electronics

MDPI

Review

Sign Language Translation: A Survey of Approaches and Techniques

Zeyu Liang, Huailing Li * and Jianping Chai

School of Data Science and Intelligent Media, Communication University of China, Beijing 100024, China; zeyuliang@cuc.edu.cn (Z.L.); jp_chai@cuc.edu.cn (J.C.)
* Correspondence: huailingli@cuc.edu.cn

Abstract: Sign language is the main communication way for deaf and hard-of-hearing (i.e., DHH) people, which is unfamiliar to most non-deaf and hard-of-hearing (non-DHH) people. To break down the communication barriers between DHH and non-DHH people and to better promote communication among DHH individuals, we have summarized the research progress on sign language translation. We provide the necessary background on sign language translation and introduce its four subtasks (i.e., sign2gloss2text, sign2text, sign2(gloss+text), and gloss2text). We distill the basic mode of sign language translation (SLT) and introduce the transformer-based framework of SLT. We analyze the main challenges of SLT and propose possible directions for its development.

Keywords: sign language translation; transformer; DHH; sign language recognition

check for updates

Citation: Liang, Z.; Li, H.; Chai, J. Sign Language Translation: A Survey of Approaches and Techniques. *Electronics* **2023**, *12*, 2678. https://doi.org/10.3390/electronics12122678

Academic Editors: Pedro Javier Herrera Caro and Jaime Duque-Domingo

Received: 13 May 2023
Revised: 9 June 2023
Accepted: 11 June 2023
Published: 15 June 2023

1. Introduction

Sign language is a special visual language for both congenital DHH and acquired DHH people, and it uses both manual and nonmanual information [1] for visual communication. Manual information includes shape, orientation, position, and motion of hands, while nonmanual information includes body posture, arm movements [2], eye gaze, lip shape, and facial expressions [3]. Sign language is not a simple word-for-word translation of spoken language but has independent grammar, semantic structure, and specific language logic [4]. Continuous changes in hand and body movements represent different units of meaning. According to statistics by the World Federation of the Deaf, the number of DHH people is 70 million, and there are over 200 sign languages in the world [5]. Therefore, improving the translation technology of sign language can bridge the communication gap between DHH and non-DHH individuals.

For the task of sign language translation (SLT), previous works [6–8] were mainly focused on sign language recognition (SLR), recognizing sign language as corresponding gloss. However, SLT is a conversion of recognized gloss into spoken language text, which is not a direct prediction [9] of the spoken language text from sign language videos. Unlike the spoken language text, gloss [10] includes the grammatical and semantic information on tense, order, and direction or position in sign language. Gloss may also include information about the repeated number of a sign. Figure 1 shows the difference between SLR and SLT.

SLR can be divided into two categories: isolated sign recognition [11–13] and continuous sign recognition [7,14,15]. The former refers to the fine-grained recognition of individual sign movements, where one video corresponds to only one gloss, and segmenting the sign videos requires a large amount of manual effort. The latter maps continuous sign language videos into one sequence of glosses, where the order of glosses in the sequence is consistent with sign language.

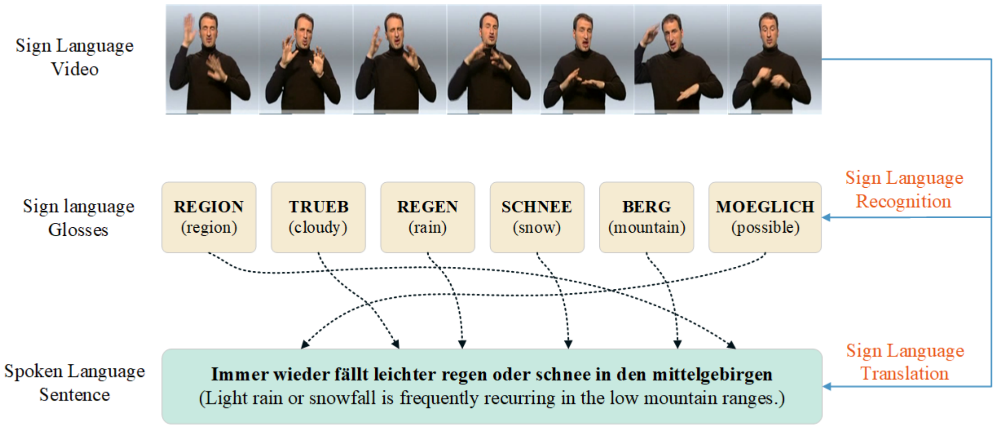
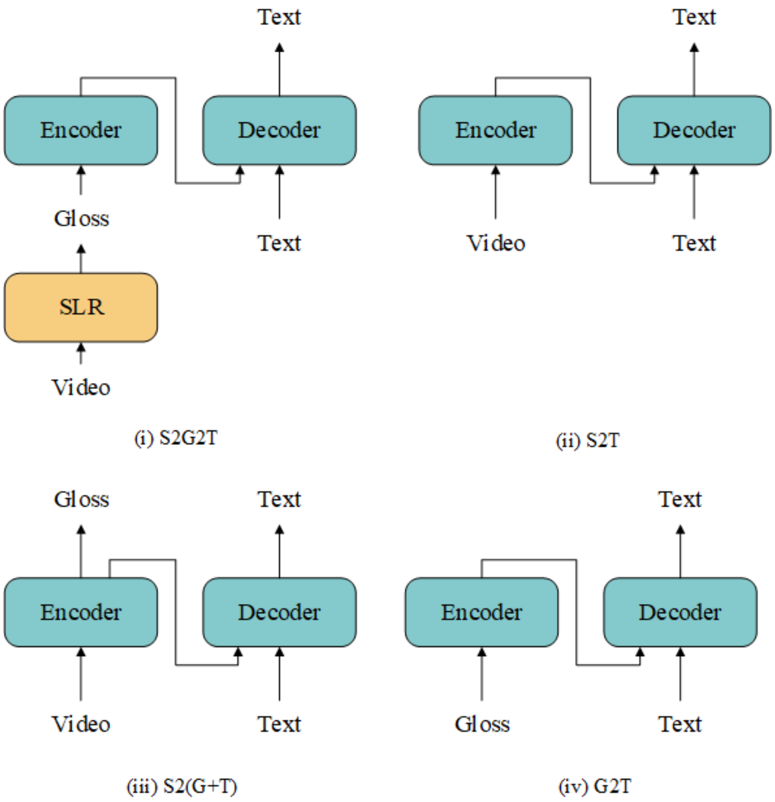


Figure 1. The difference between SLR and SLT.

With a deep neural network employed in the area of natural language processing, Camgoz et al. [9] treated SLT as ordinary spoken language text. Unlike SLR, they recognized sign language as a sequence of glosses. They aimed to generate spoken language text that non-sign language users could also understand. As shown in Figure 2, there are four common frameworks [16]:

- Sign2gloss2text (S2G2T) [17–19], which recognizes the sign language video as gloss annotations first and then translates the gloss into spoken language text.
- Sign2text (S2T), which directly generates spoken language text from sign language video end to end.
- Sign2(gloss+text) (S2(G+T)) [16,20,21], which multitasks by outputting glosses and text and can use external glosses as supervision signals.
- Gloss2text (G2T) [22–24], which can reflect the translation performance from gloss sequence to text.



