SEDA: Simple and Effective Data Augmentation for Sign Language Understanding

Sihan Tan[1,2], **Taro Miyazaki**[2], **Katsutoshi Itoyama**[1,3], **Kazuhiro Nakadai**[1]
[1]Tokyo Institute of Technology,
[2]NHK Science and Technology Research Laboratories,
[3]Honda Research Institute Japan Co., Ltd.
{tansihan, itoyama, nakadai}@ra.sc.e.titech.ac.jp, miyazaki.t-jw@nhk.or.jp

Abstract

Sign language understanding (SLU) aims to convert sign language videos into glosses that transcribe sign language word-by-word by means of another written language and generate corresponding spoken sentences, including sign language recognition (SLR) and sign language translation (SLT). SLU has been a challenging undertaking since it demands the capability of fine-grained video understanding and sequence generation. In addition, the lack of supervised training data further hinders the advancement of SLU. To narrow the modality gap between vision and language and mitigate the data scarcity problem, we propose a **Simple and Effective Data Augmentation (SEDA)** framework for end-to-end SLU. In particular, SEDA consists of two key components: data augmentations on both sign and text sides and multi-task learning with task-specific fine-tuning. Experimental results on RWTH-PHOENIX Weather 2014T demonstrate that our proposed SEDA framework significantly and consistently outperforms the baseline model and achieves a WER of 19.91, a BLEU score of 25.19, and a ROUGE score of 51.72, delivering competitive scores in both SLR and SLT.

Keywords: Sign language understanding, Data augmentation, Multi-task learning.

1. Introduction

As the native language used by deaf and hard-of-hearing individuals to communicate, sign languages (SLs) exhibit distinctive grammar and have been established as a form of natural language (Klima and Bellugi, 1979). Sign language understanding (SLU) in which SLs are understood by means of machines mainly involves two functions: sign language recognition (SLR) and sign language translation (SLT). It is a challenging undertaking that requires the model to have the capability of fine-grained video understanding and sequence generation. Unlike spoken languages, SLs involve manual and non-manual elements (e.g., the movement of the body, head, mouth, or even eyebrows). Also, the visual signal in SLs displays dramatic variability among signers, posing a huge modality gap when transforming SLs into text (Zhang et al., 2023). Insufficient supervised training data presents an additional challenge to the advancement of SLU, as it increases the risk of overfitting. To tackle these challenges, it is essential to devise inductive biases, such as novel model architectures, training strategies and objectives, facilitating knowledge transfer, and the induction of universal representations for SLU. In this paper, we aim to augment SLs data on both sign and text sides, and provide effective training, including multi-task learning.

Existing SLU methods follow the framework of neural machine translation (NMT) where the source language is spatial-temporal pixels rather than discrete tokens and the target language is spoken languages. Depending on the model architectures, annotation pairs, or final goals, SLU comprises: Sign2Gloss (Min et al., 2021; Hao et al., 2021), Sign2Gloss2Text (Yin and Read, 2020), Sign2(Gloss+Text) (Camgoz et al., 2020) and Sign2Text (Camgoz et al., 2018; Chen et al., 2022) tasks. Additionally, to boost the well-being of the sign language community and improve SLU performance, a number of studies have focused on Gloss2Text (Moryossef et al., 2021) and Text2Gloss (Miyazaki et al., 2020; Zhu et al., 2023) by transfer learning, data augmentation, etc. Following this line of study, we find that researchers seldom explore data augmentation techniques for the sign aspect, primarily concentrating on the textual component. Furthermore, constructing large-scale SL datasets with well-aligned annotations is a time-consuming and resource-consuming task (Uthus et al., 2023). For these reasons, developing a data augmentation technique for the sign side has become crucial.

In this paper, we propose a Simple and Effective Data Augmentation (SEDA) approach for SLU. The main idea is to increase the training samples and improve the model's performance by learning highly related tasks. Specificly, we adopt different sign embeddings to augment sign representations and combine preprocessed spoken texts to achieve text augmentation. The contributions of this paper can be summarized as follows:

- We propose a Simple and Effective Data Augmentation (SEDA) approach to ease the data scarcity problem in the SLU task.

