MLSLT: Towards Multilingual Sign Language Translation

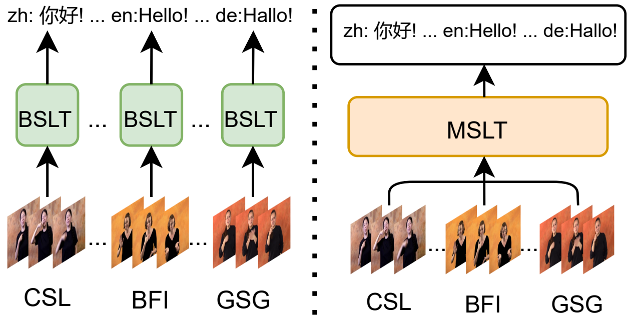
Aoxiong Yin[1] , Zhou Zhao[1]*, Weike Jin[1] , Meng Zhang[2] , Xingshan Zeng[2] , Xiaofei He[1]
[1]Zhejiang University ,[2]Huawei Noah's Ark Lab
{yinaoxiong,zhaozhou,weikejin}@zju.edu.cn
zhangmeng92@huawei.com, zxshamson@gmail.com, xiaofei_h@qq.com

Abstract

Most of the research to date focuses on bilingual sign language translation (BSLT). However, such models are inefficient in building multilingual sign language translation systems. To solve this problem, we introduce the multilingual sign language translation (MSLT) task. It aims to use a single model to complete the translation between multiple sign languages and spoken languages. Then, we propose MLSLT, the first MSLT model, which contains two novel dynamic routing mechanisms for controlling the degree of parameter sharing between different languages. Intra-layer language-specific routing controls the proportion of data flowing through shared parameters and language-specific parameters from the token level through a soft gate within the layer, and inter-layer language-specific routing controls and learns the data flow path of different languages at the language level through a soft gate between layers. In order to evaluate the performance of MLSLT, we collect the first publicly available multilingual sign language understanding dataset, Spreadthesign-Ten (SP-10), which contains up to 100 language pairs, e.g., CSL→en, GSG→zh. Experimental results show that the average performance of MLSLT outperforms the baseline MSLT model and the combination of multiple BSLT models in many cases. In addition, we also explore zero-shot translation in sign language and find that our model can achieve comparable performance to the supervised BSLT model on some language pairs. Dataset and more details are at https://mlslt.github.io/.

1. Introduction

Sign languages are the primary means of communication for an estimated 466 million deaf or hard-of-hearing people worldwide [38]. However, the difference between sign language and spoken language causes some communication barriers between them and hearing-unimpaired people, which brings inconvenience to their daily lives. This motivates researchers to design more efficient and accurate sign language translation systems [23, 24, 52, 56].


zh: 你好! ... en:Hello! ... de:Hallo!
BSLT
...
BSLT
...
BSLT
CSL
BFI
GSG
zh: 你好! ... en:Hello! ... de:Hallo!
MSLT
CSL
BFI
GSG

(a) MSLT system based on multi-models (b) MSLT system based on single model

Figure 1. An example to illustrate the advantages of the MSLT model over the BSLT model in constructing a multilingual sign language translation system.

Due to the lack of data, almost all previous studies focus on building a bilingual sign language translation (BSLT) model [2,3,27,37,53,57] (such as American Sign Language to English or British Sign Language to English). However, there are more than 300 sign languages in the world [34], and there are 389 languages spoken by more than one million people worldwide [9, 40]. So if each model can only handle one language pair, then in order to handle all situations, we need to create 116,700 models. Even if all sign languages are translated into English and then translated into other languages with the help of a machine translation system, we still need to create 300 models, which is unacceptable for practical applications.

Therefore, we aim in this work to design a single model to realize the translation between multiple sign languages and spoken languages. Figure 1 shows the advantages of the multilingual sign language translation (MSLT) model compared to the BSLT model in constructing an MSLT system. However, there are many challenges in building such a MSLT model. First, we lack an MSLT corpus. Almost all public sign language translation datasets contain only one sign language, and they are mainly concentrated

[*] Corresponding author.

in a few sign language categories such as CLS, GSG, and ASE [2, 8, 57]. Secondly, while joint training brings beneficial knowledge transfer, it also introduces language divergence and representation bottlenecks as the number of data and languages increases [6].

In order to solve the above challenges and promote progress in the field of MSLT, we construct the first large-scale parallel multilingual sign language understanding dataset, Spreadthesign-Ten (SP-10). SP-10 contains videos and corresponding spoken translations in ten sign languages collected from spreadthesign [15]. There is also a corresponding relationship between videos of different sign languages, which means that this dataset is also suitable for the multilingual text-to-video sign language generation task and the multilingual video-to-video sign language translation task.

Then, we propose the MLSLT model, an end-to-end MSLT model based on the Transformer [49] architecture. To alleviate language conflicts and representation bottlenecks, we meticulously design two end-to-end data-driven routing mechanisms, inter-layer language-specific routing (InterLSR) and intra-layer language-specific routing (IntraLSR), to help the model control parameter sharing between different languages. As shown in Figure 2, InterLSR is used to control the degree of sharing of different sign languages in the Transformer layer. Without InterLSR, the degree of sharing of all sign languages in any layer is 100%. Any sign language on each layer has a corresponding gate to control the proportion of this sign language flowing through this layer. Optimizing the state of the gate during training can adjust the network structure to maximize translation quality. IntraLSR is used to control the proportion of particular language flowing through language shared parameters and language-specific parameters in the layer. During training, each token is simply projected and then added according to the ratio calculated by the gate. In inferencing, we first linearly combine the shared feature extraction expert and the language-specific feature extraction expert, and then perform matrix multiplication, which can save half of the calculation. In addition, we also encourage shared feature extraction experts and language-specific feature extraction experts to encode different aspects of the input by using soft orthogonal subspace constraints [1].

