using Google API and manually segment sentences from the continuous word list, f) Assign start time and end time to each of the sentences.

TABLE II
DATASET COMPARISONS

	Our Dataset	Phoenix-2014T [25]	CSL-Daily [58]
Total Video Length (Hours)	60	25	100+
Number of Signers	22	9	50
Total Frame Count (Million)	1.8 (approx)	1.1	1.9
Total Sentences	24085	8257	25000
Total Vocabulary	48623	2887	-
Total words	340172	-	175000 (approx)

C. News Report Segmentation

Each news video in our dataset comprises multiple news reports. We observed significant periods of silence between these reports. Consequently, we employed a voice activity detection model to segment each video into individual news reports. Of the various voice activity detection models available, we opted for the Hugging Face Voice activity detection [2], owing to its superior performance. Following the segmentation process, we manually eliminated segments that did not feature sign interpreters. This resulted in 5,761 news report segments derived from 134 news videos. These segments were subsequently used to obtain the transcripts discussed in the following subsection.

D. Transcript Generation

The playlist lacked Bangla subtitles for the news videos, and the BTV HQ archives did not retain scripts of previously broadcasted news reports. Consequently, we required a speech-to-text model to transcribe the audio content into text. We evaluated several cloud-based speech-to-text models, including Amazon Transcribe[3], Speech to text[4], Amberscript[5], and Rev[6]. We ultimately selected the Google Speech-to-Text API[7] due to its superior performance in transcribing Bangla speech, providing starting and ending timestamps for each pronounced word. Each news report contained 3-4 sentences that were inseparable using voice activity detection due to minimal or no silence periods between sentences. Furthermore, the Google Speech-to-Text API does not support automatic punctuation marks to denote sentence endings in the Bangla language. As a result, we manually separated

[2] https://huggingface.co/pyannote/voice-activity-detection

[3] https://aws.amazon.com/pm/transcribe

[4] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text

[5] https://www.amberscript.com/en/products/api-custom-models

[6] https://www.rev.com/services/speech-to-text-apis

[7] https://cloud.google.com/speech-to-text

the sentences of each continuous news report transcription. We assigned a start and end timestamp to each sentence. This process yielded 24085 sentences from 5,761 news report segments, with each sentence assigned its timeframe.

IV. Experiment and Results

In this section, we delve into the criteria for our model selection and the comprehensive training setup employed for our Bangla SLT dataset. Post-training, we assess the performance of our dataset by presenting the BLEU [34], ROUGE [29], ROUGE-L [30], and Recall metrics at Rank K (R@K, where a higher score is preferable). Subsequently, we analyze the SLT models' performance on our dataset.

BLEU and ROUGE scores are two of the most commonly used metrics in machine translation. BLEU-n represents the weighted average translation precision up to n-grams. Generally, we use uniform weights; the weights from 1-grams to n-grams are all $1/n$. ROUGE measures the overlap of n-grams between the generated summary and the reference summary. The goal is to assess how well the generated summary captures the essential information in the reference summary.

A. Model Selection

We employ four distinct models for training on our dataset: [8] TwoStream-SLT, [10] CiCo-Sign Language Retrieval, [55] Gloss Attention for Gloss-free Sign Language Translation and [57] Gloss-free Sign Language Translation. These models were selected due to their ability to formulate an end-to-end SLT solution that does not require gloss sequence ground truth throughout the modeling process. Given the financial implications of hiring sign language experts for gloss-level annotation of the dataset, these models present a cost-effective alternative.

B. Training Setup

Our dataset is comprised of 24085 sentences. We joined all the sentences from each video, shuffled them, and divided the videos into training, validation, and testing sets in a 70:15:15 ratio. Given that we trained the dataset on four different models, we prepared four sets for each division. We trained the models in a desktop computer powered with an Nvidia 2070 RTX graphics card, 32GB RAM, and a Ryzen 8 processor.

C. CiCO-Sign Language Retrieval [10]

Cross-Lingual Contrastive Learning (CLCL) is a methodology designed to establish a shared embedding space for sign videos and text. This approach facilitates the identification of detailed sign-to-word mappings. The procedure encompasses the extraction of features from sign videos and words, the introduction of cross-lingual similarity, and the application of contrastive learning. This process enables the model to learn and understand the relationship between sign language and its corresponding textual representation.