- Intensive analysis indicates that the SEDA method improves end-to-end SLU significantly through multi-task learning and task-specific fine-tuning.

- We evaluate the effectiveness of the proposed SEDA on the widely used dataset RWTH-PHOENIX Weather 2014T (PHOENIX14T). According to the experimental results, SEDA leads to notable enhancements in the SLU models, achieving competitive results in both SLR and SLT.

2. Related Work

2.1. Sign language understanding

In previous SLU research, the SLU methods can be divided into two categories: *cascading* and *end-to-end*. The cascading style adopts intermediate supervision, such as gloss, in which each gloss presents the manual transcription of a sign to convey its intended meaning. First, the sign language recognition model recognizes the continuous glosses from the unsegmented sign videos **(Sign2Gloss)**, and then, the predicted glosses are utilized to generate the corresponding spoken sentence **(Gloss2Text)**. In the Sign2Gloss system, a common architecture involves a feature extractor and a temporal modeling mechanism, such as Connectionist Temporal Classification (CTC) (Graves et al., 2006). However, most *cascading* SLU methods inevitably introduce an information bottleneck, as these methods utilize sign glosses as intermediate supervision. When the original sign video is transformed into glosses, some spatial-temporal information is lost (Yin and Read, 2020). End-to-end training is one promising way to achieve SLU. The *end-to-end* SLU directly converts the sign videos to spoken sentences **(Sign2Text)**. Camgoz et al. (2018) formalizes this field by taking the SLU task as a neural NMT problem, demonstrating the practicality of the end-to-end method. In the following work (Camgoz et al., 2020), the sign glosses serve as the auxiliary supervision to regularize the neural encoder **(Sign2(Gloss+Text))**. Following this paradigm, we focus on the challenge of data scarcity by proposing a simple and effective data augmentation method. Besides, the data augmentation of sign language representation has rarely been explored before.

2.2. Multi-task learning

Multi-task learning aligns with the goal of increasing training samples and improving the model by learning related tasks (Zhang and Yang, 2018). Recently, the natural language processing (NLP) domain has benefited from adopting multi-task learning (e.g., multilingual translation). In Text2Gloss (Zhu et al., 2023), multilingual translation has been adopted to improve translation performance. As for SLU research, comprehensive multi-task learning experiments (i.e Sign2Gloss, Sign2Text, Gloss2Text, Text2Gloss, and MT) are conducted in (Zhang et al., 2023). These experiments offer valuable insights into how multi-task learning benefits SLT. We then combine multi-task learning with data augmentation and innovatively propose a simple and effective data augmentation (SEDA) framework for SLU. The following sections present the details of the proposed methods.

3. Methods

We applied the proposed SEDA to the sign language transformer (Camgoz et al., 2020) that widely serves as the baseline model in Sign2(Gloss+Text) and Sign2Text tasks. In this section, we first present the task definition and revisit the sign language transformer. We then give a comprehensive explanation of our proposed approaches, including data augmentation on both sign and text sides and multi-task learning.

3.1. Task definition

We formally define the setting of end-to-end Sign2(Gloss+Text). Given sign-gloss-text triples $\mathcal{D} = (S_i, G_i, T_i)_{i=1}^{N}$, where i and N represent the index of the input triple and the number of triples in the dataset, $S_i = \{s_{i,z}\}_{z=1}^{|S_i|}$ denotes sign videos comprising $|S_i|$ frames, $G_i = \{g_{i,u}\}_{u=1}^{|G_i|}$ represents a gloss sequence with $|G_i|$ gloss annotations, and $T_i = \{t_{i,w}\}_{w=1}^{|T_i|}$ is the corresponding spoken text consisting of $|T_i|$ words, and generally in SL data triples, $|S_i| \gg |G_i|$ and $|S_i| \gg |T_i|$. The end-to-end Sign2(Gloss+Text) aims to predict glosses G, the text in sign order, and generate spoken text T.

