Is context all you need? Scaling Neural Sign Language Translation to Large Domains of Discourse

Ozge Mercanoglu Sincan[1], Necati Cihan Camgoz[2], Richard Bowden[1]
[1]University of Surrey, United Kingdom
[2]Meta Reality Labs, Switzerland
{o.mercanoglusincan, r.bowden}@surrey.ac.uk, neccam@meta.com

Abstract

Sign Language Translation (SLT) is a challenging task that aims to generate spoken language sentences from sign language videos, both of which have different grammar and word/gloss order. From a Neural Machine Translation (NMT) perspective, the straightforward way of training translation models is to use sign language phrase-spoken language sentence pairs. However, human interpreters heavily rely on the context to understand the conveyed information, especially for sign language interpretation, where the vocabulary size may be significantly smaller than their spoken language equivalent.

Taking direct inspiration from how humans translate, we propose a novel multi-modal transformer architecture that tackles the translation task in a context-aware manner, as a human would. We use the context from previous sequences and confident predictions to disambiguate weaker visual cues. To achieve this we use complementary transformer encoders, namely: (1) A Video Encoder, that captures the low-level video features at the frame-level, (2) A Spotting Encoder, that models the recognized sign glosses in the video, and (3) A Context Encoder, which captures the context of the preceding sign sequences. We combine the information coming from these encoders in a final transformer decoder to generate spoken language translations.

We evaluate our approach on the recently published large-scale BOBSL dataset, which contains ∼1.2M sequences, and on the SRF dataset, which was part of the WMT-SLT 2022 challenge. We report significant improvements on state-of-the-art translation performance using contextual information, nearly doubling the reported BLEU-4 scores of baseline approaches.

1. Introduction

Sign languages are visual languages and the primary languages of Deaf communities. They are languages in their own right, as rich as any spoken language, and can vary considerably between countries with strong dialect differences within a country [16]. They have their own lexicons and grammatical constructs, thus converting between sign and spoken language is a translation problem.

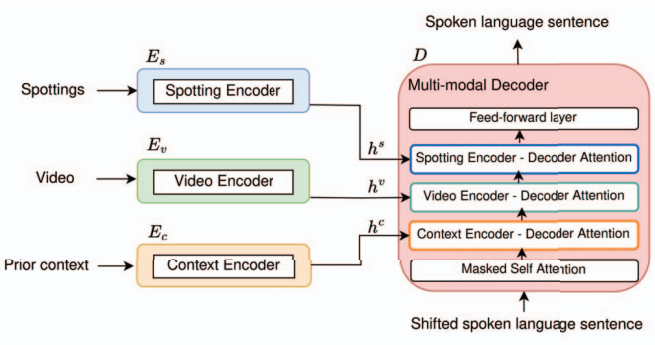

Figure 1. An overview of the proposed multi-modal sign language translation architecture.

Sign Language Recognition (SLR) [44] and Sign Language Translation (SLT) are active research areas within computer vision [5, 10, 54, 57]. While SLR focuses on the recognition of signs within a video, SLT aims to generate meaningful spoken language interpretation of a given signed phrase or vice versa. In our work, we focus on the former part of SLT, namely translating continuous sign videos into spoken language sentences.

Automatic SLT is a challenging problem for a number of reasons. Firstly, as stated, sign languages have their own grammar, they are not translated simply word-by-word by replacing words with signs [49]. Secondly sign languages contain many channels that are used in combination i.e. hand articulation, facial expression, and body posture are all used in combination and their use can vary depending on context. For example, the hand shape of a sign may change depending on the context. A good example of this would be the verb '*to give*'. The verb '*give*' is directional and the direction of the motion is subject to the placement of ob-

jects and use of space in front of the signer. But the hand shape can also change depending on the type of object being given. Thirdly, motions can be subtle or fast and this leads to motion blur. Finally, many sign can also look very similar. All of these factors make it difficult to recognize the sign that is being performed without the context in which it is used.

Human interpretation or translation of sign languages heavily relies on context, as it is fundamental to all language understanding. Consider the use of homophones in spoken language. An active listener has no issues in the disambiguation of homophones despite the fact there are no auditory cues to help. This is because we are able to use the context to disambiguate the meaning of the homophone. However, much of the SLT work to date has neglected such context, focusing largely on sentence pairs. In fact, most machine translation datasets shuffle the order of sentences, making it impossible to utilize the context from the previous sentences.

In this work, we propose a novel sign language translation architecture that incorporates important contextual information. It combines weak visual cues from a 3D convolutional backbone with strong cues from the context and sparse sign spottings. An overview of the approach can be seen in Figure 1.

We evaluate our approach on the largest available sign language dataset, BOBSL [3], which covers a wide domain of various topics. We obtain significant performance improvements by incorporating context and automatic spottings (1.27 vs. 2.88 in BLEU-4). We also evaluate our approach on the WMT-SLT 2022 challenge data, specifically the SRF partition, and surpass the reported performance of all challenge participants.