On the SP-10 dataset, we respectively implement the algorithms proposed by Camgöz et al. [3] and Johnson et al. [20] as the baseline for BSLT and MSLT. The experimental results on the SP-10 dataset show that the average BLEU and ROUGE scores obtained by MLSLT exceed the MSLT baseline and the BSLT baseline, although MLSLT uses fewer parameters than them. This fully proves the effectiveness of our proposed method and shows the great potential of multilingual sign language translation. We also explore zero-shot translation in sign language and find that our model can achieve comparable performance to the supervised BSLT model on some language pairs, which shows that our model builds an implicit bridge between the target language pairs.

Our main contributions are summarized as follows,

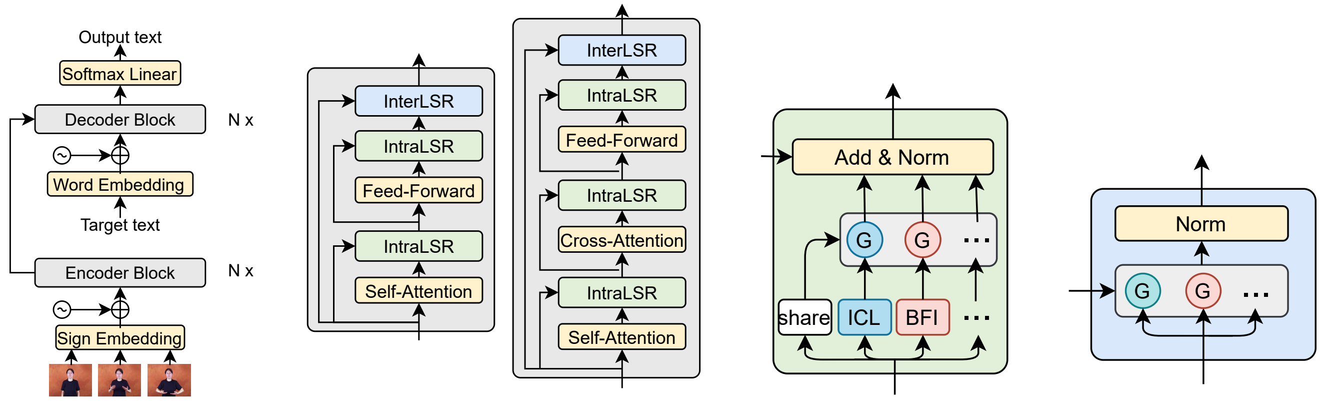
- We contribute a large-scale multilingual sign language understanding dataset suitable for multiple tasks such as multilingual sign language translation, multilingual text-to-video sign language generation, and multilingual video-to-video sign language translation.
- We are the first to explore the MSLT problem, and we propose MLSLT, an MSLT framework based on dynamic neural network. Two novel dynamic routing mechanisms are used to control parameter sharing between different sign languages.
- Extensive experimental results show that our proposed single model can perform better than the MSLT baseline and multiple BSLT models while using fewer parameters. A broad range of new baseline results can guide future research in this field.

2. Related Work

As a research field with profound social significance, the study of sign language understanding has a long history [41, 44–47]. Early research mainly used Hidden Markov Model (HMM) and other methods to model sequence information. Recently, with the continuous development of deep learning, researchers have tried to use neural network methods to handle sign language understanding tasks and achieved good results.

The early stage of sign language understanding is manifested as the sign language recognition task, and it aims to recognize sign language video as the corresponding gloss texts. The initial research direction is the isolated sign language recognition (ISLR) task [16, 31, 36], which aims to recognize isolated signs. With the development of action recognition and the proposal of some large-scale datasets [21, 26], some CNN networks [5, 10] achieve good results on ISLR. Advances in the ISRL task promote researchers to further explore the continuous sign language recognition (CSLR) task [17, 28], which attempts to recognize sign gloss sequences from continuous videos. However, due to the significant grammatical differences between sign language and spoken language, it is difficult for ordinary people to understand the sign gloss sequence.

In order to solve this problem, Camgöz et al. [2] propose a new task, sign language translation (SLT), and a new dataset, PHOENIX14T. SLT aims to translate sign language videos into corresponding spoken language texts. Then, Camgöz et al. [3] apply the Transformer network [49] to



(a) MLSLT (b) Transformer Block (c) IntraLSR (d) InterLSR

Figure 2. An overview of our proposed framework for multilingual sign language translation. (a). The MSLT Transformer. (b). The MSLT Transformer Blocks. (c). The intra-layer language-specific routing. The blue squares represent ICL-specific parameters, the blue G indicates the gate that controls the ratio of data flow through ICL parameters and shared parameters, and other languages are the same. (d). The inter-layer language-specific routing. G represents the gate that controls the ratio of data flowing through the layer and not flowing through the layer.

SLT and design a joint training method to utilize the alignment information provided by gloss. Li et al. [27] design a hierarchical structure to extract sign language features and explore SLT without gloss supervision. Orbay et al. [37] utilize adversarial, multi-task, and transfer learning to search for semi-supervised tokenization methods to reduce dependence on gloss annotations. Zhou et al. [57] propose a data enhancement method based on sign language back translation, which generates gloss texts from spoken language texts. Yin et al. [52] propose a simultaneous sign language translation method based on wait-k [32]. Gan et al. [11] use joint point data as additional supervision information during SLT to improve the model's performance.