(i) S2G2T

(ii) S2T

(iii) S2(G+T)

(iv) G2T

Figure 2. The four common protocols for SLR and SLT.

In this work, we aimed to classify and summarize the literature into three types: improving the accuracy of SLR, improving the performance of SLT through different models, and addressing the scarcity of data resources. The first type of literature introduces how to improve the performance of SLR through the task of SLR, and the second type of literature introduces how to modify the structure of the network for better capturing the visual and textual semantic information. The third type of literature aims to settle the problem of the scarcity of sign language for SLT. Finally, we introduce the most common datasets and evaluation metrics employed on the task of SLT.

We noticed that there were some other well written review articles on the same topic in the published literature [5,25–27], and our manuscript is different from the current published literature in the following aspects: Firstly, our manuscript clearly analyze the concepts of sign language recognition and translation, explain the common areas of SLR and SLT, and clarify the boundaries of SLR and SLT. SLR can be regarded as a substep of SLT, a two-step process of first recognition and then translation, but SLT can also be implemented without relying on SLR, in an end-to-end way, namely, S2T. Secondly, according to the characteristics of SLT, we analyze and compare the existing technologies and methods. We classify them into three types: improving the performance of SLR, changing the network structure for improving the performance of the translation, and solving the problem of sign language scarcity. The existing review literature may be presented in the chronological order of SLT, without classification, which is difficult for people to grasp the various problems of sign language translation. Finally, we describe the latest situation in the field of SLT, which brings readers a broad vision and brand-new inspiration. Table 1 shows the current differences between past review papers and our research.

Table 1. Current differences between past review papers and our research.

Literatures	Task	Classification	Latest Year
Minu et al. [25]	SLR	Traditional and modern	2022
Marcos et al. [5]	SLT	Year	2022
Baumgärtner et al. [26]	SLR/SLT/SLG	SLR/SLT/SLG	2019
Koller et al. [27]	SLR	None	2020
Ours	SLT	Three specific types	2023

In Section 2, we introduce the basic framework of SLT. In Section 3, we present the methods or models used for SLT. In Sections 4 and 5, we introduce the datasets and metrics employed for the task of SLT. In Section 6, we discuss the current challenges for the task of SLT.

2. The Background of SLT

SLT is a typical sequence-to-sequence problem that translates continuous sign language videos into fluent spoken language text. The celebrated encoder–decoder architecture of SLT is shown in Figure 3.

Firstly, the rich semantic information in sign language video frames is first encoded into dense vectors. Secondly, the decoder module takes the dense vectors as input and generates the target spoken text, sequentially. The framework of SLT consists of three modules:

- Spatial and word embedding layers, which map the sign language video frames and spoken text into feature vectors or dense vectors, respectively.
- A tokenization layer, which tokenizes the feature vectors.
- The encoder–decoder module, which predicts the spoken text and adjusts the network parameters through backpropagation to reduce the difference between the target text and generated text.

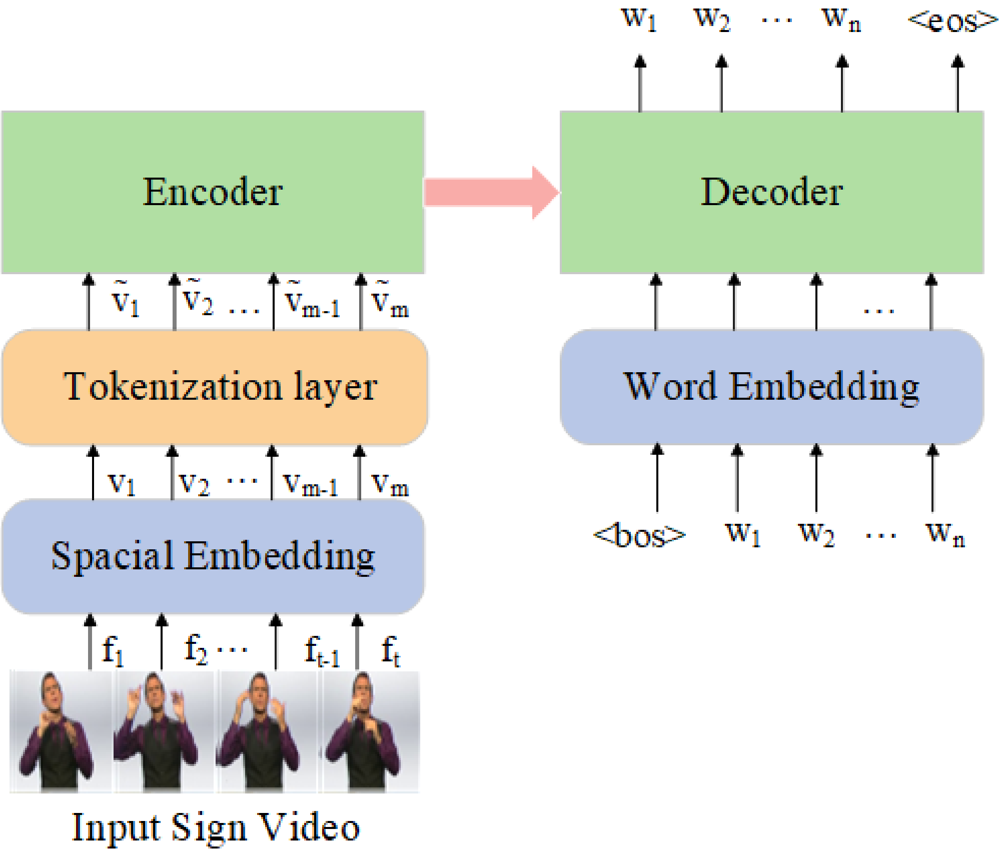


Figure 3. The encoder and decoder framework employed in SLT.

The spatial embedding layer extracts features from the visual information of sign language, and there are various network structures, such as 2D-CNN [28], 3D-CNN [29], and GCN [30,31]. In addition to the general structures, the multi-Cue network have also been applied when tailored to sign language visual information. The clue features, such as facial expressions, body posture, and gesture movements, are fused through corresponding fusion mechanisms and then fed into the tokenization layer.

The word-embedding module can learn a dense vector through a linear projection layer. For the tokenization layer, both "frame-level" and "gloss-level" token schemes are available, and RNN-HMM is a typical method for the "gloss level". The encoder–decoder module may consist of multiple RNN or LSTM cells and their variants such as Bi-LSTM or GRU cells. To address the long-term dependency issue, multiple attention mechanisms can be incorporated, such as in Bahdanau et al. [32] and Luong et al. [33]. Moreover, other structures such as graph convolutional networks and transformers [34] have also been employed in SLT.

Figure 4 shows the celebrated transformer framework employed in SLT. Firstly, the visual information is obtained through structures such as S3D [35], OpenPose [36], VGG-Net [37], and STMC [1], and undergoes spatial embedding. Next, the semantic representation of visual and textual information enters N encoder modules based on the self-attention mechanisms in the transformers. In the decoder module, the input of word embeddings enters a masked multihead attention module, with masking indicating the usage of only existing tokens when extracting contextual information to prevent overfitting of the model. Finally, the predicted words are outputted via a softmax layer.