The contributions of this paper can be summarized as:

- We propose a novel multi-modal transformer network that incorporates the context of the prior information and automatic spottings.
- We conduct extensive experiments to examine the effects of different approaches to capturing context.
- Our approach achieves state-of-the-art translation performance on two datasets, namely BOBSL, the largest publicly available sign language translation dataset, and the WMT-SLT 2022 challenge data.

The remainder of the paper is organized as follows: In Section 2, we summarize the related work. In Section 3, we describe our proposed sign language translation network. In Section 4, we provide information about the datasets we use and provide model training details. Section 5 presents the experimental results of the proposed method and we conclude the paper in Section 6.

2. Related Work

Sign Language Recognition (SLR) has seen consistent research effort from the computer vision community for decades [12, 48, 56]. The advances in models and techniques, also the release of recent isolated [3, 22, 30, 46], and continuous [23, 28] SLR datasets have led to significant improvements in the accuracy and robustness of sign language recognition systems.

SLR can be grouped into two sub-problem; isolated and continuous SLR. While the isolated SLR videos contain only a single sign, continuous SLR videos contain multiple sign sequences. After the emergence of 2D convolutional neural networks (CNNs), 2D CNNs were quickly applied to model the visual appearance in SLR [35, 39, 40, 46]. Sequence models such as the recurrent neural network (RNN) [40], long short-term memory (LSTM) [46], hidden markov model (HMM) [51] have all been used to encode temporal information. Following 2D CNNs, 3D CNNs were developed and have achieved state-of-the-art performance on a wide range of computer vision tasks, including sign language recognition [2, 22, 26, 30].

In addition to images, researchers have also used other input modalities for SLR, such as depth, skeleton, optical flow, and motion history image (MHI) to improve recognition accuracy [21, 25, 35, 46, 47]. Some studies also introduced the use of different cues such as cropped hands and faces [11, 18], or an attention mechanism [13, 47] to obtain better discriminative features.

These advances in the field of isolated SLR have also been applied to continuous SLR. Since continuous SLR videos contain multiple co-articulated signs, it is a more challenging problem. The explicit alignment between the video sequence and gloss sequence generally does not exist. In order to tackle this problem, Connectionist Temporal Classification (CTC) [17] is widely used [11, 20, 43, 59].

Sign Language Translation (SLT) is still in its infancy due to the lack of large-scale sign language translation datasets. While machine translation datasets for spoken languages contain many millions of sentence pairs such as 22.5M for English-French, and 4.5M for English-German pairs (WMT shared tasks [4]), the first public SLT dataset PHOENIX14-T [5], which was released in 2018, had only 8K sentences and its domain of discourse was limited to weather forecast. The authors handle the SLT as a Neural Machine Translation (NMT) problem and proposed the first end-to-end SLT model by combining CNNs with the attention-based encoder-decoder network with RNNs.

One of the most significant advances in NMT was the introduction of the Transformer network by Vaswani et. al [52], which is based solely on attention mechanisms and waives recurrent networks, for the sequence transduction problem. Camgoz et al. [7] applied transformer architecture to the sign language translation problem. In recent years,

transformers have become popular in SLT [15, 54, 55, 57]. Some studies tackle SLT with a two-stage approach, i.e., in the first part glosses are recognized from sign videos (Sign2Gloss), and then glosses are mapped into a spoken language sentence (Gloss2Text) [5, 55]. On the other hand, some studies deal with an end-to-end solution that predicts the spoken language sentence from sign video inputs [7, 57].

Zhou et al. [57] proposed a two-stage approach, but unlike others, their approach is based on back-translation. They convert spoken language text to sign sequences with both text-to-gloss and gloss-to-sign steps to generate synthetic data. They used the synthetic samples as additional data and trained an end-to-end SLT method based on the transformer. Zhou et. al [58] and Camgoz et. al [6] also utilized multiple cues for the SLT task, such as hands and face. To the best of our knowledge, and perhaps surprisingly, using context has not been exploited in the literature. However, Papastratis et. al [36] did use the previous sentence to initialize the hidden state of a BLSTM for predictions of the next video sequence to improve recognition accuracy in a continuous SLR. They obtain slightly better results when the context-aware gloss predictions were fed into the transformer for SLT.

Datasets: PHOENIX14-T [5] became the most commonly used dataset in the literature. The performance on this dataset is generally satisfactory to provide a usable translation, e.g., Chen et. al [10] obtained 28.39 in terms of BLEU-4 score. However, due to its limited domain of discourse, models trained on PHOENIX14-T have little real-world applicability. To address this, researchers released several datasets in recent years [3, 8, 33]. The largest to date is BOBSL [3], a broadcast interpretation-based large-scale British Sign Language (BSL) dataset. Their SLT baseline is based on the transformer network and obtains only 1.0 in terms of BLEU-4. Recently, Swiss German Sign Language (DSGS) broadcast datasets were introduced in the first SLT-WMT shared task [33], where all the submissions scored under 0.56 in terms of BLEU-4. Yin et. al [54] collect the first multi-lingual dataset for multiple sign language translations and proposed the first multi-lingual SLT model. Although significant progress has been made in the area of SLT, there is still room for further improvement.