3.2. Sign language transformers

The sign language transformer follows the encoder-decoder paradigm, with Transformer (Vaswani et al., 2017) as its backbone. It consists of five modules: a sign embedding, a transformer encoder, a CTC classifier, a word embedding, and a transformer decoder. In our sign language transformer, we introduce label smoothing to CTC training loss, aiming to mitigate the overfitting issue, and a new sign embedding to extract informative sign features.

Sign embedding. Replacing the sign embedding in (Camgoz et al., 2020), we adopt the retrained one from self-mutual knowledge distillation (SMKD) model (Hao et al., 2021) followed by 1D-CNN layers to extract the informative sign representations. Here, we denote the new sign embedding as spatial-temporal embedding. During training, the parameters of the new pre-trained spatial-temporal embedding are frozen. We formulate this operation as:

$$f_i = \text{Spatial-temporalEmbedding}(S_i), \tag{1}$$

where f_i is the sign representation from the newly introduced spatial-temporal embedding.

Transformer encoder. The sign language transformer encoder intending to predict sign glosses G contains multi-layer transformer networks. The inputs of the transformer encoder are embedding sequences of tokens, such as the sign feature f_i from the spatial-temporal sign embedding. Unlike traditional seq2seq models, transformer networks do not employ recurrence or convolution mechanisms, which means they do not inherently contain positional information within sequences. To tackle this concern, we adopt the positional encoding introduced in (Vaswani et al., 2017) and add temporal order information into the embedded representations in the following manner:

$$\hat{f}_i = f_i + \text{PositionalEncoding}, \tag{2}$$

where the PositionalEncoding is pre-defined to generate a distinct vector in the shape of a sine wave that has been phase-shifted for each time step. Furthermore, $\hat{f}_i$ is modeled using self-attention and projected into contextual representations $h(\mathcal{S}_i)$ that are fed forward to the transformer decoder to generate the target spoken text.

CTC with label smoothing. Sign language transformer employs glosses as auxiliary supervisions to train the transformer encoder. In the CTC-based Sign2Gloss tasks, CTC introduces the $blank$ label, representing the silence or transition between two consecutive glosses. The extended glosses can be defined as $\mathcal{G}^* = (g_{i,1}, ..., g_{i,|G_i|}) \cup \{blank\} \in R^l$, where l is the total number of labels. The CTC is utilized to compute the $p(\mathcal{G}^*|h(\mathcal{S}_i))$, marginalizing over all possible $h(\mathcal{S}_i)$ to $\mathcal{G}^*$ alignments as:

$$p(\mathcal{G}^*|h(\mathcal{S}_i)) = \sum_{\pi \in \mathcal{B}} p(\pi|h(\mathcal{S}_i)), \tag{3}$$

where π is a path and $\mathcal{B}$ is the collection of all possible paths that lead to $\mathcal{G}^*$. The CTC loss in Sign2Gloss is defined as:

$$\mathcal{L}_{ctc} = 1 - p(\mathcal{G}^*|h(\mathcal{S}_i)). \tag{4}$$

While CTC-based methods offer notable training convenience, as indicated in previous works (Min et al., 2021; Tan et al., 2023), they are prone to overfitting during training. Moreover, SLs are low-resource languages, this fact also poses the risk of overfitting. To mitigate the overfitting problem, we add a regularization term to the CTC objective function, which consists of the Kullback-Leibler (KL) divergence between the network's predicted distribution P and a uniform distribution $\mathcal{Q}$ over labels.

$$\mathcal{L}_{\mathcal{R}} = (1-\alpha)\mathcal{L}_{ctc} + \alpha \sum_{t=1}^{T} D_{KL}(P||\mathcal{Q}) \tag{5}$$

Training the transformer encoder networks using CTC with label smoothing encourages the differences between the logits of the correct class and the logits of the incorrect classes to be a constant dependent on the weight ratio α.