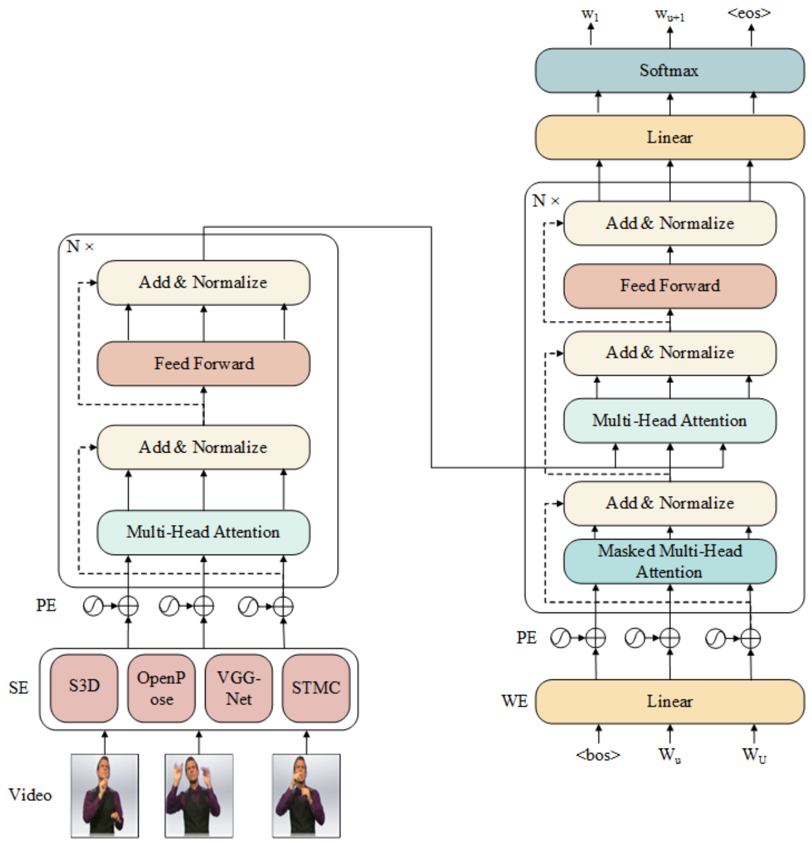


Figure 4. The transformer framework employed in SLT.

3. Literature Review of SLT

In this section, we classify the SLT into three types: improving the performance of SLR, network structure for improving the performance of translation, and solving the problem of the scarcity of sign language. The first type of literature aims to improve the performance of SLT by improving the performance of SLR. The second type of literature improves the performance of SLT by modifying the structure of the network. The third type of literature aims to solve the problem of the scarcity of sign language.

3.1. Improving the Performance of SLR

SLR can be regarded as a substep of SLT, a two-step process of first recognition and then translation, but SLT can also be implemented without relying on SLR in an end-to-end way, namely, S2T. Experiments have shown that improving the performance of sign language recognition is an effective way to improve sign language translation. In this section, we review these solutions in detail.

To better capture the global visual semantic information, He et al. [38] employed the faster R-CNN model to locate and recognize hand gestures in sign language videos. They combined a 3D-CNN network with the LSTM-based encoder–decoder framework for SLR. To meet the high accurate demand of sign language video segmentation, Li et al. [4] proposed a temporal semantic pyramid model to partition the video into segments with different levels of granularity. They employed the time-based hierarchical feature learning method and attention mechanisms to learn local information and non-local contextual information. To explore the precise action boundaries and learn the temporal cues in sign language videos, Guo et al. [39] proposed a hierarchical fusion model to obtain the visual information with different visual granularities. First, a 3D-CNN framework and a Kinect device were used for extracting the RGB features and skeleton descriptors, separately. Then, an adaptive clip summarization (ACS) framework was proposed for automatically selecting key clips or frames of variable sizes. Next, multilayer LSTMs were employed to learn features at the frame, clip, and viseme/signeme levels. Finally, a query-adaptive model was designed to generate the target spoken text. To fully leverage the significant information of body postures and positions, Gan et al. [40] proposed a skeleton-aware model, which considered the skeletons as a corresponding representation of human postures, and they sliced the video into clips. To improve the robustness of the proposed model, Kim et al. [41] proposed a robust key-point normalization method which normalized the position of key points through the neck–shoulder framework. The normalized key-points were then used as input sequence for a transformer network. For the area of German sign language, there exists an obvious problem, where some signs have the same hand gesture but differ only in lip shape. To settle this problem, Zheng et al. [3] proposed a semantic focus model for extracting facial expression features for German sign language.

Unlike some works [9,42–44] that focus more on specific appearance features, Rodriguez and Martinez [2] focused more on the motion variations in sign language and used optical flow images instead of RGB frames for SLT. They first used a 3D-CNN framework to obtain the optical flow representation and extract spatial motion patterns. Then, they used bidirectional recurrent neural networks for the motion analysis and for learning the nontemporal relationships of optical flow. Finally, they combined the gestural attention mechanism with spatiotemporal descriptors to enhance the correlation with the spoken language units. Different from some works [14,42,45] that only consider a single feature such as hand features, Camgoz et al. [46] took different articulators' information separately. They designed a multichannel transformer model to improve the performance of the feature extraction. Zhou et al. [1] proposed a spatiotemporal multicue model (STMC) to explore the hidden information contained in sign language videos.Unlike some common approaches that process multiple aspects of visual information contained in sign language, Kan et al. [47] proposed a novel model using graph neural networks to transform sign language features into hierarchical spatiotemporal graphs. Their hierarchical spatiotemporal graphs consisted of high-level and fine-level graphs, the former representing the states of the hands and face, while the latter represented more detailed patterns of hand joints and facial regions. Table 2 shows the performance of some celebrated models mentioned above.

Table 2. The performance of some celebrated models on the PHOENIX14 dataset.

Model	Dev		Test	
	WER	**del/ins**	**WER**	**del/ins**
SMC [1]	22.7	7.6/3.8	22.4	7.4/3.5
STMC [1]	21.1	7.7/3.4	20.7	7.4/2.6
HST-GNN [47]	19.5	-	19.8	-
Hybrid CNN-HMM [8]	31.6	-	32.5	-
CTF [48]	37.9	12.8/5.2	37.8	11.9/5.6
DenseTCN [49]	35.9	10.7/5.1	36.5	10.5/5.5
Song et al. [50]	38.1	12.7/5.5	38.3	11.9/5.6

3.2. Network Structure for Improving the Performance of Translation

To improve the performance of SLT, many researchers have proposed various advanced network frameworks for capturing deep visual and text semantic information. In this section, we review these solutions in detail.

Fang et al. [51] proposed a DeepASL system for American Sign Language (ASL) translation. First, DeepASL treated hand shape, relative position, and hand movement as the skeleton information of American sign language (ASL), employing a leap motion sensor device worn by the user. Then, a hierarchical bidirectional model was used to capture semantic information and generate word-level translation. Finally, a connectionist temporal classification was employed for sentence-level translation. Koller et al. [8] proposed a hybrid framework for sequence encoding and employed a Bayesian framework to generate sign language text. Wang et al. [48] proposed the connectionist temporal fusion (CTF) framework for SLT. First, the C3D-ResNet was employed to extract visual information from the video clips. Then, the visual information was sent into temporal convolution (TCOV) and bidirectional GRU (Bi-GRU) modules to obtain the long-term and short-term transitions, respectively. Next, a fusion module (FL) was employed to connect the modules and learn complementary relationships. Finally, a connectionist temporal fusion (CTF) mechanism was designed to generate sentences. Although their proposed model could solve the frame-level alignment problem, it could not address the correspondence between the jumbled word order and visual content. Therefore, Guo et al. [43] proposed a hierarchical-LSTM (H-LSTM) model, which embedded features at the frame-level, clip-level, and viseme-level. They used a C3D network [29] to extract the visual features and utilized an online adaptive key-segment mining method to remove irrelevant frames. They proposed three pooling strategies to reduce less important clips.