The extraction of sign features is achieved using a sign encoder and a Transformer. Text features, on the other hand, are generated via a lower-cased byte pair encoding (BPE) representation and an additional Transformer. The encoders are initialized with CLIP's image and text encoders to enhance transfer capability. The performance of various datasets and a comparison with our BTVSL dataset are illustrated in the table referenced as Tab. III below.

TABLE III
CiCo Model [10]: Performance metrics for models trained on different datasets

$Dataset$	R@1	R@5	R@10	$MedR$
How2Sign	56.6	69.9	74.7	1.0
PHOENIX2014T	69.5	86.6	92.1	1.0
BTVSL	40.1	51.45	65.12	4.4

The evaluation metrics utilized include Recall at 1 (R@1), Recall at 5 (R@5), Recall at 10 (R@10), and Median Rank ($MedR$). The performance of the CiCo Model varied across different datasets. For the "How2Sign" dataset, the model achieved a R@1 of 56.6%, R@5 of 69.9%, R@10 of 74.7%, and a Median Rank of 1.0. The model exhibited superior performance on the "PHOENIX2014T" dataset, with a R@1 of 69.5%, R@5 of 86.6%, R@10 of 92.1%, and a Median Rank of 1.0. However, the performance was relatively lower on the "BTVSL" dataset, with a R@1 of 40.1%, R@5 of 51.45%, R@10 of 65.12%, and a Median Rank of 4.4. These results suggest that the effectiveness of the CiCo Model is contingent on the specific dataset, with the "PHOENIX2014T" dataset yielding robust results.

D. TwoStream-SLT [8]

Two-stream SLT is a system that processes sign language through two distinct information streams: a visual stream for sign language gestures (video or image) and a linguistic stream for the signed content's transcriptions or semantic representations. In the context of our BTVSL dataset, the visual stream involves collecting data from various signers, preprocessing video frames or images to extract relevant features, and training a gesture recognition model to link visual patterns with specific Bangla sign gestures. Concurrently, the linguistic stream transcribes the sign language into Bangla text or semantic representations. This is followed by applying Natural Language Processing (NLP) techniques to enhance the understanding of the linguistic content. The integration phase ensures the temporal alignment of the visual and linguistic streams, synchronizing the recognized gestures with their corresponding linguistic elements. The final translation model uses this integrated information to generate translated outputs, which could be spoken Bangla or written text. Compared to other datasets, the Bangla dataset's performance is demonstrated in the table referenced as Tab. IV below.

The performance of the two-stream SLT model varied across different datasets. For the Phoenix-2014T dataset, the model achieved a ROUGE score of 52.01, with BLEU scores