However, the current research mainly focuses on BSLT and a few sign languages such as CLS or GSG. There are few studies on MSLT and other sign languages. Therefore, we try to explore the MSLT task. Multilingual translation has made some progress in the field of neural machine translation [6, 20, 51, 55], and how to share parameters between different languages is one of the most important issues in this field. High-quality sign language datasets are essential for sign language research. A summary of the publicly available sign language translation datasets is shown in Table 1. It can be seen that the previous datasets have very few types of sign language and only one language pair. To this end, we propose a multilingual sign language understanding dataset with ten sign languages and 100 types of language pairs.

3. Methods

3.1. Preliminaries

SLT is often considered a sequence-to-sequence learning problem. Given a source video sentence $V' = (v_1, v_2, ..., v_T)$ with T frames, SLT can be formulated as learning the conditional probability $p(Y'|V')$ of generating a spoken language sentence $Y' = (y_1, y_2, ..., y_M)$ with M words. When performing MSLT, we follow the practice of Johnson et al. [20] by adding language tags to the beginning of the sequence to indicate the language category of source and target, i.e. changing V' to $V = (lang_{src}, v_1, v_2, ..., v_T)$ and Y' to $Y = (lang_{tgt}, y_1, y_2, ..., y_M)$. To unify the standard, we abandon some traditional sign language abbreviations in this paper. All sign language categories are represented by three letters that conform to the ISO639-3 [19] standard (for example, changing ASL to ASE, changing DGS to GSG), and all spoken language categories are represented by a 2-letter code that conforms to the ISO639-1 [18] standard. An overview of our proposed framework for MSLT is shown in Figure 2. In the remainder of this section, we will introduce each component of our approach in detail.

3.2. Embedding Layer

Similar to the general sequence-to-sequence learning task, we first embed the source video and target text. For video embedding, we first use a CNN network [48] to extract video features for each frame and then use a linear layer to project it into a denser space. For text embedding, we first use MultiBPEmb [14], which is a multilingual BPE [43] sub-word segmentation model learned on the Wikipedia dataset using the SentencePiece [25] tool to segment text into sub-words. Using the BPE algorithm to split longer words into sub-words can allow generalized morphological variants or compound words. In addition, the use of sub-words can alleviate the data sparsity problem in multilingual vocabulary representation because different languages may share the same root or affix. We use the pre-trained sub-word embedding in MultiBPEmb as initialization and use a linear layer to adjust the dimensionality of the sub-word vector. We formulate these operations as:

$$f_t = CNN_\theta(v_t)W_1 + b1 \tag{1}$$

$$w_m = Emb(y_m)W_2 + b_2 \tag{2}$$

where θ denotes the parameters of the CNN network. To provide temporal position information, we use the same position encoding method as the vanilla Transformer.

3.3. Intra-layer Language-specific Routing

As shown in Figure 2b and Figure 2c, we use intra-layer language-specific routing (IntraLSR) to control the proportion of each language flowing through the shared parameters and language-specific parameters in the Transformer layer. IntraLSR replaces the position of the add & norm layer in the original Transformer, and forms a new sub-layer with structures such as self-attention layer, feed-forward layer and cross-attention layer. Since the proportion of shared knowledge in different languages should be different, each language in IntraLSR will learn a soft gate based on the output $e^l \in \mathbb{R}^d$ of the previous sub-layer. These gates give IntraLSR the ability to control the proportion of data flowing through shared paths or language-specific paths freely. We formulate these operations as:

$$h^s = f(e^l)W^s, h^u = f(e^l)W^u \tag{3}$$

$$h = g_u(e^l)h^u + (1 - g_u(e^l))h^s \tag{4}$$

$$e^{l+1} = LayerNorm(h + e^l) \tag{5}$$

where $f(\cdot)$ represents the first layer in the sub-layer such as the self-attention layer, $W^s \in \mathbb{R}^{d\times d}$ represents a weight matrix shared by all languages, and $W^u \in \mathbb{R}^{d\times d}$ represents a weight matrix unique to each language. $g_u(\cdot)$ represents the gate unique to each language. We use an MLP network to calculate the degree of opening and closing of the gate based on the output of the previous sub-layer. The detailed calculation method is as follows:

$$g_u(e^l) = \sigma((relu(e^lW_3 + b_3) + e^l)W_4 + b_4) \tag{6}$$

where $\sigma(\cdot)$ is the logistic-sigmoid function, $W_3 \in \mathbb{R}^{d\times d}$ and $W_4 \in \mathbb{R}^{d\times 1}$ are trainable parameters, and d represents the dimensionality of the hidden layer in Transformer.

In the inference stage, in order to reduce the amount of calculation, we adjust the calculation method of h as follows:

$$h = f(e^l)(g_u(e^l)W^u + (1 - g_u(e^l))W^s) \tag{7}$$

The adjusted calculation method can reduce one matrix multiplication operation while ensuring that the calculation result will not change.

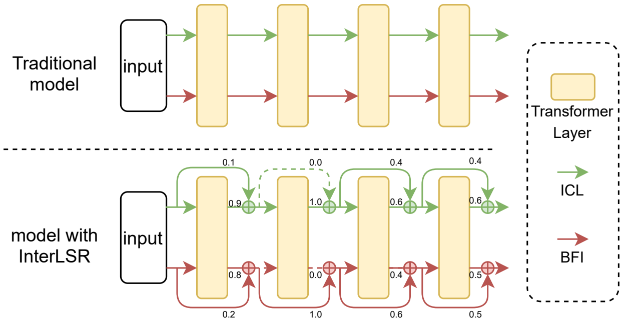


Figure 3. An example to illustrate the difference between the model with the InterLSR module and the traditional model.