To the best of our knowledge, Camgoz et al. [9] were the first to take the sign language videos into spoken language text with an end-to-end model. Their model employed a 2D-CNN framework for spatial embedding, an RNN-HMM framework for word segmentation, and a sequence-to-sequence attention model for sequence mapping. To address the issues of gradient vanishing, Arvanitis et al. [52] employed a gated recurrent unit (GRU [53])-based seq2seq framework [32,54,55] for SLT. They employed three different Luong attention mechanisms [33] to calculate the weight parameters of hidden states. Guo et al. [49] proposed a dense temporal convolutional network (DenseTCN) that captured the details of sign language movements from short-term to long-term network. They used a 3D-CNN framework for extracting the visual representation and designed a temporal convolution (TC) framework [56] to capture local context information. Inspired by DenseNet [57], they expanded the network into a dense hierarchical structure to capture global context information and computed the CTC loss at the top of each layer to optimize the parameters. Camgoz et al. [16] argued that using the glosses as inputs could undermine the performance of the SLT system. To settle these issues, they proposed a transformer model which took the SLR and SLT in an end-to-end way, without requiring an explicit gloss representation. Yin and Read [58] proposed an STMC-Transformer framework for SLT, where the gloss sequence was identified by a transformer-based encoder–decoder network [16]. After exploring spatial and temporal cues in STMC [1], Yin and Read [17] predicted the gloss sequence using Bi-LSTM and CTC frameworks and generated spoken text using a transformer framework. To overcome the difficulty of modeling long-term dependencies and consuming a large quantity of resources, Zheng et al. [59] proposed a frame stream density compression framework and a temporal convolution and dynamic hierarchical model for SLT. Voskou et al. [60] proposed an SLT architecture with a novel layer in the transformer network that avoided using gloss sequences as an explicit representation. Qin et al. [61] proposed the video transformer net (VTN) framework for the task of SLR and SLT. Their framework was a lightweight SLT architecture that used the Resnet-34 and transformer framework as encoder and decoder, respectively. To address the weakly supervised problem in SLT, Song et al. [50] proposed a parallel temporal encoder framework to extract both global and local information simultaneously. To address the problem of multilanguage SLT (MSLT), Yin et al. [62] proposed a transformer model for translating different types of sign language into corresponding texts in an end-to-end way. To capture the nonlocal and global semantic information of the video and spoken language text, Guo et al. [63] proposed a locality-aware transformer (LAT) framework for the task of SLT. To acquire the multicue information of sign language, Zhou et al. [20] proposed a spatial–temporal multicue (STMC) framework for the task of SLT. Table 3 shows the performance of some of the celebrated models mentioned above.

Table 3. The performance of some celebrated models on the PHOENIX14T dataset.

Model	Dev							Test						
	BL-1	**BL-2**	**BL-3**	**BL-4**	**ROUGE**	**METEOR**	**WER**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**ROUGE**	**METEOR**	**WER**
TSPNet-Sequential (Li et al.) [4]	-	-	-	-	-	-	-	35.65	22.8	16.6	12.97	34.77		
TSPNet-Joint (Li et al.) [4]	-	-	-	-	-	-	-	36.1	23.12	16.88	13.41	34.96		
SANet (Gan et al.) [40]	56.6	41.5	31.2	23.5	54.2	-	-	57.3	42.4	32.2	24.8	54.8		
Multistream (Zheng et al.) [3]	-	-	-	10.76	34.81	-	-	-	-	-	10.73	34.75		
Camgoz et al. [46]	-	-	-	19.51	45.9	-	-	-	-	-	18.51	43.57		
Kan et al. [47]	46.1	33.4	27.5	22.6	-	-	-	45.2	34.7	27.1	22.3	-		
S2G2T (Camgoz et al.) [9]	42.88	30.3.	23.02	18.4	44.14	-	-	43.29	30.39	22.82	18.13	43.8		
Multiregion (Zheng et al.) [3]	-	-	-	10.94	34.96	-	-	-	-	-	10.89	34.88		
Sign2Text (Camgoz et al.) [9]	45.54	32.6	25.3	20.69	-	-	-	45.34	32.31	24.83	20.17	-		
Best Recog. Sign2(Gloss+Text) [16]	46.56	34.03	26.83	22.12	-	-	24.61	47.2	34.46	26.75	21.8	-		24.49
Best Trans. Sign2(Gloss+Text) [16]	47.26	34.4	27.05	22.38	-	-	24.98	46.61	33.73	26.19	21.32	-		26.16
STMC-Transformer [17]	48.27	35.2	27.47	22.47	46.31	44.95	-	48.73	36.53	29.03	24.0	46.77	45.78	
STMC-Transformer Ens. [17]	50.31	37.6	29.81	24.68	48.7	47.45	-	50.63	38.36	30.58	25.4	48.78	47.6	
Zheng et al. [59]	31.43	19.12	13.4	10.35	32.76	-	-	31.86	19.51	13.81	10.73	32.99		
Voskou et al. (Yin and Read) [60]	49.12	36.29	28.34	23.23	-	-	-	48.61	35.97	28.37	23.65	-		
Multitask (Orbay and Akarun) [64]	-	-	-	-	-	-	-	37.22	23.88	17.08	13.25	36.28		
BN-TIN-Transf.2+BT [19]	49.33	36.43	28.66	23.51	49.53	-	-	48.55	36.13	28.47	23.51	49.35		
BN-TIN-Transf.+SignBT [19]	51.11	37.9	29.8	24.45	50.29	-	-	50.8	37.75	29.72	24.32	49.54		
BERT2RND (Coster et al.) [21]	-	-	-	22.47	-	-	36.59	-	-	-	22.25	-		35.76
BERT2BERT (Coster et al.) [21]	-	-	-	21.26	-	-	40.99	-	-	-	21.16	-		39.99
mBART-50 (Coster et al.) [21]	-	-	-	17.06	-	-	40.25	-	-	-	16.64	-		39.43
BERT2RNDff Sign2Text [65]	-	-	-	21.58	47.36	-	-	-	-	-	21.39	46.67		
BERT2RNDff Sign2(Gloss+Text) [65]	-	-	-	21.97	47.54	-	-	-	-	-	21.52	47		
Zhao et al. [66]	35.85	24.77	18.65	15.08	38.96	22.0	70.0	36.71	25.4	18.86	15.18	38.85	21.0	72
ConSLT (Fu et al.) [67]	50.47	37.54	29.62	24.31	-	-	-	51.29	38.62	30.79	25.48	-		
Sign2Gloss2Text [68]	50.36	37.5	29.69	24.63	50.23	-	-	49.94	37.28	29.67	24.6	49.59		
Sign2Text (Chen et al.) [68]	53.95	41.12	33.14	27.61	53.1	-	-	53.97	41.75	33.84	28.39	52.65		
TIN-SLT (Cao et al.) [24]	-	-	-	-	-	-	-	51.06	38.85	31.23	26.13	48.56	47.83	

3.3. Solving the Problem of the Scarcity of Sign Language

As we all know, recording high-quality sign language data in large quantities is extremely expensive, and the sign–German pairs of RWTH-PHOENIX-Weather-2014T has less than 9000 samples [19], which is an order of magnitude smaller than the task of neural machine translation [69]. Data scarcity is a major challenge and bottleneck for the task of SLT. To settle this issue, many solutions have emerged, such as backtranslation, data augmentation, transfer learning, and leveraging generative models. In this section, we review these solutions in detail.