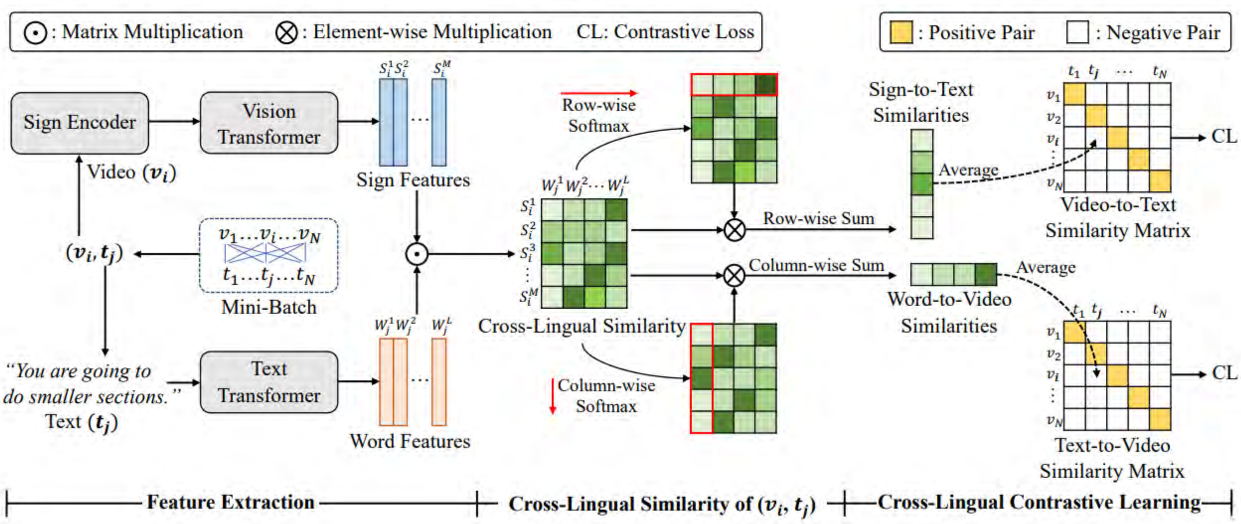

⊙: Matrix Multiplication
⊗: Element-wise Multiplication
CL: Contrastive Loss
: Positive Pair
: Negative Pair
Sign Encoder
Vision Transformer
Video (v_i)
$s_i^1 s_i^2 \cdots s_i^M$
Sign Features
(v_i, t_j)
$v_1 \ldots v_i \ldots v_N$
$t_1 \ldots t_j \ldots t_N$
Mini-Batch
"You are going to do smaller sections."
Text (t_j)
Text Transformer
$w_j^1 w_j^2 \cdots w_j^L$
Word Features
Row-wise Softmax
Cross-Lingual Similarity
Column-wise Softmax
Row-wise Sum
Column-wise Sum
Sign-to-Text Similarities
Average
Word-to-Video Similarities
Average
Video-to-Text Similarity Matrix
Text-to-Video Similarity Matrix
CL
CL
Feature Extraction
Cross-Lingual Similarity of (v_i, t_j)
Cross-Lingual Contrastive Learning

Fig. 3. Illustration of cross-lingual contrastive learning

ranging from 52.35 (BLEU-1) to 26.47 (BLEU-4). The model demonstrated superior performance on the CSL-Daily dataset, with a ROUGE score of 54.08 and BLEU scores ranging from 54.32 (BLEU-1) to 28.66 (BLEU-4). However, the model's performance was relatively lower on the BTVSL dataset, yielding a ROUGE score of 30.23 and BLEU scores ranging from 28.11 (BLEU-1) to 18.56 (BLEU-4). These metrics provide a comprehensive evaluation of the model's effectiveness in generating translations, highlighting variations in performance across different datasets. The robust scores on the CSL-Daily dataset suggest strong translation capabilities, while the lower scores on the BTVSL dataset indicate potential challenges in handling diverse datasets.

E. Gloss-free Sign Language Tranlsation [57]

In line with our primary objective of constructing a comprehensive pipeline to convert Bangla sign language videos into Bangla text, bypassing the use of intermediate gloss, we have chosen to employ the Gloss-free Sign Language Translation model [57]. This model has been expressly incorporated to handle our BTVSL dataset.

In the Gloss-free Sign Language Translation (GSFLT) model, we employ a 2D-CNN component that leverages the ResNet18 [18] architecture, pre-trained on ImageNet [13]. Adhering to the configuration of a prior study [58], we set the stride size to 2/4 and the kernel size to 6/2 for the Conv1D/Maxpooling layers within the temporal blocks. The Transformer encoder and decoder each comprise three layers, with a hidden size of 1024 and a feed-forward size of 4096. Each layer incorporates eight attention heads. A dropout rate of 0.1 is implemented to mitigate the risk of overfitting.

TABLE IV
TwoStream-SLT [8]: Performance metrics for models trained on different datasets

Dataset	*ROUGE*	*BLEU − 1*	*BLEU − 2*	*BLEU − 3*	*BLEU − 4*
Phoenix-2014T	52.01	52.35	39.76	31.85	26.47
CSL-Daily	54.08	54.32	41.99	34.15	28.66
BTVSL	30.23	28.11	25.33	20.12	18.56