3.4. Inter-layer Language-specific Routing

Inter-layer language-specific routing (InterLSR) is designed to control the degree of sharing between different languages when flowing through the Transformer layer. Figure 3 vividly shows the difference between the model with the InterLSR module and the traditional model and how the InterLSR module works. The figure shows that the degree of parameter sharing between different languages in the traditional model is 100%. However, because the differences between languages are objective, complete parameter sharing will make it difficult for the model to capture language-specific patterns. Although the IntraLSR we designed can partially alleviate this problem, it mainly controls the flow of data at the token level. In addition, we also need a way to generate unique data transmission paths for different languages from the language level, which is why we proposed InterLSR. Its calculation method is as follows:

$$\alpha = \sigma(\mathsf{E}^{lang}W_5 + b_5) \tag{8}$$

$$z^{l+1} = LN(\alpha z^l + (1 - \alpha)o^{l+1}) \tag{9}$$

where $\mathsf{E}^{lang} \in \mathbb{R}^{d_g}$ represents language embedding, $W_5 \in \mathbb{R}^{d_g\times 1}$ represents trainable parameters, d_g is the dimension of language embedding vector, $\sigma(\cdot)$ is the the logistic-sigmoid function, z^l represents the output of the previous layer, and o^{l+1} represents the original output of the current layer.

3.5. Training of MLSLT

In the MLSLT model, we use label smoothed [33] cross-entropy loss to optimize the MSLT task:

$$\hat{y_m} = y_m(1 - \varepsilon) + \varepsilon/K \tag{10}$$

$$\mathcal{L}_{ce} = -\sum_{m=1}^{M} \hat{y_m} log(P(y_m|y_{1:m-1}, V; \theta)) \tag{11}$$

where y_m represents the one-hot vector with the target category being 1, and the other categories being 0, K is the

Table 1. Summary of publicly available sign language translation datasets. To the best of our knowledge, SP-10 is the first publicly available multilingual sign language understanding dataset, which contains up to 100 types of language pairs, such as CSL→en, GSG→zh, etc. The pose in the table indicates whether the dataset provides human body posture data, and Transparent indicates whether the dataset includes videos with a transparent background. This feature allows researchers to freely replace the video background to obtain a more robust model.

Dataset	Language	Vocab	Duration (h)	Signers	Language Pairs	Transcription	Pose	Transparent
Boston104 (2006) [54]	ASE	104	0.145	3	1 (ASE→en)	✓	✗	✗
SIGNUM (2010) [50]	GSG	450	55	25	1 (GSG→de)	✓	✗	✗
NCSLGR (2012) [35]	ASE	1.8k	5.3	4	1 (ASE→en)	✓	✗	✗
BSL Corpus (2013) [42]	BFI	5k	-	249	1 (BFI→de)	✓	✗	✗
Video-Based CSL (2018) [17]	CSL	178	100	50	1 (CSL→zh)	✓	✓	✗
Public DGS Corpus (2020) [13]	GSG	-	50	327	1 (GSG→de)	✓	✓	✗
PHOENIX14T (2020) [37]	GSG	3k	11	9	1 (GSG→de)	✓	✗	✗
CSL-Daily (2021) [57]	CSL	2.3k	20.79	10	1 (CSL→zh)	✓	✓	✗
How2Sign (2021) [8]	ASE	16k	79	11	1 (ASE→en)	✓	✓	✗
SP-10 (ours)	Multilingual	16.7k	14	79	100 (Multilingual)	✓	✓	✓

vocabulary size, ε is a hyperparameter, and θ is the trainable parameter in the model.

In addition, we also use soft orthogonal subspace constraint loss [1] to encourage language-sharing weight matrix W^s and language-specific weight W^u to encode different aspects of the input:

$$\mathcal{L}_o = \frac{1}{L}\sum_{i=1}^{L} \left\| (W^s)^\top W_i^u \right\|_F^2 \tag{12}$$

Therefore, the total loss function to train the MLSLT model is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_o \tag{13}$$

where λ_1 and λ_2 are hyperparameters used to balance losses.

Table 2. Key statistics of the SP-10 dataset.

	Train	DEV	Test
samples	830	142	214
segments	8300	1420	2021
duration (h)	9.9	1.7	2.4
frames/segment	112.4	113.7	119.1

4. Multilingual Sign Language Understanding Dataset

As discussed in section 1, the lack of datasets hinders progress in multilingual sign language translation. To solve this problem, we propose the Spreadthesign-Ten(SP-10) dataset, which contains videos and corresponding spoken translations in ten sign languages collected from spreadthesign [1]. Almost all the samples in the dataset will contain ten videos of different sign languages and ten translation texts of spoken language.

Table 1 shows the information of the current publicly available sign language translation datasets and the comparison between SP-10 and them. All videos in SP-10 have a resolution of 320×240 and a frame rate of 25FPS. We distinguish different signers by using the face recognition toolbox [2] to perform face recognition on all videos. We use the OpenPose toolbox [4] to extract 2-dimensional (2D) pose estimation data from the original video data. Each pose annotation data contains 25 body joint points, 70 face joint points, and 42 hand joint points. We utilize Robust Video Matting (RVM) toolbox [30] to matte the original video data to obtain a video with transparent background. We provide more examples of the SP-10 dataset in the supplementary material.