Orbay et al. [64] considered that using gloss as supervision data in SLT could improve translation performance. Since the quantity of gloss data is limited and expensive to obtain, they explored semisupervised labeling methods. They proposed two labeling methods: the first method utilized OPENPose [70] to extract hands from video frames, then hand shape recognition was performed using a 2D-CNN. The second approach employed a pretrained model for action recognition. Experiments showed that frame-level labeling may be better than scarce gloss, 3D-CNNs may be more effective for SLT in the future, and the labeling of the right-hand information contributed more to translation quality. Because of a lack of corpora between Myanmar sign language (MSL) and Myanmar language, Moe et al. [71] investigated unsupervised neural machine translation (U-NMT) in Myanmar. To settle the issue of sparse data annotations, Albanie et al. [72] proposed an automatic annotation method, and they proposed the BSL-1K dataset. To improve the performance of SLT, Zhou et al. [19] used a large number of monolingual texts to augment the dataset of SLT. Inspired by backtranslation models [73], they introduced the SignBT algorithm, which added newly generated parallel sample pairs to the dataset. To settle the scarcity of sign language, Nunnari et al. [74] proposed a data augmentation model, which could help eliminate the background and personal effects of the signer. Gomez et al. [75] proposed a transformer model for text-to-sign gloss translation. Their proposed model recognized syntactic information and enhanced the discriminative power for low-resource SLT tasks without significantly increasing model complexity.

To address the lack of parallel corpus pairs, Coster et al. [21] proposed a frozen pretrained transformer (FPT) model, and they initialize the transformer translation model with pretrained BERT-based and mBART-50 models. Coster et al. [65] continued their research on the effectiveness of frozen pretrained model, and they found that the performance improvement was not due to the changes of their proposed model but rather to the written language corpora. To settle the issue of small SL datasets, Zhao et al. [66] aimed to learn the linguistic characteristics of spoken language to improve translation performance. They proposed a framework consisting of a verification model that queried whether words existed in sign language videos, a pretrained conditional sentence generation module that combined the existing words into multiple sentences, and a cross-modal reranking model was employed to select the best-fit sentence.

To address the low-resource problem, Fu et al. [67] proposed a novel contrastive learning model for the task of SLT. They fed the recognized gloss twice to the transformer translation network and used the hidden layer representations as two types of "positive examples". Correspondingly, they randomly selected K tokens from the vocabulary as "negative examples", which were not present in the current sentence. With the progress of transfer learning in areas of speech recognition, Mocialov et al. [76] introduced transfer learning into the low-resource task of SLT. They designed two transfer learning techniques for language modeling and used the large corpus Penn Treebank to import the English language knowledge into stacked LSTM models. Chen et al. [68] proposed a transfer learning strategy for SLT. Their model was a progressive pretraining approach that utilized a large quantity of external data from a general domain to pretrain the model step by step. As opposed to the research that focuses on improving SLR, Cao et al. [24] aimed to improve the translation part in SLT. They proposed a task-aware instruction network (TIN) that leveraged a pretrained model and a large number of unlabeled corpora to enhance translation performance. To reduce the differences between gloss and text, they proposed a data augmentation strategy that performed upsampling at the token level, sentence level, and dataset level. Moryossef et al. [22] focused on the gloss–text task in SLT and proposed two rule-based augmentation strategies to address the problem of scarce resources. They proposed general rules and language-specific rules to generate pseudo-parallel gloss–text pairs, which were then used for backtranslation to improve the model's performance. Ye et al. [77] proposed a domain text generation model for the gloss–text translation task, which could generate large-scale spoken language text for backtranslation (BT). To settle the problem of data scarcity and the modality gap between sign video and text, Zhang et al. [78] proposed a novel framework for the task of SLT. To obtain more data resources, they paid attention to the task of machine translation. Similarly, Ye et al. [79] proposed a cross-modality data augmentation framework to settle the problem of the modality gap between sign and text. Table 3 shows the performance of some of the celebrated models mentioned above.

4. Datasets

In this section, we introduce the datasets employed for SLT and SLR. Firstly, the dataset employed for the task of SLT contains sign language videos, sign language gloss, and spoken language text. Secondly, the dataset employed for the subtask of SLT (gloss2text) contains spoken language text and sign language gloss. Thirdly, the dataset employed for SLR contains sign language videos and sign language gloss. Additionally, we introduce multilanguage SLT datasets that contain various sign languages and spoken languages. Table 4 shows the dataset of some of the celebrated models mentioned above.

Currently, RWTH-PHOENIX-Weather-2014T [9] is the most widely used dataset compared by most of baseline models in SLT. PHOENIX14T consists of German sign language videos, sign language gloss, and spoken language text. PHOENIX14T is segmented into parallel sentences, where each German video (consisting of multiple sign language frames) corresponds to multiple gloss and German sentences. The sign language videos of PHOENIX14T were collected from nine different signers, the size of the gloss vocabulary is 1066, and the size of the vocabulary of the German spoken language text is 2887. The visuals for PHOENIX14T were sourced from German weather news, while the German sign language annotations were provided by deaf specialists and the German text was sourced from news speakers.

CSL-Daily [19] is a celebrated sign language dataset for Chinese SLT. It covers various themes such as family life, school life, medical care, etc. The CSL-Daily dataset includes sign language videos from 10 native signers whose signing expressions are both normative and natural. The sign gloss vocabulary in this dataset consists of 2000 words, and the spoken Chinese text vocabulary consists of 2343 words, which were created under the guidance of sign language linguistics experts and sign language teachers.

RWTH-PHOENIX-Weather 2014 (PHOENIX14) [80] is a celebrated sign language dataset for SLT. The sign language videos of PHOENIX14 are from weather news programs, which were collected from nine signers with a gloss vocabulary of 1081.

ASLG-PC12 [81] is a US sign language dataset created semi-automatically using rule-based methods, and the dataset does not include sign language videos. ASLG-PC12 contains a massive number of gloss–text pairs, and it can be employed for the task of G2T. The size of the vocabulary of spoken language text and American sign gloss are 21,600 and 15,782, respectively.

Spreadthesign-Ten (SP-10) [62] is a multilingual sign language dataset for SLR. The sign language videos and corresponding texts of SP-10 were collected from 10 different languages. Each data point of SP-10 includes 10 sign language videos and 10 spoken language translation texts.

Table 4. The performance of some celebrated models on the PHOENIX14T dataset.