3. Method

Most sign translation datasets and especially those based on broadcast interpretation [3, 5, 33], contain a set of consecutive sign phrase videos $(V_1, ..., V_M)$ and spoken language sentences $(S_1, ..., S_N)$. In some datasets, such as Phoenix2014T [5], sign phrase videos and their spoken language translations are paired and the order of the pairs are shuffled and distributed between training and evaluation sets. Unfortunately, this destroys the context of the sentence. Datasets like BOBSL [3] release the video and sentence sets with only weak alignment. Although this is generally regarded as a weakness, making subsequent learning from the data more challenging, it has a fundamental advantage that we make use of in this work: it allows the use of context to improve the translation.

Given an input video $V = (x_1, x_2, ..., x_T)$ with T frames, the aim of a sign language translation is to learn the conditional probability $p(S|V)$ in order to generate a spoken language sequence $S = (w_1, w_2, ..., w_U)$ with U words.

We propose to take advantage of the contextual information that comes from the preceding context, $S_C = (S_{n-1}, S_{n-2}, S_{n-3}, ...)$. We also make use of sparse sign spottings, $Sp = (g_1, ..., g_K)$, automatically recognized from the current video V using a state-of-the-art model. Thus, we extend the classical translation formalization to one of learning the conditional probability $p(S|V, S_C, Sp)$. This conditioning allows weak and ambiguous visual cues in V to be disambiguated based on context.

Our translation network is based on a transformer architecture and contains three separate encoders, E_v, E_c, E_s for each of the different input cues, i.e., video, context, and spottings, and a multimodal decoder, D, which learns the mapping between all input source representations and the target spoken language sentence. A detailed overview of our model is shown in Figure 2.

3.1. Embedding Layers

Following the classic neural machine translation methods, we first project source and target sequences to a dense continuous space via embedding layers. In order to represent video sequences, we utilize pretrained CNNs. For linguistic concepts that originate from written text in the form of the preceding and target spoken language sentences and spotted sign glosses, we use word embedding layers.

Sign Embedding: To convert a given video, V, to its feature representation, we use the I3D model [9] as a backbone due to its recent success in sign recognition tasks. We first divide the videos into smaller video clips, $c_t = (x_t, ..., x_{t+L-1})$ of size L. In our experiments we use a window size of $L = 16$ to obtain the sign video embedding:

$$f_t = \text{SignEmbedding}(c_t) \tag{1}$$

We stride $SignEmbedding$ over the full video V with the step size of 4, thus yielding a final feature set of $f_{1:\frac{T-L}{4}+1}$.

We considered two types of features as the output of our sign embedding layer, namely a) 1024-dimensional representation that is extracted from the last layer before classification, and b) C-dimensional class probabilities after the softmax activation function. We conduct experiments using both of these feature representations in our translation pipeline and compare their performance.

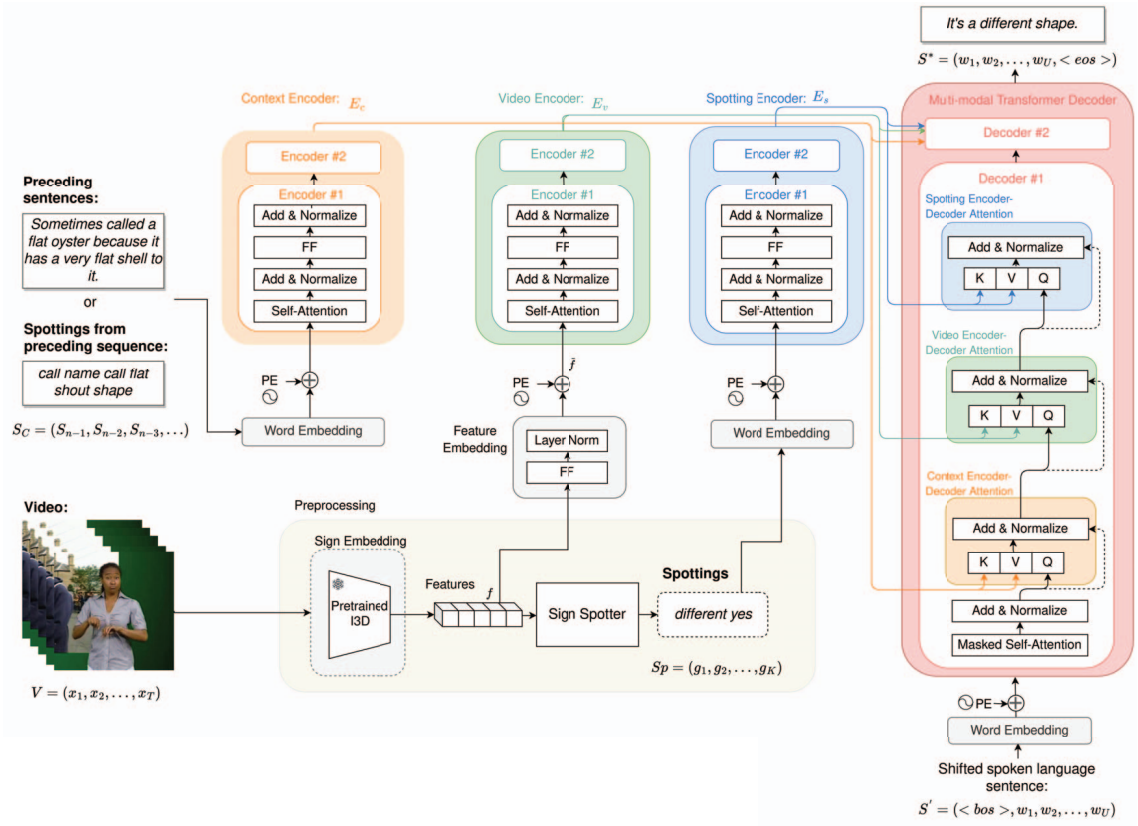

Figure 2. A detailed overview of the proposed multi-modal sign language translation architecture.