Our division and statistics of the dataset are shown in Table 2. We randomly divide the data set according to id. Each sample in the training set has ten video and spoken translation texts in different sign languages. Some samples in the test set lack some types of sign language data. More detailed dataset division distribution images and data distribution data for different sign languages are also provided in the appendix.

5. Experiments

5.1. Implementation Details

The input images are resized to 600×600, and we use EfficientNet [48] pre-trained on Imagenet [7] to extract fea-

[1] The copyright of the data belongs to the non-profit association European Sign Language Centre. For related terms of service, please refer to https://www.spreadthesign.com

[2] https://github.com/ageitgey/face_recognition

Table 3. Comparison of the translation results from multiple sign languages to English on the SP-10 dataset. Among them, Single represents the models trained in a single language pair using the method in [3], and Multi represents the multilingual translation model trained using the method in [20].

Part / Metrics	Method	Param ($\times 10^7$)	CSL	UKL	RSL	BQN	ICL	GSG	ISE	SWL	LLS	BFI	Mean
Dev / BLEU4	Single [3]	10 × 10	4.13	**5.65**	4.75	**4.60**	6.24	**5.61**	6.57	**6.88**	5.22	6.9	5.65
	Multi [20]	10	2.46	3.14	2.93	2.21	3.44	2.71	3.18	2.89	1.81	3.49	2.83
	MLSLT (ours)	10	**5.16**	5.42	**4.95**	3.28	**6.76**	5.18	**7.05**	6.33	**6.08**	**7.03**	**5.72**
Dev / ROUGE	Single [3]	10 × 10	30.98	33.58	**32.16**	**29.71**	34.03	32.83	33.49	35.66	30.77	34.64	32.78
	Multi [20]	10	28.50	28.93	30.01	24.66	29.91	29.75	28.33	31.01	27.7	32.42	29.12
	MLSLT (ours)	10	**34.59**	**34.04**	31.62	27.98	**35.29**	**33.5**	**37.96**	**36.02**	**34.48**	**37.25**	**34.27**
Test / BLEU4	Single [3]	10 × 10	3.47	2.94	2.99	2.46	**4.32**	3.60	3.23	**3.83**	3.83	3.72	3.43
	Multi [20]	10	2.28	2.38	2.06	1.10	1.38	1.82	2.09	2.13	2.68	3.27	2.12
	MLSLT (ours)	10	**5.19**	**4.18**	**3.66**	**2.85**	3.93	**4.97**	**6.70**	3.70	**5.72**	**5.73**	**4.66**
Test / ROUGE	Single [3]	10 × 10	30.16	**34.12**	31.32	**27.94**	32.03	31.36	29.42	31.72	29.02	31.60	30.86
	Multi [20]	10	29.37	28.63	29.57	23.95	28.53	29.36	29.30	29.83	30.03	30.76	28.93
	MLSLT (ours)	10	**33.33**	34.07	**31.54**	25.75	**33.25**	**32.13**	**35.37**	**33.09**	**33.11**	**35.34**	**32.70**

Table 4. Comparison of the translation results from British Sign Language to multiple spoken languages on the SP-10 dataset.

Part / Metrics	Method	Param ($\times 10^7$)	zh	uk	ru	bg	is	de	it	sv	lt	en	Mean
Dev/BLEU4	Single [3]	10 × 10	0.90	4.93	1.10	4.90	0.80	**6.18**	4.15	3.85	**3.21**	6.90	3.69
	MLSLT (ours)	10	**4.56**	**4.95**	**3.34**	**6.42**	**5.11**	5.78	**4.16**	**5.52**	2.79	**7.03**	**4.96**
Dev/ROUGE	Single [3]	10 × 10	35.21	30.77	25.28	30.85	27.46	33.29	27.51	32.52	**36.11**	34.64	31.36
	MLSLT (ours)	10	**36.76**	**31.97**	**32.80**	**33.03**	**34.24**	**33.60**	**29.88**	**33.95**	35.20	**37.25**	**33.87**
Test/BLEU4	Single [3]	10 × 10	1.03	**1.69**	1.10	1.10	0.90	**3.27**	1.38	1.11	2.19	3.72	1.75
	MLSLT (ours)	10	**5.35**	1.59	**1.12**	1.10	**2.54**	1.73	**1.39**	**2.63**	**2.32**	**5.73**	**2.55**
Test/ROUGE	Single [3]	10 × 10	**36.86**	27.43	25.43	27.45	28.34	**31.68**	26.97	28.95	32.24	31.60	29.69
	MLSLT (ours)	10	34.28	**29.78**	**31.67**	**28.73**	**30.22**	31.38	**27.14**	**32.91**	**33.25**	**35.34**	**31.47**

tures. The dimensions of the linear layer in the sign embedding layer and the word embedding layer are both set to 512. The hidden size, feed-forward size, and the number of attention heads in the Transformer layer are set to 512, 2048, and 8, respectively. Our model contains three encoder and decoder layers, and the dropout rate is set to 0.1 to prevent overfitting. We use Xavier initialization [12] to initialize our network. Following [2], we use BLEU [39] score and ROUGE [29] score to evaluate the performance of sign language translation.