Transformer decoder. The sign language transformer decoder aims to generate the spoken sentence based on the contextual representation $h(\mathcal{S}_i)$. It consists of cross-attention and self-attention layers. We introduce cross-entropy loss as the objective function of spoken language sentence generation.

$$\mathcal{L}_{\mathcal{T}} = -\sum_{u=1}^{|T_i|} log\mathcal{P}(t_{i,u}|t_{i,<u}, h(\mathcal{S}_i)) \tag{6}$$

3.3. Data augmentation

Data augmentation is a common technique used to relieve the risk of overfitting due to data scarcity. One commonly used data augmentation method involves original data with some minor changes. We apply the SEDA framework to the sign language transformer.

Sign representation augmentation. Instead of augmenting the sign frames directly, our proposed SEDA focuses on sign feature augmentation, that is, the same sign frames are processed by different sign embeddings. Given the sign video frames $S_i = \{s_{i,z}\}_{z=1}^{|S_i|}$, we propagate the S_i to different embedding layers (i.e., spatial embedding from (Camgoz et al., 2020) and newly introduced spatial-temporal sign embedding) separately to obtain multiple sign features. By taking this approach, we can obtain $f_i \in \mathcal{F}$ from spatial-temporal embedding, where $\mathcal{F} = \{f_1, f_2, ..., f_N\}$, and $f'_i \in \mathcal{F}'$ from the original spatial embedding in the sign language transformer, where $\mathcal{F}' = \{f'_1, f'_2, ..., f'_N\}$.

Spoken text augmentation. Inspired by the combining preprocessing methods in (Zhu et al., 2023),

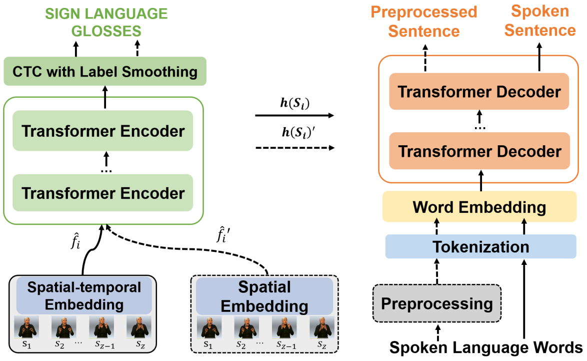

Figure 1: Overview of the proposed SEDA framework. The same sign frames will be forwarded to different sign embeddings to obtain multiple sign features. During multi-task learning, the sign features from spatial-temporal embedding are used to predict glosses and original spoken sentences, as shown by the solid line. Meanwhile the sign features from spatial embedding are fed to the model to generate glosses and preprocessed sentences, which is presented by the dotted line.

we apply preprocessing techniques to the spoken sentence T_i. We conduct lemmatization and alphabet normalization on the PHOENIX14T dataset (Camgoz et al., 2018) and combine the processed data with the original annotations. Lemmatization is the linguistic process of reducing words to their base or root form. Alphabet normalization is employed to convert specific letters, such as ü, ö, ä, and ß, into their corresponding counterparts. The processed spoken text, denoted as T_i', is then paired with the copied gloss sequence G_i to become a new training dataset on the target side. Once the augmented sign features and preprocessed spoken text annotations are obtained, we are able to construct new data triples $\mathcal{D}_1 = (f_i, G_i, T_i)_{i=1}^N$ and $\mathcal{D}_2 = (f_i', G_i, T_i')_{i=1}^N$.

3.4. Multi-task learning

The augmented data triples, represented as $\mathcal{D}_1$ and $\mathcal{D}_2$, are then mixed up and fed to the sign language transformer one after another. As shown in Fig. 1, when presented with the input f_i, the sign language transformer encoder is trained to predict G_i, and the transformer decoder is trained to generate T_i. The same procedure applies to the input f_i', where the sign language transformer encoder predicts G_i, and the transformer decoder generates T_i'.