Feature Embedding: To avoid biases caused by dimensionality we project the extracted feature representations into the same size denser space using a linear layer. We also employ layer normalization to transform them to be of the same scale. This feature embedding operation can be formalized as:

$$\hat{f}_t = \text{FeatureEmbedding}(f_t) \tag{2}$$

Word Embedding: We first tokenize our spoken language sentences using a pretrained BERT model [14]. More specifically we employ the *BERT-base-cased* and *BERT-base-german-cased* from the Huggingface's Transformer library [53], which uses WordPiece tokenization. The word embedding layer is shared between all the text input cues, such as spottings, context sentences, and shifted target sentences.

Positional Encoding: In order to provide sequential order information to our networks we use the standard positional encoding method as proposed in [52] in the form of shifted sine and cosine waves. This is added after the feature and word embedding layers. This positional encoding can be formalized as:

$$\bar{f}_t = \text{PositionalEncoding}(\hat{f}_t) \tag{3}$$

3.2. Translation Network

After embedding layers, positionally encoded features and word vectors are sent to the transformer encoders. Our encoders have a stack of two identical layers each of which has a multi-headed self-attention and a fully connected feed-forward layer. Each of these two sub-layers is followed by a residual connection and layer normalization.

Video-Encoder: The video encoder network, E_v, takes the positionally encoded feature vectors $\bar{f}_{1:\frac{T-L}{4}+1}$ that come from the feature embedding layer and produces a spatial-temporal representation $h^v_{1:\frac{T-L}{4}+1}$ that captures the motion and content of the video.

Context-Encoder: The context encoder, E_c, takes positionally encoded-word embedding results from preceding context, S_C, and produces representations, h^c, which capture the context in which the currently signed phrase is performed.

Spotting-Encoder: The spotting encoder network, E_s, takes positionally encoded spotting embeddings Sp and produces representations h^s that correspond to confident but sparse sign detections that have been spotted in the current video, V, that we are attempting to translate. See Section 4.2 for details of the sign spotting technique [32] we used in our experiments.

Decoder: After encoding each input modality, the output of the encoder layers h^c, h^v, h^s, and the positionally encoded and shifted spoken sentence is then sent to the transformer decoder D. We extend the classical transformer decoder architecture [52] by introducing several encoder-decoder attention layers, which combine and enrich the representations coming from complementary cues of information. The network flow can be formalized as:

$$\begin{aligned} h^c &= \text{ContextEncoder}(S_C) \\ h^v &= \text{VideoEncoder}(V) \\ h^s &= \text{SpottingEncoder}(Sp) \\ S^* &= \text{Decoder}(h^c, h^v, h^s, S') \end{aligned} \tag{4}$$

where S' and S^* correspond to the shifted and predicted target sentences, respectively. In words, firstly, the word embeddings extracted from the shifted spoken language embedding S' are passed to the masked self-attention layer. Then, our first encoder-decoder attention layer takes outputs of the masked self-attention and context encoder, h^c. The output of the context encoder-decoder attention is sent to the video encoder-decoder attention to be used as a query, while the key and the value come from the video encoder, h^v. In a similar way, the spotting encoder-decoder attention performs attention operations over h^s and the previous layer. Finally, the last representation of the transformer decoder is projected to the space of the target vocabulary using a linear layer to predict the target spoken language sentence S^*, one word at a time.

We train our network using cross-entropy loss as proposed in [52], by comparing the predicted target sentence S^* against the ground truth sentence S at the word level.

4. Dataset and Implementation Details

4.1. Datasets

SRF is a Swiss German Sign Language (DSGS) dataset that was recently released for the WMT-SLT 2022 challenge [33] as one of the training corpora. It contains daily news and weather forecast broadcast. It includes 16 hours of sign footage, divided into 29 episodes, performed by three signers. In total 7,071 subtitles were manually aligned by Deaf annotators. Separate development and test sets were provided in the WMT-SLT. We use the SRF dataset for training and used the official development and test sets for the evaluation of the model to be able to compare our approach against the methods presented in the challenge.

BOBSL [3] is a large-scale British Sign Language (BSL) dataset that consists of BSL-interpreted BBC broadcast footage covering a wide range of topics. The dataset has an approximate duration of 1,400 hours and contains around 1.2M sentences. While the training and validation set's subtitles are audio-aligned, the test data is manually aligned and contains 20,870 sentences with a vocabulary size of 13,641.