5.2. Training and Inference

We use a single NVIDIA RTX 2080ti GPU with a batch size of 32 to train our model. We use the Adam [22] optimizer with a learning rate of 5×10^{-4} (β_1=0.9, β_2=0.998, $\epsilon = 10^{-8}$), and the weight decays to 10^{-3}. When the BLEU score of the validation set stops increasing 9 times, the learning rate will decrease by a factor of 0.5 until it reaches 10^{-7}. During training, λ_1 and λ_2 are set to 1 and 0.1, respectively. The rate of label smoothing [33] ε=0.2. For decoding in the inference process, we use the beam search strategy with beam size = 3 and the length normalization penalty $\alpha = 1$

5.3. Experimental Results

We set up two comparison benchmarks. We use the method in [3] to train the model as the benchmark for the BSLT model, and we use the method in [20] to train the model as the benchmark for the MSLT model. To make a fair comparison, we control the parameters of our model and report the parameters of all models in the experimental results.

5.3.1 Many to One

In this section, we explore the application of our method when there is multiple source sign language and a single target language. The experimental results of multiple sign languages translation to English are shown in Table 3. We can see that the performance of our model surpasses the MSLT baseline on all language pairs, which proves that our model can better handle MSLT tasks. In addition, the performance of our model on most language pairs and the average performance of our model exceed the BSLT baseline, although the model parameters are only one-tenth of theirs. This shows that the method we designed can help to share knowledge between different languages and minimize conflicts. On a few language pairs, the translation performance of our model is a little lower than the bilingual sign language translation baseline (for example, BQN→en). A reasonable assumption is that the difference between BQN and other sign languages is relatively more enormous. The experimental results of the other nine spoken languages as target languages are provided in the appendix.

5.3.2 One to Many

In this section, we explore the application of our method when there is a single source sign language and multiple target languages. The experimental results of the translation of British Sign Language into a variety of spoken languages are shown in Table 4. The performance of the MSLT baseline is very poor under this setting, so we removed it from the table and directly compared our model with the BLST baseline. It can be seen that MLSLT is higher than the BSLT baseline on both the BLEU and ROUGE metrics, which shows that the method we designed is also applicable to the many to one situation. It is worth noting that the performance of the two models is relatively poor in some language pairs (for example, BFI→ru). A reasonable hypothesis is that the grammar of Russian is more complex and has a sizeable grammatical difference with British Sign Language, which makes it difficult for the model to generate smooth Russian sentences. The experimental results of the other nine sign languages as source languages are also provided in the appendix.

Table 5. Comparison of many-to-many translation results and BSLT baseline results.

Language Pairs	Single		MLSLT		Param ratio
	BLEU4	ROUGE	BLEU4	ROUGE	
2×2	2.43	30.05	**3.62**	**33.76**	4:1
3×3	2.53	30.54	**3.34**	**32.72**	9:1
4×4	2.24	29.55	**2.72**	**31.69**	16:1
5×5	2.37	29.00	**2.81**	**31.25**	25:1
6×6	2.26	30.01	**3.05**	**31.58**	36:1
7×7	2.02	29.17	**2.49**	**29.96**	49:1
8×8	1.78	28.87	**2.04**	**30.00**	64:1
9×9	1.71	28.92	**2.09**	**30.01**	81:1
10×10	1.66	29.27	**1.88**	**30.26**	100:1

5.3.3 Many to Many

In this section, we report the experimental results of a model with multiple sign language inputs and multiple spoken language outputs, which is the most difficult situation. In order to judge where the performance limit of our model is, we gradually increase the number of language pairs that the model needs to process and compare the average result of MLSLT with the BSLT baseline. The experimental results are shown in Table 5. It can be seen that our model can achieve better results than the BSLT baseline at the beginning, and the gap between them gradually becomes smaller as the number of language pairs increases. This is because we need to use a fixed model capacity to handle more and more language pairs, so the model will gradually encounter a capacity bottleneck. However, even when processing 100 language pairs, our model can achieve performance that exceeds the BSLT baseline, while our parameter amount is only one percent of it. This shows that the method we designed allows the model to efficiently use the existing parameters. Increasing the parameters of the model also help alleviate the capacity bottleneck problem.

Table 6. Comparison of zero-shot translation results and BSLT baseline results.

Language Pair	BLEU4			ROUGE		
	Single	Zero-shot	diff	Single	Zero-shot	diff
CSL→en	3.47	2.69	-0.78	30.16	30.68	0.52
UKL→en	2.94	2.65	-0.29	34.12	30.62	-3.5
RSL→en	2.99	2.40	-0.59	31.32	30.20	-1.12
BQN→en	2.46	1.03	-1.43	27.94	26.05	-1.89
ICL→en	4.32	1.78	-2.54	32.03	30.12	-1.91
GSG→en	3.60	3.44	-0.16	31.36	30.37	-0.99
ISE→en	3.23	1.69	-1.54	29.42	29.65	0.23
SWL→en	3.83	3.29	-0.54	31.72	31.96	0.24
LLS→en	3.83	3.14	-0.69	29.02	31.47	2.45
BFI→en	3.72	3.36	-0.36	31.60	31.92	0.32

5.3.4 Zero-Shot Translation

An interesting benefit of building a multilingual sign language translation model is that it allows performing zero-shot translation between a language pair for which no explicit parallel training data has been seen without any modification to the model. When performing zero-shot translation, although there is no training data supervision between the input sign language and the target language, the existing training data builds an implicit bridge between them to achieve correct translation. The model will try to map all sign languages to a shared semantic space, and then the model will learn how to translate sentences with the same semantics into different spoken texts. In Table 6, we show the performance of zero-shot translation from multiple sign languages to English and the performance comparison with the supervised translation model. All zero-shot models are trained with $4 \times (4 - 1)$ language pairs. It can be seen that even in the absence of supervised data, our model can still achieve reasonable translation performance through zero-shot translation. Even in some language pairs, the zero-shot translation model is a little bit better than the BSLT baseline. The language pairs used during training have an important impact on the performance of zero-shot translation. Related analysis experiments are provided in the appendix.