The networks are trained by minimizing the joint loss term $\mathcal{L}$, which is the weighted sum of the translation loss $\mathcal{L}_\mathcal{T}$ and the gloss prediction loss $\mathcal{L}_\mathcal{R}$, as follows:

$$\mathcal{L} = \lambda_R \mathcal{L}_\mathcal{R} + \lambda_T \mathcal{L}_\mathcal{T}, \tag{7}$$

where λ_R and λ_T are hyperparameters that determine recognition and translation loss function weight during training. Since our final goal is to predict G_i and generate T_i, we then fine-tune the network using $\mathcal{D}_1 = (f_i, G_i, T_i)_{i=1}^N$.

4. Experiments

4.1. Experimental setup

To evaluate the effectiveness of the proposed SEDA framework, we conducted ablation experiments.

Model setting. For training hyper-parameters, we start mainly from the setting for the sign language transformer[1]. In particular, we keep $\alpha = 0.01$, $\lambda_R = 5.0$, and $\lambda_T = 1.0$, which is empirically decided. As suggested in (Zhu et al., 2023), the model performance increases when the number of encoders or decoders is reduced compared to the original transformer architecture in SL translation scenarios. We performed extensive experiments. As the results indicated, we maintained the encoder depth at 2 and the decoder depth at 4.

Dataset. We worked on the widely utilized PHOENIX14T dataset and augmented the spoken texts (Zhu et al., 2023). The details of the augmented information are shown in Table 2. Note that we used $\mathcal{D}_1$ for the development and test.

Evaluation metrics. We report the experimental results mainly on the Sign2 (Gloss+Text) task, including the Sign2Gloss and the total Sign2 (Gloss+Text) results. The most common measure of Sign2Gloss performance is the word error rate (WER), which can be calculated as:

$$\mathrm{WER} = \frac{S + D + I}{S + D + C}, \tag{8}$$

where S, D, I, and C indicate the number of **S**ubstitutions, **D**eletions, **I**nsertions, and **C**orrections, respectively. For SLT task, we use standard metrics commonly used in machine translation, including tokenized BLEU (Papineni et al., 2002) with 4-grams and the Rouge-L F1 (ROUGE) (Lin, 2004).

4.2. Experimental results

We evaluated the proposed SEDA framework on augmented PHOENIX14T. The main results are listed in Table 1. On the PHOENIX14T development set, the proposed SEDA surpassed the baseline by 9.93 WER, 4.24 BLEU, and 4.65 ROUGE. It also outperformed the state-of-the-art end-to-end or cascading approaches.

4.3. Discussion

Introducing high-quality spatial-temporal sign embedding improves SLR and SLT. Replacing

[1] `https://github.com/neccam/slt`

Table 1: End-to-end SLU performance on PHOENIX14T dataset

	DEV			TEST		
Methods	WER↓	BLEU-4↑	ROUGE↑	WER↓	BLEU-4↑	ROUGE↑
Previous research (end-to-end)						
Joint-SLRT (Camgoz et al., 2020)	24.98	22.38	–	26.16	21.32	–
Sign Back Translation (Zhou et al., 2021)	22.70	23.90	50.29	24.45	24.34	49.54
STMC-T*(Zhou et al., 2022)	–	24.09	48.24	–	23.65	46.65
Previous research (cascading)						
STMC-Transformer*(Yin and Read, 2020)	**19.60**	22.47	48.70	**21.00**	22.47	48.78
Ours (end-to-end)						
Baseline	29.84	20.95	47.07	28.67	21.70	47.82
+High-quality Spatial-temporal Sign Embedding	21.40	22.28	48.81	22.59	22.86	48.97
+ CTC Label Smoothing	21.56	23.05	48.86	22.05	22.40	47.58
+ Multi-task Learning	20.36	23.88	50.57	21.79	23.34	49.71
+ Fine-tune	**19.91**	25.19	**51.72**	**21.51**	**24.89**	**51.61**
+ Gloss-less fine-tune	–	**25.35**	51.40	–	24.75	50.77

$*$ denotes using extra clues (keypoints)

Table 2: Statistics of preprocessed PHOENIX14T

	Original Text			**Preprocessed text**		
	Train	Dev	Test	Train	Dev	Test
Instance	7,096	519	642	7,096	519	642
Vocab.	1,066	393	411	2,216	793	836
tot. words	99,081	6,820	7,816	99,081	6,820	7,816
tot.OOVs	–	57	60	–	39	38

the original sign embedding, we introduce the retrained one from the SKMD model and adopt 1D-CNN layers to extract the spatial-temporal sign information. This replacement delivers notable enhancements in both SLR and SLT (–8.44 WER, +1.33 BLEU, +1.74 ROUGE on the dev set). Adding a regularization term to the CTC, we observe an improvement in SLT (+ 0.77 BLEU on the dev set).