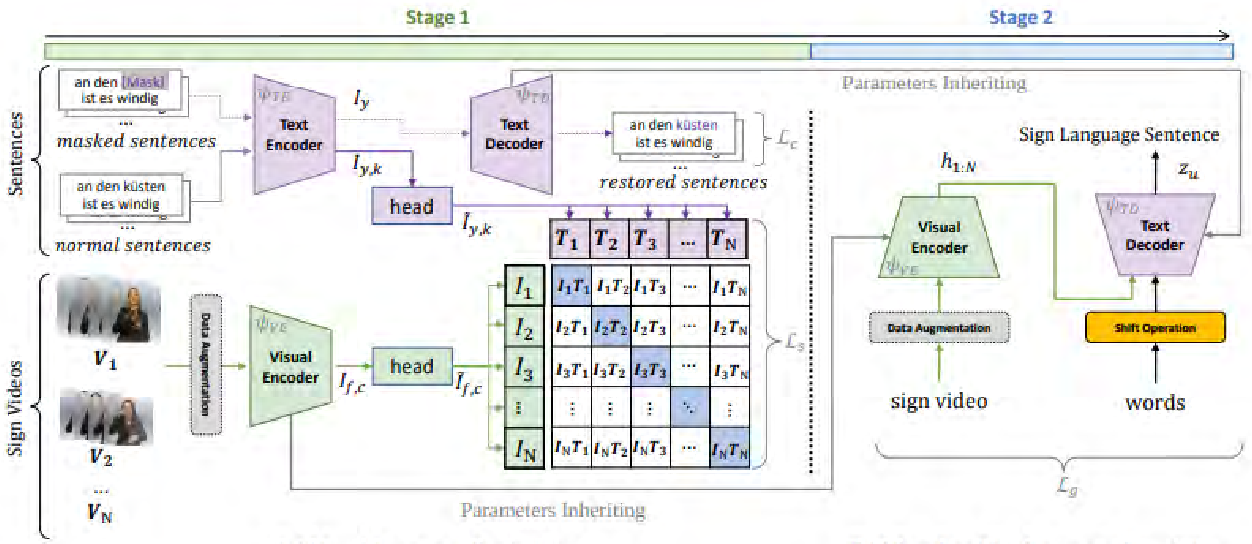
TABLE V
Gloss-free SLT [57]: Performance metrics for models trained on different datasets

Dataset	*ROUGE*	*BLEU − 1*	*BLEU − 2*	*BLEU − 3*	*BLEU − 4*
Phoenix-2014T	44.08	33.56	26.74	22.12	43.72
CSL-Daily	39.20	25.02	16.35	11.07	36.70
BTVSL	30.23	23.11	16.33	12.16	25.16

We undertake distinct pretraining tasks on the training subset of the BTVSL dataset. The mini-batch size is configured to 16, with Automatic Mixed Precision (AMP) technology employed to augment the batch size. Given the constraints associated with the original Bangla dataset, input sequences are initially resized to 224x224 during both the training and inference stages. Stochastic Gradient Descent (SGD) is utilized as the optimizer. The learning rate is subject to decay following a cosine schedule [32], ranging from a maximum of 0.01 to a minimum of 1e-5. This process is sustained for a total of 80 epochs.

Tab. V provides a comparative analysis of the results obtained from different datasets trained using this architecture.

The performance of the Gloss-free Sign Language Translation model, as evaluated by ROUGE and BLEU metrics, varies across different datasets. For the Phoenix-2014T dataset, the model achieved a ROUGE score of 44.08. The CSL-Daily and BTVSL datasets yielded lower ROUGE scores of 39.20 and 30.23, respectively. Regarding BLEU metrics ranging from BLEU-1 to BLEU-4, the Phoenix-2014T dataset outperformed the CSL-Daily and BTVSL datasets across all n-gram orders.



(a) Visual-Language Pre-training **(b) Gloss-free Sign Language Translation**

Fig. 4. Method Overview: GFSLT-VLP improves the SLT by (a) performing Visual-Language Pretraining in stage 1 first, and then (b) transferring parameters of the pre-trained Visual Encoder and Textual Decoder in stage 2. Wherein N indicates the number of samples in a mini-batch

F. Gloss Attention for Gloss-free Sign Language Translation [55]

The GASLT model is implemented using the PyTorch framework [35], based on the open-source code from [27] and [5]. The model follows the Transformer architecture with 512 hidden units, 8 heads, and 2 layers for the encoder and decoder. The gloss attention parameter (N) is set to 7. To prevent overfitting, dropout with 0.5 drop rates is applied to both the encoder and decoder layers.

We use the pre-trained I3D model from TSPNet [28] to extract visual features for a fair comparison. Visual features for models other than TSPNet are extracted using a sliding window of eight with a stride of two. The network is initialized with Xavier initialization [15]. Label smoothed cross-entropy loss [48] optimizes the SLT task with a smoothing parameter (ε) set to 0.4. During training, the batch size is set to 64.