4.2. Sign Spotter

Momeni et al. [32] released automatically extracted dense annotations for the BOBSL dataset. We use these annotations as the spotting input on the BOBSL experiments.

The key idea is that a set of video clips with a particular sign must have a correlation at the time when the sign is performed. Taking inspiration from [32], we create similar automatic dense annotations for the SRF dataset by correlating the I3D features and examplar subtitles. To do this, firstly we lemmatize and lowercase each word in the subtitle sentences and extract a vocabulary list. German language has compound words by concatenating two words. In order to reduce the number of singletons in the vocabulary list, we use the compound-split library [1]. Then, for each word w, we take a reference video clip V_0 that contains w in its subtitle sentence. We choose random $N = 9$ positive video examples $V_1, V_2, .., V_N$ that contain the word w in their subtitles, and $3 * N$ negative video examples that do not contain w to avoid annotating non-lexical signs. We compute the cosine similarities between reference and examplar video features. We apply a voting scheme among the videos with cosine similarity above 0.5 to find the location of the given word in the reference video.

4.3. Implementation Details

Sign Embeddings: For full-body video inputs, we pretrain two different I3D models which we call BSL-I3D and DGS-I3D on two different sign language datasets, namely BOBSL [3] and MeineDGS [29]. While training the BSL-I3D model we use the annotations released with the dataset [3] which has a vocabulary size of 2,281. For MeineDGS we use the linguistic annotation available with the dataset. In order to obtain a similar size vocabulary of 2,301 classes, we choose classes that have more than 12 occurrences.

We resize the input images to 224×224 and follow the training instructions of [3] with some small modifications; we use the Swish activation function instead of ReLU and change the learning rate scheduler to reduce on a plateau. We also use label smoothing of 0.1 in order to help reduce overfitting.

Training and Network Details: Our model is implemented using PyTorch [38]. We use the Adam [27] optimizer with an initial learning rate of 3×10^{-4} (β_1 = 0.9, β_2 = 0.999, ϵ = 10^{-8}) with a batch size 16 on SRF; and learning rate 6×10^{-4} with batch size 64 on the BOBSL dataset. We reduce the learning rate by a factor of 0.7, if the BLEU-4 score does not increase for five epochs. This step continues until the learning rate drops below 10^{-5}.

For transformer encoders and the decoder, we use two layers with 8 heads. We conduct an ablation study to choose the size of the hidden layers and the feed-forward layers (in section 5.1). We choose 512 and 1024, respectively. We use 0.1 for the dropout rate.

During training, we use a greedy search to evaluate translation performance on the development set. At inference, we evaluated both a greedy search and a beam search (decoding size of 2 and 3) for our video-to-text approach. However, we did not observe a significant improvement in scores. Therefore, we provide greedy search performances on both validation and test set.

Metrics: We use BLEU-1, BLEU-4 [37], ROUGE [31], and CHRF [41] scores, which are commonly used metrics for machine translation, to evaluate the performance of our model. As ROUGE, we use ROUGE-L F1 score; as BLEU score we use the sacreBLUE [42] implementation.

5. Experiment Results

We run our experiments in an end-to-end manner on two recently released sign language datasets, namely the SRF partition of WMT-SLT [33] and BOBSL [3], which is the largest sign language dataset available. For each dataset, we train baseline models that have one encoder and one decoder, and take only one input source, i.e., the preceding context (using the preceding spoken sentence or preceding spottings), current spotting, or video. We name our single modality models as *Context-to-Text, Spot-to-Text, Video-to-Text*, respectively.

Then, we investigate the impact of integrating context information to the *Spot-to-Text* or *Video-to-Text* approaches by adding a context encoder and using a dual-mode transformer decoder with the related encoder-decoder attention layers. Finally, we investigate using all sources simultaneously to gain more information. We combine all three sources using three separate encoders and a decoder. We name our final model as *Context+Video+Spot-to-Text*.

5.1. Experiments on SRF partition of WMT-SLT

Video-to-Text: We evaluate our *Video-to-Text* model which takes only the video source and tries to generate spoken language in an end-to-end manner.

First, we conduct ablations studies on the SRF partition of the WMT-SLT dataset using different types of input channels for the *Video-to-Text* model. We run our experiments with different numbers of hidden size (HS) and feed-forward (FF) units, with 64×128, 128×256, 256×512, 512×1024, 512×2048. We obtain similar results with 512×1024 and 512×2048, where 512×1024 is slightly better. Therefore, for the rest of our experiments, we use 512×1024 parameters for HS$\times$FF.

Table 1 shows our ablation experiments against the baseline [34] on the WMT-SLT development set. We repeat each experiment 3 times and report the mean and standard deviation of scores. All our experiments outperform the baseline.

We do not observe any significant difference between the BSL or DGS-pretrained I3D model on the WMT-SLT. Our best score, obtained using BSL-I3D features, was 1.51 in terms of BLEU-4. On the other hand, class probabilities obtain lower BLEU scores than feature embeddings. Therefore we use BSL-I3D features going forward for our video encoder.