Model	**Dev**						**Test**					
	BL-1	**BL-2**	**BL-3**	**BL-4**	**ROUGE**	**METEOR3**	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**ROUGE**	**METEOR3**
Camgoz et al. [46]	44.4	31.83	24.61	20.16	46.02	-	44.13	31.47	23.89	19.26	45.45	-
Camgoz et al. [16]	50.69	38.16.	30.53	25.35	-	-	48.9	36.88	29.45	24.54	-	-
Transformer [58]	-	-	-	-	-	-	47.69	35.52	28.1	23.32	46.58	44.85
Transformer Ens. [58]	-	-	-	-	-	-	48.4	36.9	29.7	24.9	48.51	46.24
Transformer [17]	49.05	36.2.	28.53	23.52	47.36	46.09	47.69	35.52	28.17	23.32	46.58	44.85
Transformer Ens. [17]	48.85.	36.62	29.23	24.38	49.01	46.96	48.4	36.9	29.7	24.9	48.51	46.24
mBART w/ CC25 [68]	54.01	41.41	33.5.	28.19	53.79	-	52.65	39.99	32.07	26.7	52.54	-
TIN-SLT [24]	52.35.	39.03.	30.83	25.38	48.82	48.4	52.77	40.08	32.09	26.55	49.43	49.36
BT-tuned [22]	-	-	-	-	-	-	-	-	-	22.02	-	-
General-tuned [22]	-	-	-	-	-	-	-	-	-	23.35	-	-
Specific-tuned [22]	-	-	-	-	-	-	-	-	-	23.17	-	-
Transformer+Scaling BT [77]	48.68	37.94.	30.58	25.56	-	-	47.7	37.09	29.92	25.04	-	-

5. Metrics

SLT has different evaluation metrics for different subtasks but can generally be divided into SLR and SLT. The evaluation metrics for SLR include WER (word error rate) [82], Acc, and recall, while the evaluation metrics for SLT mainly include BLEU, ROUGE [83], METEOR [84], WER, CIDEr [85], PER [86], TER [87], NIST [88], among others.

The WER is an evaluation metric from NLP, which refers to the minimum total sum of recognized gloss sequences to the reference sequence, including substitutions, insertions, and deletions.

BLEU is a celebrated evaluation metric employed for SLT, and it assesses the performance of each model through the similarity of generated text and a reference translation. The BLEU score is calculated by computing the frequency of shared n-grams [89] between predicted text and reference sentence, with scores ranging from zero to one to indicate the degree of similarity. The BLEU score is one if the predicted text and the reference sentence are completely similar. Depending on the length of n-grams, BLEU is divided into BLEU-1, BLEU-2, BLEU-3, and BLEU-4.

The ROUGE index evaluates the translation performance by measuring the matching degree between the predicted and reference results, and differs from BLEU by emphasizing the index of recall rather than the index of precision for evaluating the quality of translation.

METEOR uses WordNet [90] and combines multiple evaluation metrics such as word matching, word-order matching, synonymy matching, stemming matching, and noun-phrase matching. Table 5 shows the metrics of some of the celebrated models mentioned above.

Table 5. The dataset employed in SLR and SLT.

Dataset	Language	Video	Gloss	Text	Signs	Running Glosses	Signers	Duration (h)
RWTH-Phoenix-Weather [91]	DGS	✓	✓	✓	911	21,822	7	3.25
BSL [92]	BSL	✓	✓	✓	5k	-	249	-
S-pot [93]	Suvi	✓	✓	✓	1211	-	5	-
RWTH-Phoenix-Weather-2014 [80]	DGS	✓	✓	✓	1081	65,227	9	-
DGS Korpus [94]	DGS	✓	✓	✓	-	-	330	-
GSL [95]	GSL	✓	✓	✓	310	-	7	-
RWTH-Phoenix-2014T [9]	DGS	✓	✓	✓	1066	67,781	9	11
How2Sign [96]	ASL	✓	✓	✓	16,000	-	11	79
CSL-Daily [19]	CSL	✓	✓	✓	2000	20,654	10	20.62
SIGNUM [97]	DGS	✓	✓	✗	455	-	25	55
RWTH-BOSTON-104 [98]	ASL	✓	✓	✗	104	-	3	0.145
Devisign-G [99]	CSL	✓	✓	✗	36	-	8	-
USTC CSL [44]	CSL	✓	✓	✗	178	-	50	100
WLASL [100]	ASL	✓	✓	✗	2000	-	119	-
ASLG-PC12 [81]	ASL	✗	✓	✓	-	-		-

6. Conclusions

SLT is a typical multimodal cross-disciplinary task, which plays an important role in promoting communication in the deaf community and between non-DHH and DHH individuals. At present, SLT faces various challenges, such as the scarcity of dataset resources and the difficulty of obtaining high-quality SLT resources. In this work, we combined the background of SLT to classify the task of SLT and describe the pipeline of SLT. We extracted the typical framework of SLT and analyzed it with specific examples. For the latest literature and methods of SLT, we classified and summarized them into three types: improving the accuracy of SLR, improving the performance of SLT through different models, and addressing the scarcity of data resources. In addition, we introduced the most commonly used datasets and evaluation metrics employed for the task of SLT.

Currently, there are three feasible directions for improving the performance of SLT: firstly, fine-tuning the model obtained from a deep neural network; secondly, introducing data resources from other fields, such as gesture recognition and machine translation; finally, using generative models to produce the required SLT data resources. However, the ultimate step for improving translation performance through these methods is to make targeted innovations suitable for the domain of SLT, which may be the most important aspect of performance improvement.

Author Contributions: Writing—original draft preparation, H.L. and Z.L.; writing—review and editing, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fundamental Research Funds for the Central Universities (CUC220F002), and the National Key R&D Program of China (2022YFC3302100).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, H.; Zhou, W.; Zhou, Y.; Li, H. Spatial-temporal multi-cue network for continuous sign language recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020 ; Volume 34, pp. 13009–13016.
2. Rodriguez, J.; Martínez, F. How important is motion in sign language translation? *IET Comput. Vis.* **2021**, *15*, 224–234. [CrossRef]
3. Zheng, J.; Chen, Y.; Wu, C.; Shi, X.; Kamal, S.M. Enhancing neural sign language translation by highlighting the facial expression information. *Neurocomputing* **2021**, *464*, 462–472. [CrossRef]
4. Li, D.; Xu, C.; Yu, X.; Zhang, K.; Swift, B.; Suominen, H.; Li, H. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12034–12045.
5. Núñez-Marcos, A.; Perez-de Viñaspre, O.; Labaka, G. A survey on Sign Language machine translation. *Expert Syst. Appl.* **2022**, *213*, 118993. [CrossRef]
6. Cui, R.; Liu, H.; Zhang, C. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7361–7369.
7. Cihan Camgoz, N.; Hadfield, S.; Koller, O.; Bowden, R. Subunets: End-to-end hand shape and continuous sign language recognition. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3056–3065.
8. Koller, O.; Zargaran, S.; Ney, H.; Bowden, R. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *Int. J. Comput. Vis.* **2018**, *126*, 1311–1325. [CrossRef]
9. Camgoz, N.C.; Hadfield, S.; Koller, O.; Ney, H.; Bowden, R. Neural sign language translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7784–7793.
10. Ananthanarayana, T.; Srivastava, P.; Chintha, A.; Santha, A.; Landy, B.; Panaro, J.; Webster, A.; Kotecha, N.; Sah, S.; Sarchet, T.; et al. Deep learning methods for sign language translation. *ACM Trans. Access. Comput. (TACCESS)* **2021**, *14*, 1–30. [CrossRef]
11. Zhang, J.; Zhou, W.; Xie, C.; Pu, J.; Li, H. Chinese sign language recognition with adaptive HMM. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
12. Hu, H.; Zhou, W.; Li, H. Hand-model-aware sign language recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 1558–1566.
13. Wu, D.; Pigou, L.; Kindermans, P.J.; Le, N.D.H.; Shao, L.; Dambre, J.; Odobez, J.M. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1583–1597. [CrossRef]