The Adam optimizer [48] is utilized with an initial learning rate of 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-8}$. Weight decay is set to 10^{-3}. Similar plateau learning rate scheduling as in [6] is employed, with adjustments to patience (9) and decrease factor (0.5). Both translation cross-entropy loss and knowledge transfer loss L_{kt} weights are set to one. All experiments use the same random seed.

The table referenced as Tab. VI provides a comparative analysis of the results obtained from different datasets trained using this architecture.

The table illustrates the performance metrics of the Gloss Attention for Gloss-free Sign Language Translation (SLT) model on various datasets. Across the Phoenix-2014T dataset, the model exhibits a substantial ROUGE-L score of 39.86, indicating a strong correlation between the generated translations and reference sign language sequences. Additionally, the BLEU scores for n-gram precision (BLEU-1 to BLEU-4) are consistently high, underscoring the model's effectiveness in capturing unigram and higher-order gram overlaps.

TABLE VI
GLOSS ATTENTION FOR GLOSS-FREE SLT [55]: PERFORMANCE METRICS FOR MODELS TRAINED ON DIFFERENT DATASETS

Dataset	$ROUGE-L$	$BLEU-1$	$BLEU-2$	$BLEU-3$	$BLEU-4$
Phoenix-2014T	39.86	39.07	26.74	21.86	15.74
CSL-Daily	20.35	19.90	9.94	5.98	4.07
SP-10 [54]	16.98	21.72	10.92	6.61	4.35
BTVSL	22.14	20.11	9.67	6.32	3.25

On the CSL-Daily dataset, the model displays a lower ROUGE-L score of 20.35, suggesting a comparatively weaker correlation with reference gloss sequences. Correspondingly, the BLEU scores for this dataset are also lower than those observed on Phoenix-2014T, indicating that the model's performance is less robust in this context.

For the SP-10 dataset, the model achieves a moderate ROUGE-L score of 16.98, with BLEU scores at an intermediate level. This suggests a reasonable level of precision in capturing n-gram overlaps, showcasing the model's performance in this specific sign language dataset.

On the BTVSL dataset, the model performs well, evidenced by a ROUGE-L score of 22.14, indicating a good alignment with reference sign language sequences. However, the BLEU scores are relatively lower than Phoenix-2014T, suggesting some challenges in capturing specific n-gram overlaps on this dataset. But compared to the previous models on the same datasets, we can see that the BTVSL dataset almost outperforms two of the three prominent datasets in this model.

G. Discussion on Results

In the CiCO-Sign Language Retrieval model, our dataset performs with a score of $R@1$ of 40.1%, $R@5$ of 51.45%, $R@10$ of 65.12%, and a Median Rank of 4.4. This performance may indicate potential challenges or unique aspects of the Bangla Sign Language as captured by the BTVSL dataset. Our dataset was prepared using a publicly available YouTube channel, while other datasets were typically prepared in a dedicated studio setting. This has led to differences in video resolution, the accuracy of sentence-level mapping, and the diversity of the dataset. These factors could have contributed to the observed variations in model performance across different datasets.

In the context of the TwoStream-SLT model, the summarization scores on the BTVSL dataset were comparatively lower, with a ROUGE score of 30.23 and BLEU scores ranging from 28.11 (BLEU-1) to 18.56 (BLEU-4). The BTVSL dataset presents unique challenges in summarization, potentially due to the diversity of sign language expressions it encompasses. Potential areas for improvement could include refining the processing of Bangla transcriptions in the linguistic stream and adapting the visual stream to better capture the intricacies of diverse sign gestures within the BTVSL dataset to enhance performance. Furthermore, strengthening the synchronization between the visible and linguistic elements could improve summarization.

The Gloss-free Sign Language Translation (GSFLT) model's performance on the BTVSL dataset, as evaluated by ROUGE and BLEU metrics, indicates room for improvement. The model achieved a ROUGE score of 30.23 and BLEU scores ranging from 23.11 (BLEU-1) to 25.16 (BLEU-4). These results suggest that the model encounters challenges in directly translating Bangla sign language videos into text without using an intermediate gloss. This could be due to unique linguistic nuances present in the BTVSL dataset. Further investigation into these specific linguistic nuances and subsequent refinement of the translation model could be beneficial to enhance translation quality.