	Size	BLEU-1	BLEU-4	CHRF
Baseline [34]		-	0.58	-
BSL-P	2281	14.26 ± 0.47	1.01 ± 0.2	17.0 ± 0.17
DGS-P	2301	14.6 ± 0.55	1.03 ± 0.08	17.03 ± 0.47
BSL-F	1024	15.86 ± 0.2	1.23 ± 0.25	17.27 ± 0.15
DGS-F	1024	15.14 ± 0.44	1.17± 0.08	17.13 ± 0.12

Table 1. Evaluation of different features for SLT on WMT-SLT development set. BSL-F: BSL-I3D features, DGS-F: DGS-I3D features, BSL-P: BSL-I3D class probabilities, DGS-P: DGS-I3D class probabilities.

	BLEU-1	BLEU-4	CHRF	ROUGE
MSMUNICH [15]	-	0.56	17.4	-
SLT-UPC [50]	-	0.5	12.3	-
SLATTIC [45]	-	0.25	19.2	-
Baseline [33]	-	0.12	5.5	-
DFKI-MLT [19]	-	0.11	6.8	-
NJUPT-MTT	-	0.10	14.6	-
DFKI-SLT [24]	-	0.08	18.2	-
Ours				
- Video-to-Text	14.43	0.81	18.18	5.60
- Context-to-Text	12.80	0.69	14.48	3.73
- Context+Video-to-Text	14.33	1.00	18.12	6.00
- Spot-to-Text	22.11	1.87	22.23	11.17
- Context+Video+Spot-to-Text	31.36	3.93	24.69	17.65

Table 2. Comparison with the literature on the full WMT-SLT test set.

Table 2 shows the comparison of our approach against the participants of the WMT-SLT shared task [33]. All approaches are based on Transformer architectures [52]. Similar to our *Video-to-Text*, MSMUNICH [15] also uses an I3D model for feature extraction and obtained the highest score of 0.56 in BLEU-4. While they use an I3D model pretrained on BSL-1K [2], we pretrained our I3D on the BOBSL [3] which provides better feature representation and a slight improvement.

Context-to-Text: Here, we are testing how well a network can guess the content of a sentence given the context of the preceding sentence. To do this, we need ordered data. Although the development and test data of the SRF partition of WMT-SLT consists of segments extracted from several episodes, the segments contain consecutive numbers for each episode. Therefore, we used sorted segments to evaluate our *Context-to-Text* model. As can be seen in Table 2, *Context-to-Text*, which takes only the previous sentence as a source, performs worse than our *Video-to-Text*. However, its BLEU-4 performance is still superior, and CHRF performance is competitive to the literature, which verifies that contextual information provides important cues for translation tasks.

Context+Video-to-Text: Next, we combine context and video sources by including a context encoder, a video en-

	Val				Test			
	BLEU-1	**BLEU-4**	**ROUGE**	**CHRF**	**BLEU-1**	**BLEU-4**	**ROUGE**	**CHRF**
Context-to-Text								
- 1 preceding sentence	13.50	0.45	5.59	10.0	13.41	0.45	6.10	10.4
- 2 preceding sentence	13.21	0.52	5.11	10.2	13.34	0.42	5.54	10.4
- 3 preceding sentence	13.32	0.51	5.36	10.2	13.0	0.43	5.59	10.5
- Max 10 spottings	13.77	0.74	6.33	10.88	12.90	0.60	6.01	10.76
- Max 20 spottings	13.88	0.75	6.36	10.9	12.96	0.56	6.07	10.66
Spot-to-Text	21.97	2.25	8.52	19.4	21.63	2.21	9.45	19.7
Context+Spot-to-Text	22.77	2.56	9.98	19.9	21.68	2.43	10.0	19.72

Table 3. Performance of our text-to-text models on the BOBSL dataset.

	Val				Test			
	BLEU-1	**BLEU-4**	**ROUGE**	**CHRF**	**BLEU-1**	**BLEU-4**	**ROUGE**	**CHRF**
Albenie et. al [3]	-	-	-	-	12.78	1.00	10.16	-
Video-to-Text								
- trained with 274K	15.15	1.02	12.71	19.7	12.68	0.83	8.32	17.9
- trained with 1M	18.8	1.28	7.91	17.7	17.71	1.27	8.9	18.8
Context+Video-to-Text								
- 1 preceding sentence	20.18	1.53	9.13	18.2	19.11	1.51	9.94	19.3
- 2 preceding sentence	19.14	1.52	8.97	18.0	18.15	1.41	9.56	18.9
- 3 preceding sentence	20.05	1.56	9.08	18.1	18.82	1.48	9.64	19.1
- Max 10 spottings	20.84	1.71	10.03	18.21	19.05	1.50	9.95	18.94
Context+Video+Spot-to-Text								
-with 1 preceding sentence	25.06	2.73	11.12	22.6	24.07	2.81	12.07	23.7
-with max 10 spottings	25.94	3.07	12.27	23.69	24.29	2.88	12.41	24.53

Table 4. Impact of the integrating context and spottings information to video-to-text approaches on the BOBSL dataset.

coder, and a decoder, which we call *Context+Video-to-Text*. Incorporating context information besides video features improved our translation results as we expected.