14. Koller, O.; Ney, H.; Bowden, R. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3793–3802.
15. Yin, Q.; Tao, W.; Liu, X.; Hong, Y. Neural Sign Language Translation with SF-Transformer. In Proceedings of the 2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI), Guangzhou, China, 4–6 March 2022; pp. 64–68.
16. Camgoz, N.C.; Koller, O.; Hadfield, S.; Bowden, R. 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, Seattle, WA, USA, 13–19 June 2020; pp. 10023–10033.
17. Yin, K.; Read, J. Better sign language translation with STMC-transformer. *arXiv* **2020**, arXiv:2004.00588.
18. Kumar, S.S.; Wangyal, T.; Saboo, V.; Srinath, R. Time series neural networks for real time sign language translation. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 243–248.
19. Zhou, H.; Zhou, W.; Qi, W.; Pu, J.; Li, H. Improving sign language translation with monolingual data by sign back-translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1316–1325.
20. Zhou, H.; Zhou, W.; Zhou, Y.; Li, H. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Trans. Multimed.* **2021**, *24*, 768–779. [CrossRef]
21. De Coster, M.; D'Oosterlinck, K.; Pizurica, M.; Rabaey, P.; Verlinden, S.; Van Herreweghe, M.; Dambre, J. Frozen pretrained transformers for neural sign language translation. In Proceedings of the 18th Biennial Machine Translation Summit (MT Summit 2021), Virtual, 16–20 August 2021; pp. 88–97.
22. Moryossef, A.; Yin, K.; Neubig, G.; Goldberg, Y. Data augmentation for sign language gloss translation. *arXiv* **2021**, arXiv:2105.07476.
23. Zhang, X.; Duh, K. Approaching sign language gloss translation as a low-resource machine translation task. In Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL), Virtual, 20 August 2021; pp. 60–70.
24. Cao, Y.; Li, W.; Li, X.; Chen, M.; Chen, G.; Hu, L.; Li, Z.; Kai, H. Explore more guidance: A task-aware instruction network for sign language translation enhanced with data augmentation. *arXiv* **2022**, arXiv:2204.05953.
25. Minu, R. A Extensive Survey on Sign Language Recognition Methods. In Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 23–25 February 2023; pp. 613–619.
26. Baumgärtner, L.; Jauss, S.; Maucher, J.; Zimmermann, G. Automated Sign Language Translation: The Role of Artificial Intelligence Now and in the Future. In Proceedings of the CHIRA, Virtual Event, 5–6 November 2020; pp. 170–177.
27. Koller, O. Quantitative survey of the state of the art in sign language recognition. *arXiv* **2020**, arXiv:2008.09918.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
30. Liang, Z.; Du, J. Sequence to sequence learning for joint extraction of entities and relations. *Neurocomputing* **2022**, *501*, 480–488. [CrossRef]
31. Liang, Z.; Du, J.; Shao, Y.; Ji, H. Gated graph neural attention networks for abstractive summarization. *Neurocomputing* **2021**, *431*, 128–136. [CrossRef]
32. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
33. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
35. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 305–321.
36. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef]
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
38. He, S. Research of a sign language translation system based on deep learning. In Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Dublin, Ireland, 17–19 October 2019; pp. 392–396.
39. Guo, D.; Zhou, W.; Li, A.; Li, H.; Wang, M. Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation. *IEEE Trans. Image Process.* **2019**, *29*, 1575–1590. [CrossRef]
40. Gan, S.; Yin, Y.; Jiang, Z.; Xie, L.; Lu, S. Skeleton-aware neural sign language translation. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4353–4361.

41. Kim, S.; Kim, C.J.; Park, H.M.; Jeong, Y.; Jang, J.Y.; Jung, H. Robust keypoint normalization method for Korean sign language translation using transformer. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 21–23 October 2020; pp. 1303–1305.
42. Ko, S.K.; Kim, C.J.; Jung, H.; Cho, C. Neural sign language translation based on human keypoint estimation. *Appl. Sci.* **2019**, *9*, 2683. [CrossRef]
43. Guo, D.; Zhou, W.; Li, H.; Wang, M. Hierarchical LSTM for sign language translation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018; Volume 32.
44. Huang, J.; Zhou, W.; Zhang, Q.; Li, H.; Li, W. Video-based sign language recognition without temporal segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018; Volume 32.
45. Parton, B.S. Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence. *J. Deaf Stud. Deaf Educ.* **2006**, *11*, 94–101. [CrossRef] [PubMed]
46. Camgoz, N.C.; Koller, O.; Hadfield, S.; Bowden, R. Multi-channel transformers for multi-articulatory sign language translation. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part IV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 301–319.
47. Kan, J.; Hu, K.; Hagenbuchner, M.; Tsoi, A.C.; Bennamoun, M.; Wang, Z. Sign language translation with hierarchical spatio-temporal graph neural network. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 3367–3376.
48. Wang, S.; Guo, D.; Zhou, W.G.; Zha, Z.J.; Wang, M. Connectionist temporal fusion for sign language translation. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 26 October 2018; pp. 1483–1491.
49. Guo, D.; Wang, S.; Tian, Q.; Wang, M. Dense Temporal Convolution Network for Sign Language Translation. In Proceedings of the IJCAI, Macao, China, 10–16 August 2019; pp. 744–750.
50. Song, P.; Guo, D.; Xin, H.; Wang, M. Parallel temporal encoder for sign language translation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1915–1919.
51. Fang, B.; Co, J.; Zhang, M. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems, Delft, The Netherlands, 6–8 November 2017; pp. 1–13.
52. Arvanitis, N.; Constantinopoulos, C.; Kosmopoulos, D. Translation of sign language glosses to text using sequence-to-sequence attention models. In Proceedings of the 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Sorrento, Italy, 26–29 November 2019; pp. 296–302.
53. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
54. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
55. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *arXiv* **2014**, arXiv:1409.3215.
56. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165.
57. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
58. Yin, K.; Read, J. Attention Is All You Sign: Sign Language Translation with Transformers. Available online: https://www.slrtp.com/papers/extended_abstracts/SLRTP.EA.12.009.paper.pdf (accessed on 12 May 2023).
59. Zheng, J.; Zhao, Z.; Chen, M.; Chen, J.; Wu, C.; Chen, Y.; Shi, X.; Tong, Y. An improved sign language translation model with explainable adaptations for processing long sign sentences. *Comput. Intell. Neurosci.* **2020**, *2020*, 8816125. [CrossRef] [PubMed]
60. Voskou, A.; Panousis, K.P.; Kosmopoulos, D.; Metaxas, D.N.; Chatzis, S. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11946–11955.
61. Qin, W.; Mei, X.; Chen, Y.; Zhang, Q.; Yao, Y.; Hu, S. Sign language recognition and translation method based on VTN. In Proceedings of the 2021 International Conference on Digital Society and Intelligent Systems (DSInS), Chengdu, China, 19–21 November 2021; pp. 111–115.
62. Yin, A.; Zhao, Z.; Jin, W.; Zhang, M.; Zeng, X.; He, X. MLSLT: Towards Multilingual Sign Language Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5109–5119.
63. Guo, Z.; Hou, Y.; Hou, C.; Yin, W. Locality-Aware Transformer for Video-Based Sign Language Translation. *IEEE Signal Process. Lett.* **2023**, *30*, 364–368. [CrossRef]
64. Orbay, A.; Akarun, L. Neural sign language translation by learning tokenization. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 222–228.
65. De Coster, M.; Dambre, J. Leveraging frozen pretrained written language models for neural sign language translation. *Information* **2022**, *13*, 220. [CrossRef]