The final model, Gloss Attention for Gloss-free Sign Language Translation, shows that the BTVSL dataset can outperform datasets even with the constraints the datasets have. From the ROUGE-L and BLEU metrics, we find that the Phoenix-2014T dataset tops the metrics with around a score of 39. The BTVSL dataset does better than the CSL-Daily and SP-10 datasets with an improvement of $+2$ in the score for the CSL-Daily and $+6$ in the score for the SP-10 in this model. This result shows that no matter what models we use, the BTVSL dataset almost always performs around an average BLEU score of 20.

Overall insights indicate a common challenge across models, with BTVSL consistently presenting challenges, producing an average score in all the models, indicating the need for dataset-specific optimizations. The diverse nature of sign language expressions in BTVSL may require model adjustments to capture better and understand the varied linguistic and visual elements. Future improvements could involve more extensive data collection within the Bangla Sign Language context, including a broader range of sign gestures and linguistic variations. Fine-tuning models specifically for the nuances of BTVSL and potentially incorporating user feedback for continuous refinement may lead to enhanced model performance.

V. Conclusion and Future Works

We present a novel dataset from online videos for Bangla sign language translation. This resource generates continuous sentences and has the potential for further exploration. Future work includes using Generative Adversarial Networks (GANs) to convert these sentences into continuous video sequences, creating realistic sign language avatars. Despite current inaccuracies, we believe improvements can be made through precise video annotations.

Our dataset is a dynamic tool in the evolving field of sign language translation, with potential for enhancement. One strategy is to use image processing networks to add depth modalities, improving the translation of sign language expressions.

We aim to develop a deep learning-based framework for seamless translation between Bangla texts and sign language. The goal is to map texts directly to a sign language interpreter's skeleton data and vice versa. If successful, this approach could revolutionize interpretation, enabling Bangla sign language users to communicate more easily with others and promoting inclusivity. This work contributes to sign language translation, facilitating more accessible communication for diverse linguistic communities.

Acknowledgements

This research work is performed at the Bangladesh University of Engineering and Technology (BUET) and is supported by the ICT Innovation Fund from the Government of the People's Republic of Bangladesh and CodeCrafters International Ltd., Bangladesh.

References

[1] S. Albanie, G. Varol, L. Momeni, T. Afouras, H. Bull, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, et al. Bobsl: Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.

[2] K. Bantupalli and Y. Xie. American sign language recognition using deep learning and computer vision. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4896–4899. IEEE, 2018.

[3] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018.

[4] N. C. Camgoz, O. Koller, S. Hadfield, and R. 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.

[5] N. C. Camgoz, O. Koller, S. Hadfield, and R. 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*, page 10023–10033, 2020.

[6] N. C. Camgoz, O. Koller, S. Hadfield, and R. 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*, page 10023–10033, 2020.

[7] N. C. Camgöz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021.

[8] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056, 2022.

[9] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai. Fully convolutional networks for continuous sign language recognition. In *European Conference on Computer Vision*, pages 697–714. Springer, 2020.

[10] Y. Cheng, F. Wei, J. Bao, D. Chen, and W. 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.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

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

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, page 248–255. IEEE, 2009.

[14] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. Speech recognition techniques for a sign language recognition system. *hand*, 60:80, 2007.

[15] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, page 249–256, Mar. 2010.

[16] I. Gruber, Z. Krnoul, M. Hrúz, J. Kanis, and M. Bohacek. Mutual support of data modalities in the task of sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3424–3433, 2021.

[17] A. Hasib, S. S. Khan, J. F. Eva, M. Khatun, A. Haque, N. Shahrin, R. Rahman, H. Murad, M. Islam, M. R. Hussein, et al. Bdsl 49: A comprehensive dataset of bangla sign language. *arXiv preprint arXiv:2208.06827*, 2022.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 770–778, 2016.

[19] J. Huang, W. Zhou, H. Li, and W. Li. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2822–2832, 2018.

[20] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[21] Z. Huang, W. Xu, and K. Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[22] M. Islam, S. S. S. Mousumi, N. A. Jessan, A. Rabby, S. Abujar, S. A. Hossain, et al. Ishara-bochon: The first multipurpose open access dataset for bangla sign language isolated digits. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 420–428. Springer, 2018.