Table 7. Results of ablation experiments on the SP-10 dataset. We gradually add the methods we mentioned earlier to verify their effectiveness.

Model	BLEU1	BLEU4	ROUGE
Native SLT	26.01	2.27	29.80
+IntraLSR	32.35	5.19	33.33
+InterLSR	30.49	3.58	32.20
+IntraLSR+InterLSR (MLSLT)	**34.43**	**6.30**	**35.76**

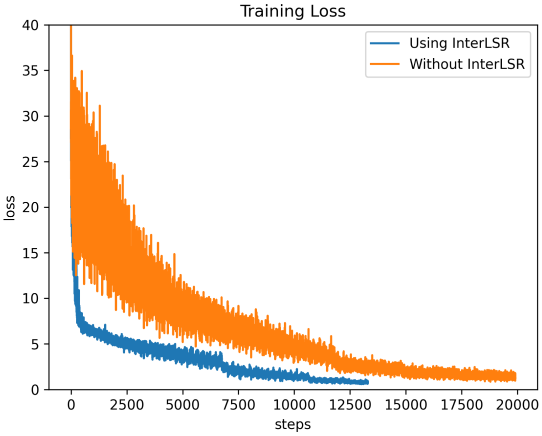


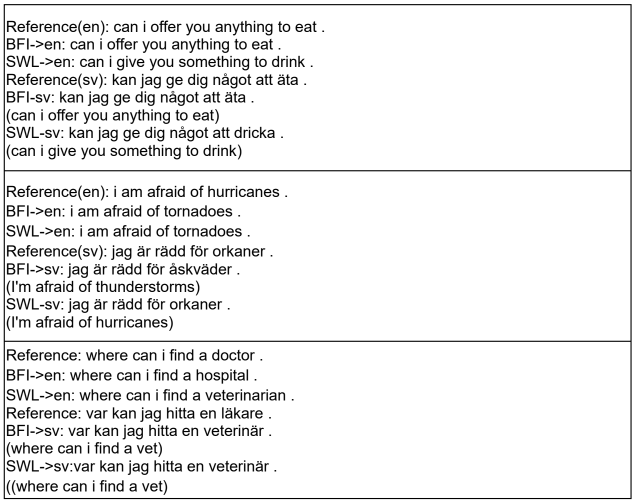
Figure 4. The impact of using and not using InterLSR on model optimization.

5.4. Ablation Study

In this subsection, we will introduce the results of our ablation experiments on the SP-10 dataset and analyze the effectiveness of our proposed method through the experimental results. The experimental results are shown in Table 7. As shown in the third row of Table 7, the performance of the model after the addition of IntraLSR has been significantly prompted. As shown in the fourth row of Table 7, the performance improvement of the corresponding model of InterLSR alone is helpful but not as obvious as IntraLSR. Adding IntraLSR and InterLSR at the same time can achieve the best results. In addition, we also found that adding InterLSR can help the model converge faster and more stably during training. The training loss curves of the two models with and without InterLSR are shown in Figure 4. It can be seen that the jitter of the model training loss after adding InterLSR is much smaller, and the training is finished about 8k steps earlier than the model without InterLSR.

5.5. Qualitative Results

In this section, we show some qualitative results of our model. As shown in Figure 5, we compare the translation results output by MLSLT with the reference text and provide English translations for other languages. Our model can output sentences with relatively complete semantics but will confuse some words that are similar in sign language. As shown in the first line, our model recognizes ”eat” as ”drink” because these two words are very similar in SWL. In addition, the ”hurricanes” and ”doctor” in the second and third rows are also for similar reasons.



Reference(en): can i offer you anything to eat .
BFI->en: can i offer you anything to eat .
SWL->en: can i give you something to drink .
Reference(sv): kan jag ge dig något att äta .
BFI-sv: kan jag ge dig något att äta .
(can i offer you anything to eat)
SWL-sv: kan jag ge dig något att dricka .
(can i give you something to drink)

Reference(en): i am afraid of hurricanes .
BFI->en: i am afraid of tornadoes .
SWL->en: i am afraid of tornadoes .
Reference(sv): jag är rädd för orkaner .
BFI->sv: jag är rädd för åskväder .
(I'm afraid of thunderstorms)
SWL-sv: jag är rädd för orkaner .
(I'm afraid of hurricanes)

Reference: where can i find a doctor .
BFI->en: where can i find a hospital .
SWL->en: where can i find a veterinarian .
Reference: var kan jag hitta en läkare .
BFI->sv: var kan jag hitta en veterinär .
(where can i find a vet)
SWL->sv:var kan jag hitta en veterinär .
((where can i find a vet)

Figure 5. Comparison of the spoken language generated by the MLSLT model and the reference text.

6. Conclusion

In this paper, we introduce a challenging task, multilingual sign language translation (MSLT), and propose the first MSLT model, MLSLT. Compared with the previous research, we try to use a single model to complete the translation between multiple language pairs. To reduce the conflict between different languages, we propose two novel dynamic routing mechanisms. They dynamically adjust the flow of data from the language level and token level, respectively. To evaluate the effectiveness of our proposed method, we created the first publicly available multilingual sign language understanding dataset, SP-10. Compared with the previous dataset, SP-10 contains more language pairs, and the pairing information between different sign languages creates possibilities for multilingual text-to-video generation tasks and video-to-video translation tasks. We conduct extensive experiments on this dataset to support future research and prove the effectiveness of our proposed method. We discuss the limitations and potential negative effects of our work in the appendix.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62072397, and Zhejiang Natural Science Foundation LR19F020006. This work was also partially supported by the Huawei Noah’s Ark Lab.