Spot-to-Text: In the literature, ground truth sign glosses are used to train a text-to-text translation model to create an upper bound for end-to-end translation [5]. Motivated by this, we created spottings as described in 4.2 using our BSL-I3D model. The trained *Spot-to-Text* model achieves significantly better translation performance compared to other single-modality architectures.

Context+Video+Spot-to-Text: Finally, we integrate automatically created spottings as input to the spotting encoder. The performance gain is significant when compared to *Context+Video-to-Text* and *Spot-to-Text*, showing the benefits of the incorporation of complementary information cues. However, this result should be taken as an upper bound on performance as the spotting approach requires a prior over the spoken word occurrence. This artificially inflates the performance but as can be seen, the potential benefits of accurate spotting on translation are significant.

5.2. Experiments on BOBSL

Context-to-Text: We evaluate training *Context-to-Text* with two different types of data on the BOBSL; a) preceding sentences and b) preceding spottings. As can be seen in Table 3, using only the preceding text data leads to poor translation. Firstly, we experiment with increasing the context by providing more preceding text. While using more sentences provides a slight improvement in terms of the BLEU-4 and CHRF scores on the validation set, it did not help in the other scores or on the test set. In the experiments with preceding spottings, we experiment with different numbers of spottings. We use the spottings from up to 3 previous sentences since we do not see a significant improvement when we include more prior sentences in the previous experiments. We obtain better results when we use the spottings of 3 prior sentences, but limit the maximum number of spottings to just 10.

Spot-to-Text: We utilized the sign spottings [32] of the BOBSL to evaluate our *Spot-to-Text* model. We train our model using all automatic annotations without any thresholding, which obtains 21.63 and 2.21 for BLEU-1 and BLEU-4 on the test set, respectively.

Video-to-Text: In [3], the authors provide an SLT baseline that is trained on a subset of the BOBSL training set. They created their new training set for sign language translation by filtering the sentences that contain high-confidence automatic spottings. They selected words that occur at least 50 times in the training set and constructed sentences by filtering according to this vocabulary. They also discard sentences with over 30 words, yielding 274K sentences. To provide a comparison, we first train our *Video-to-Text* network on this subset. However, transformers tend to get better results with more data. Therefore, we

also train our model using all sentences as in the *"version v1_2"* of the BOBSL dataset for which the training set contains about 1M sentences. In this experiment, our BLEU-4 score increased to 1.27 from 0.83 as seen in Table 4.

Context+Spot-to-Text: Firstly, we combine context and spotting sources by having a context encoder, a spotting encoder, and a decoder, which we call *Context+Spot-to-Text*. We set the maximum number of spottings to 10. *Context+Spot-to-Text* achieved better results than *Spot-to-Text* (2.21 vs 2.43 BLEU-4 score in the test set).

Context+Video-to-Text: Then, we evaluate the *Context+Video-to-Text* model. We use all training videos and all validation videos in our multi-modal experiments. As can be seen in Table 4, when using prior spoken text as input for context-encoder, our *Context+Video-to-Text* model achieves a significant improvement over our *Video-to-Text* model on both the manually aligned test set (1.27 vs. 1.51 BLEU-4, and 12.68 vs. 19.11 BLEU-1) and validation set (1.28 vs. 1.53 BLEU-4, and 18.8 vs. 20.18 BLEU-1). We also investigate using a different number of preceding sentences. Similar to *Context-to-Text* experiments, increasing the number of preceding sentences does not improve the translation quality. On the other hand, we experiment with the preceding spottings for the context-encoder. Although we obtain a much better result in the validation set (1.53 vs. 1.71 BLEU-4), we get similar results in the test set. This shows that using either the preceding sentences or preceding spottings provides similar context and helps to provide better translation.

Context+Video+Spot-to-Text : Finally, we train our transformer using all modalities. Our final approach is able to surpass all previous models and obtains state-of-the-art on the BOBSL dataset test set, with 2.81 for BLEU-4 and 24.07 for BLEU-1.

Qualitative results: In this section, we share translations produced by the proposed model using different modalities and discuss our qualitative findings. As shown in Table 5, we compare our *Video-to-Text*, *Context+Video-to-Text* and *Context+Video+Spot-to-Text* to better analyze the contribution of using the preceding context and current spottings. The results show that although translation quality is not perfect, context information helps us to get closer to the true meaning when compared to *Video-to-Text*. As shown in the first example, the ground truth translation is *"He lost nearly 200 sheep during the prolonged heavy snow in April."*. While *Video-to-Text* model is able to infer only *"sheep"* correctly, *Context+Video-to-Text* model produces *"Two sheep have been killed by the weather."*, which is a closer meaning.

Ex#1 GT:	He lost nearly 200 sheep during the prolonged heavy snow in April.
V2T:	The sheep are rounded up and the autumn begins to drift away.
(C+V)2T:	The two sheep have been killed, and the two have been killed by the weather.
(C+V+S)2T	And the sheep are in the middle of April, and they're all farmed in the winter.
Ex#2 GT:	You can see it's quite a different shape...
V2T:	It's a very different story.
(C+V)2T:	It's a very different shape.
(C+V+S)2T	It's a different shape.
Ex#3 GT:	With the crops on the farm, summer is a busy time of year with harvest just around the corner.
V2T:	It's a real dramatic change in the night and it's a real labour of love.
C+V2T:	It's a very busy time of year, but it's a very busy time of year.
(C+V+S)2T:	During the summer, the farm is busy grazing and the farm is busy harvesting.