66. Zhao, J.; Qi, W.; Zhou, W.; Duan, N.; Zhou, M.; Li, H. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Trans. Multimed.* **2021**, *24*, 2662–2672. [CrossRef]
67. Fu, B.; Ye, P.; Zhang, L.; Yu, P.; Hu, C.; Chen, Y.; Shi, X. ConSLT: A token-level contrastive framework for sign language translation. *arXiv* **2022**, arXiv:2204.04916.
68. Chen, Y.; Wei, F.; Sun, X.; Wu, Z.; Lin, S. A simple multi-modality transfer learning baseline for sign language translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5120–5130.
69. Barrault, L.; Bojar, O.; Costa-Jussa, M.R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; et al. Findings of the 2019 Conference on Machine Translation (WMT19). In Proceedings of the Fourth Conference on Machine Translation, Florence, Italy, 1–2 August 2019.
70. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
71. Moe, S.Z.; Thu, Y.K.; Thant, H.A.; Min, N.W.; Supnithi, T. Unsupervised Neural Machine Translation between Myanmar Sign Language and Myanmar Language. Ph.D. Thesis, MERAL Portal, Mandalay, Myanmar, 2020.
72. Albanie, S.; Varol, G.; Momeni, L.; Afouras, T.; Chung, J.S.; Fox, N.; Zisserman, A. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 35–53.
73. Sennrich, R.; Haddow, B.; Birch, A. Improving neural machine translation models with monolingual data. *arXiv* **2015**, arXiv:1511.06709.
74. Nunnari, F.; España-Bonet, C.; Avramidis, E. A data augmentation approach for sign-language-to-text translation in-the-wild. In Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021), Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Zaragoza, Spain, 1–3 September 2021.
75. Gómez, S.E.; McGill, E.; Saggion, H. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021), Online, 6 September 2021; pp. 18–27.
76. Mocialov, B.; Turner, G.; Hastie, H. Transfer learning for british sign language modelling. *arXiv* **2020**, arXiv:2006.02144.
77. Ye, J.; Jiao, W.; Wang, X.; Tu, Z. Scaling Back-Translation with Domain Text Generation for Sign Language Gloss Translation. *arXiv* **2022**, arXiv:2210.07054.
78. Zhang, B.; Müller, M.; Sennrich, R. SLTUNET: A simple unified model for sign language translation. *arXiv* **2023**, arXiv:2305.01778.
79. Ye, J.; Jiao, W.; Wang, X.; Tu, Z.; Xiong, H. Cross-modality Data Augmentation for End-to-End Sign Language Translation. *arXiv* **2023**, arXiv:2305.11096.
80. Koller, O.; Forster, J.; Ney, H. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Comput. Vis. Image Underst.* **2015**, *141*, 108–125. [CrossRef]
81. Othman, A.; Jemni, M. English-asl gloss parallel corpus 2012: Aslg-pc12. In *Sign-lang@ LREC 2012*; European Language Resources Association (ELRA): Paris, France, 2012; pp. 151–154.
82. Su, K.Y.; Wu, M.W.; Chang, J.S. A new quantitative quality measure for machine translation systems. In Proceedings of the COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992.
83. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Toronto, ON, Canada, 2004; pp. 74–81.
84. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; Association for Computational Linguistics: Toronto, ON, Canada, 2005; pp. 65–72.
85. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
86. Tillmann, C.; Vogel, S.; Ney, H.; Zubiaga, A.; Sawaf, H. Accelerated DP based search for statistical translation. In Proceedings of the Eurospeech, Rhodes, Greece, 22–25 September 1997.
87. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
88. Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, USA, 24–27 March 2002; pp. 138–145.
89. Shannon, C.E. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **2001**, *5*, 3–55. [CrossRef]
90. Oram, P. WordNet: An electronic lexical database. Christiane Fellbaum (Ed.). Cambridge, MA: MIT Press, 1998. Pp. 423. *Appl. Psycholinguist.* **2001**, *22*, 131–134. [CrossRef]
91. Forster, J.; Schmidt, C.; Hoyoux, T.; Koller, O.; Zelle, U.; Piater, J.H.; Ney, H. RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In Proceedings of the LREC, Istanbul, Turkey, 23–25 May 2012; Volume 9, pp. 3785–3789.

92. Schembri, A.; Fenlon, J.; Rentelis, R.; Reynolds, S.; Cormier, K. Building the British sign language corpus. *Lang. Doc. Conserv.* **2013**, *7*, 136–154.
93. Viitaniemi, V.; Jantunen, T.; Savolainen, L.; Karppa, M.; Laaksonen, J. S-pot–a benchmark in spotting signs within continuous signing. In Proceedings of the LREC Proceedings, Reykjavik, Iceland, 26–31 May 2014; European Language Resources Association (LREC): Paris, France, 2014.
94. Hanke, T.; Schulder, M.; Konrad, R.; Jahn, E. Extending the Public DGS Corpus in size and depth. In Proceedings of the Sign-Lang@ LREC, Marseille, France, 11–16 May 2020; European Language Resources Association (ELRA): Paris, France, 2020; pp. 75–82.
95. Adaloglou, N.; Chatzis, T.; Papastratis, I.; Stergioulas, A.; Papadopoulos, G.T.; Zacharopoulou, V.; Xydopoulos, G.J.; Atzakas, K.; Papazachariou, D.; Daras, P. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Trans. Multimed.* **2021**, *24*, 1750–1762. [CrossRef]
96. Duarte, A.; Palaskar, S.; Ventura, L.; Ghadiyaram, D.; DeHaan, K.; Metze, F.; Torres, J.; Giro-i Nieto, X. How2sign: A large-scale multimodal dataset for continuous american sign language. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2735–2744.
97. Von Agris, U.; Kraiss, K.F. Towards a video corpus for signer-independent continuous sign language recognition. *Gesture Hum. Comput. Interact. Simul. Lisbon Port. May* **2007**, *11*, 2.
98. Dreuw, P.; Neidle, C.; Athitsos, V.; Sclaroff, S.; Ney, H. Benchmark Databases for Video-Based Automatic Sign Language Recognition. In Proceedings of the LREC, Marrakech, Morocco, 26 May–1 June 2008.
99. Chai, X.; Wang, H.; Chen, X. The devisign large vocabulary of chinese sign language database and baseline evaluations. In *Technical Report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)*; Institute of Computing Technology: Beijing, China, 2014.
100. Li, D.; Rodriguez, C.; Yu, X.; Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1459–1469.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.