References

[1] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 343–351, 2016. 2, 5

[2] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 2, 6

[3] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE, 2020. 1, 2, 6

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017. 2

[6] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5):99:1–99:38, 2020. 2, 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 5

[8] Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giró-i-Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2735–2744. Computer Vision Foundation / IEEE, 2021. 2, 5

[9] Ethnologue. Ethnologue languages of the world. https://www.ethnologue.com/, 2021. 1

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019. 2

[11] Shiwei Gan, Yafeng Yin, Zhiwei Jiang, Lei Xie, and Sanglu Lu. Skeleton-aware neural sign language translation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4353–4361. ACM, 2021. 3

[12] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org, 2010. 6

[13] Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 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, 2020. 5

[14] Benjamin Heinzerling and Michael Strube. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). 3

[15] Marlene Hilzensauer and Klaudia Krammer. A multilingual dictionary for sign languages:" spreadthesign. *ICERI2015 Proceedings*, pages 7826–7834, 2015. 2

[16] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign language recognition using 3d convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo, ICME 2015, Turin, Italy, June 29 - July 3, 2015*, pages 1–6. IEEE Computer Society, 2015. 2

[17] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2257–2264. AAAI Press, 2018. 2, 5

[18] ISO Central Secretary. Codes for the representation of names of languages — part 1: Alpha-2 code. Standard ISO 639-1:2002, International Organization for Standardization, Geneva, CH, 2002. 3

[19] ISO Central Secretary. Codes for the representation of names of languages — part 3: Alpha-3 code for comprehensive coverage of languages. Standard ISO 639-3:2007, International Organization for Standardization, Geneva, CH, 2007. 3

[20] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351, 2017. 2, 3, 6

[21] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 100. BMVA Press, 2019. 2

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on*

Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 6

[23] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3416–3424. IEEE Computer Society, 2017. 1

[24] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms. *Int. J. Comput. Vis.*, 126(12):1311–1325, 2018. 1

[25] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics, 2018. 3

[26] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020. 2

[27] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 3

[28] 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. 2

[29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[30] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. *CoRR*, abs/2108.11515, 2021. 5

[31] Tao Liu, Wengang Zhou, and Houqiang Li. Sign language recognition with long short-term memory. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 2871–2875. IEEE, 2016. 2

[32] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, pages 3025–3036. Association for Computational Linguistics, 2019. 3

[33] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705, 2019. 4, 6

[34] United Nations. International day of sign languages 23 september. https://www.un.org/en/observances/sign-languages-day, 2021. 1

[35] Carol Neidle and Christian Vogler. A new web interface to facilitate access to corpora: Development of the asllrp data access interface (dai). In *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*, volume 3. Citeseer, 2012. 5

[36] Sylvie CW Ong and Surendra Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(06):873–891, 2005. 2

[37] Alptekin Orbay and Lale Akarun. Neural sign language translation by learning tokenization. In *15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Buenos Aires, Argentina, November 16-20, 2020*, pages 222–228. IEEE, 2020. 1, 3, 5

[38] World Health Organization. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2021. 1

[39] 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002. 6

[40] Suzanne Romaine. Language endangerment and language death: the future of language diversity. In *The Routledge Handbook of Ecolinguistics*, pages 40–55. Routledge, 2017. 1

[41] Hirohiko Sagawa, Hiroshi Sakou, and Masahiro Abe. Sign Language Translation System Using Continuous DP Matching. *Journal of Machine Vision and Applications*, pages 339–342, 1992. 2

[42] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. Building the british sign language corpus. *Language Documentation & Conservation*, 7:136–154, 2013. 5

[43] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. 3

[44] T Starner and A Pentland. Real-time American Sign Language recognition from video using hidden Markov models. In *Proceedings of International Symposium on Computer Vision - ISCV*, pages 265–270, 1995. 2

[45] 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. 2

[46] S Tamura and S Kawasaki. Recognition of sign language motion images. *Pattern Recognition*, 21(4):343–353, 1988. 2

[47] Shinichi Tamura and Shingo Kawasaki. Recognition of sign language motion images. *Pattern Recognition*, 21(4):343–353, 1988. 2

[48] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. 3, 5

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 2

[50] U Von Agris and KF Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 243–246, 2010. 5

[51] Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[52] Aoxiong Yin, Zhou Zhao, Jinglin Liu, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simulslt: End-to-end simultaneous sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4118–4127, 2021. 1, 3

[53] Kayo Yin and Jesse Read. Better sign language translation with stmc-transformer. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5975–5989. International Committee on Computational Linguistics, 2020. 1

[54] Morteza Zahedi, Philippe Dreuw, David Rybach, Thomas Deselaers, and Hermann Ney. Continuous sign language recognition-approaches from speech recognition and available data resources. In *Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pages 21–24, 2006. 5

[55] Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. Share or not? learning to schedule language-specific capacity for multilingual translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[56] Jihai Zhang, Wengang Zhou, and Houqiang Li. A threshold-based HMM-DTW approach for continuous sign language recognition. In *International Conference on Internet Multimedia Computing and Service, ICIMCS '14, Xiamen, China, July 10-12, 2014*, page 237. ACM, 2014. 1

[57] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1316–1325. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 5