[23] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3413–3423, 2021.

[24] S.-K. Ko, J. G. Son, and H. Jung. Sign language recognition with recurrent neural network using human keypoint detection. In *Proceedings of the 2018 conference on research in adaptive and convergent systems*, pages 326–328, 2018.

[25] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[26] O. Koller, S. Zargaran, H. Ney, and R. Bowden. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms. *International Journal of Computer Vision*, 126(12):1311–1325, 2018.

[27] J. Kreutzer, J. Bastings, and S. Riezler. Joey nmt: A minimalist nmt toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, page 109–114. Association for Computational Linguistics, Nov. 2019.

[28] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In *Advances in Neural Information Processing Systems*, volume 33, page 12034–12045. Curran Associates, Inc., 2020.

[29] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[30] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, page 605–612, Barcelona, Spain, 2004.

[31] K. Lin, X. Wang, L. Zhu, K. Sun, B. Zhang, and Y. Yang. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada, July 2023. Association for Computational Linguistics.

[32] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint*, 2016.

[33] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri. A modified lstm model for continuous sign language recognition using leap motion. *IEEE Sensors Journal*, 19(16):7056–7063, 2019.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, page 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[36] J. Pu, W. Zhou, and H. Li. Dilated convolutional network with iterative optimization for continuous sign language recognition. In *IJCAI*, volume 3, page 7, 2018.

[37] J. Pu, W. Zhou, and H. 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.

[38] A. M. Rafi, N. Nawal, N. S. N. Bayev, L. Nima, C. Shahnaz, and S. A. Fattah. Image-based bengali sign language alphabet recognition for deaf and dumb community. In *2019 IEEE global humanitarian technology conference (GHTC)*, pages 1–7. IEEE, 2019.

[39] G. A. Rao, K. Syamala, P. Kishore, and A. Sastry. Deep convolutional neural networks for sign language recognition. In *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*, pages 194–197. IEEE, 2018.

[40] R. Rastgoo, K. Kiani, and S. Escalera. Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, 150:113336, 2020.

[41] B. Saunders, N. C. Camgoz, and R. Bowden. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer, 2020.

[42] B. Saunders, N. C. Camgoz, and R. Bowden. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *International journal of computer vision*, 129(7):2113–2135, 2021.

[43] B. Saunders, N. C. Camgoz, and R. Bowden. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929, 2021.

[44] S. Sharma and K. Kumar. Asl-3dcnn: American sign language recognition technique using 3-d convolutional neural networks. *Multimedia Tools and Applications*, 80(17):26319–26331, 2021.

[45] S. Stoll, N. C. Camgöz, S. Hadfield, and R. Bowden. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. University of Surrey, 2018.

[46] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908, 2020.

[47] M. Sundermeyer, R. Schlüter, and H. Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,

page 2818–2826, 2016.

[49] A. Voskou, K. P. Panousis, D. Kosmopoulos, D. N. Metaxas, and S. Chatzis. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955, 2021.

[50] A. Wadhawan and P. Kumar. Deep learning-based sign language recognition system for static signs. *Neural computing and applications*, 32(12):7957–7968, 2020.

[51] F. Wen, Z. Zhang, T. He, and C. Lee. Ai enabled sign language recognition and vr space bidirectional communication using triboelectric smart glove. *Nature communications*, 12(1):1–13, 2021.

[52] R. Wilbur and A. C. Kak. Purdue rvl-slll american sign language database. *none*, 2006.

[53] Q. Xiao, M. Qin, and Y. Yin. Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks*, 125:41–55, 2020.

[54] A. Yin, Z. Zhao, W. Jin, M. Zhang, X. Zeng, and X. He. Mlslt: Towards multilingual sign language translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, page 5099–5109, New Orleans, LA, USA, June 18-24 2022. IEEE.

[55] A. Yin, T. Zhong, L. Tang, W. Jin, T. Jin, and Z. Zhao. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2551–2562, 2023.

[56] K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, and M. Alikhani. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*, 2021.

[57] B. Zhou, Z. Chen, A. Clapés, J. Wan, Y. Liang, S. Escalera, Z. Lei, and D. Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881, October 2023.

[58] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. 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.