Table 5. Qualitative results of the proposed method on the BOBSL. V2T: *Video-to-Text*, (C+V)2T: *Context+Video-to-Text*, (C+V+S)2T : *Context+Video+Spot-to-Text*.

6. Conclusion

In this paper, we have proposed a novel multi-modal transformer architecture for context-aware sign language translation. Our approach utilizes complementary transformer encoders, including a spotting and video encoder for modeling the current sign phrase and a context encoder for capturing the context of preceding sign sequences. These encoders are then combined in a final transformer decoder to generate spoken language translations. We evaluate our approach on two sign language datasets with large domains of discourse and obtain state-of-the-art results by doubling the BLEU-4 score. We hope this work will encourage the exploration of new model ideas on large-scale sign language translation. A future direction may include exploring the leverage of context, such as to alleviate the local ambiguity for similar signs, or to improve spottings performance.

Acknowledgement

This work was supported by the EPSRC project ExTOL (EP/R03298X/1), SNSF project 'SMILE II' (CRSII5 193686), European Union's Horizon2020 programme ('EASIER' grant agreement 101016982) and the Innosuisse IICT Flagship (PFFS-21-47). This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains.

References

[1] Compound-split library. https://pypi.org/project/compound-split/, 2022. Accessed: 2022-12-01.

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *European conference on computer vision*, pages 35–53. Springer, 2020.

[3] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.

[4] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58, 2014.

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

[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319. Springer, 2020.

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

[8] Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard 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.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

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

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

[12] Helen Cooper, Brian Holt, and Richard Bowden. Sign language recognition. In *Visual analysis of humans*, pages 539–562. Springer, 2011.

[13] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Isolated sign recognition from rgb video using pose flow and self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3441–3450, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, and Oscar Koller. Clean text and full-body transformer: Microsoft's submission to the wmt22 shared task on sign language translation. *arXiv preprint arXiv:2210.13326*, 2022.

[16] Karen Emmorey. *Language, cognition, and the brain: Insights from sign language research*. Psychology Press, 2001.

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

[18] Ivan Gruber, Zdenek Krnoul, Marek Hrúz, Jakub Kanis, and Matyas 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.

[19] Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. Dfki-mlt at wmt-slt22: Spatio-temporal sign language representation and translation. In *Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics*, 2022.

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

[21] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11087–11096, 2021.

[22] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping 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.

[23] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[24] Lorenz Hufe and Eleftherios Avramidis. Experimental machine translation of the swiss german sign language via 3d augmentation of body keypoints. In *Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics*, 2022.

[25] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun 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.

[26] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[29] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, and Olg Jeziorski. Meine dgs–annotiert. Öffentliches korpus der deutschen geb ärdensprache, 2. release/my dgs–annotated. public corpus. 2019.

[30] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6205–6214, 2020.

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

[32] Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. Automatic dense annotation of large-vocabulary sign language videos. In *European Conference on Computer Vision*, pages 671–690. Springer, 2022.

[33] Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, HfH Zurich, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, et al. Findings of the first wmt shared task on sign language translation. pages 744–772, 2022.

[34] Mathias Müller, Annette Rios, and Amit Moryossef. Sockeye baseline models for sign language translation. `https://github.com/bricksdont/sign-sockeye-baselines`, 2022.

[35] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[36] Ilias Papastratis, Kosmas Dimitropoulos, and Petros Daras. Continuous sign language recognition through a context-aware generative adversarial network. *Sensors*, 21(7):2437, 2021.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[39] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In *European conference on computer vision*, pages 572–578. Springer, 2014.

[40] Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, 126(2):430–439, 2018.

[41] Maja Popović. Chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

[42] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, Oct. 2018. Association for Computational Linguistics.

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

[44] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

[45] Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. Ttic's wmt-slt 22 sign language translation system. In *Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics*, 2022.

[46] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[47] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Using motion history images with 3d convolutional networks in isolated sign language recognition. *IEEE Access*, 10:18608–18618, 2022.

[48] Thad Starner, Joshua Weaver, and Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1371–1375, 1998.

[49] Rachel Sutton-Spence and Bencie Woll. *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999.

[50] Laia Tarrés, Gerard I Gállego, Xavier Giró-i Nieto, and Jordi Torres. Tackling low-resourced sign language translation: Upc at wmt-slt 22. *arXiv preprint arXiv:2212.01140*, 2022.

[51] Anil Osman Tur and Hacer Yalim Keles. Evaluation of hidden markov models using deep cnn features in isolated sign recognition. *Multimedia Tools and Applications*, 80(13):19137–19155, 2021.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[54] Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119, 2022.

[55] Kayo Yin and Jesse Read. Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, 2020.

[56] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 279–286, 2011.

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

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

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