Multi-task learning enhances both SLR and SLT. The sharing of parameters through multi-task learning, using augmented dataset, facilitates knowledge transfer. As shown in Table 1, the multi-task learning achieves a quality boost (–1.2 WER, + 0.83 BLEU, +1.71 ROUGE on the dev set). Sharing the mixed parameters benefits tasks but lacks of task-specific characteristics. For this, we performed fine-tuning in the following.

Mixing shared parameters with task-specific parameters further provides quality gains. We further conduct task-specific fine-tuning using the data triples $\mathcal{D}_1 = (f_i, G_i, T_i)_{i=1}^{N}$. Here gloss-less fine-tuning refers to using the multi-task learning applied model, and we fine-tune the model to do the Sign2Text task without glosses. By task-specific fine-tuning, SLR and SLT tasks undergo a dramatic improvement (Fine-tune: –0.45 WER, + 1.31 BLEU, + 1.15 ROUGE; Gloss-less fine-tune: +1.47 BLEU, +0.83 ROUGE on the dev set).

5. Conclusion

In this paper, we propose a simple and effective data augmentation (SEDA) method to mitigate the data scarcity problems in end-to-end sign language understanding (SLU). The SEDA approach includes adopting different sign embeddings, combining preprocessed spoken texts, and a multi-task learning strategy. The former two methods increase the amount of training data, especially the sign representations, which has rarely been conducted before. Multi-task learning narrows the gap between vision and language by sharing mixed parameters. Experimental results on the widely utilized PHOENIX14T dataset indicate that our proposed SEDA benefits the end-to-end SLU, surpassing the baseline by 9.93 WER, 4.24 BLEU score, and 4.65 ROUGE score and achieving competitive results in both sign language recognition (SLR) and translation (SLT) tasks.

6. Limitations

While our SEDA framework significantly benefits the end-to-end SLU on the PHOENIX14T dataset, it still faces the limitation that more datasets, such as the German sign language dataset (Public DGS Corpus (Hanke et al., 2020)) or Chinese sign language dataset (CSL-Daily (Zhou et al., 2021)), are needed to demonstrate the universality of the proposed method. We will adopt multiple datasets and conduct more detailed analyses in future work.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant No. JP19KK0260, JP20H00475 and JP23K11160.

8. Bibliographical References

Necati Cihan Camgoz et al. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Necati Cihan Camgoz et al. 2020. 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 (CVPR)*.

Yutong Chen et al. 2022. 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.

Alex Graves et al. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Thomas Hanke et al. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Aiming Hao et al. 2021. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11303–11312.

Edward S Klima and Ursula Bellugi. 1979. *The signs of language*. Harvard University Press.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yuecong Min et al. 2021. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11542–11551.

Taro Miyazaki et al. 2020. Machine translation from spoken language to sign language using pre-trained language model as encoder. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 139–144, Marseille, France. European Language Resources Association (ELRA).

Amit Moryossef et al. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.

Kishore Papineni et al. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Sihan Tan et al. 2023. Improving sign language understanding introducing label smoothing. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 113–118. IEEE.

David Uthus et al. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *arXiv preprint arXiv:2306.15162*.

Ashish Vaswani et al. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. *arXiv preprint arXiv:2004.00588*.

Biao Zhang et al. 2023. Sltunet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*.

Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43.

Hao Zhou et al. 2021. 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.

Hao Zhou et al. 2022. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24:768–779.

Dele Zhu et al. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.