Research Statement

Ronglai Zuo
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
rzuo@cse.ust.hk

1. Introduction

According to the statistics of World Health Organization (WHO), over 5% of the world population (about 430 million people) are suffering from hearing loss. By 2050, the number would increase to 700 million. Sign languages, also known as signed languages, are the primary communication method among the deaf and hard-of-hearing people. As a kind of visual language, sign languages use both manual (*e.g.*, hand movement) and non-manual (*e.g.*, expression and mouthing) parameters to convey information. Besides, sign languages also have unique grammatical rules and vocabulary which are usually different with their spoken language counterparts (*e.g.*, American sign language *vs.* English, Chinese sign language *vs.* Chinese). These characteristics of sign languages result in a **two-way** communication gap between the deaf and hearing. Over the past decade, we have witnessed the technical explosion in computer vision and natural language processing, however, few attention has been paid on the deaf community. During my PhD study, I devote myself to narrowing this two-way communication gap by developing intelligent systems for both sign language understanding [1–7] and generation [8].

There are several research directions in sign language understanding, such as sign language recognition (SLR), sign language translation (SLT), sign spotting [9], and sign retrieval [10]. My works on sign language understanding focus on two primary directions: SLR and SLT. The objective of SLR is to transcribe a sign video into its constituent glosses[1]. According to the number of glosses in a sign sentence, SLR can be further categorized into isolated SLR (ISLR), in which each sign sentence only consists of a single gloss, and continuous SLR (CSLR), in which each sign sentence may consist of multiple glosses. Taking a step further, SLT aims at translating sign languages into spoken languages. Studies on sign language understanding can help the hearing understand the information conveyed by sign languages. As a reversed process, sign language generation (SLG) aims at translating spoken languages into sign languages, allowing two-way communication between the deaf and hearing together with SLT. The relationship of these research directions is illustrated in Figure 1. Below I will first introduce more details on my research experience in Section 2 and then discuss several future research directions in Section 3.

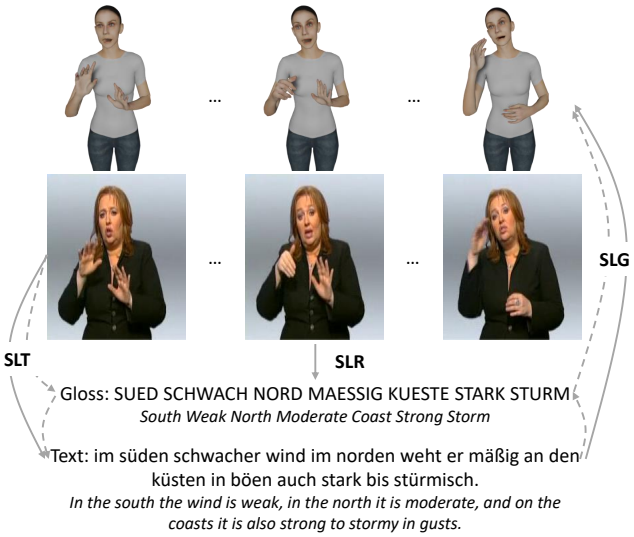

Figure 1. Relationship of sign language recognition (SLR), translation (SLT), and generation (SLG). There are various output formats of SLG systems such as keypoints, video frames, and avatars. Here we use 3D avatars to represent SLG outputs as done in our work [8]. Dashed lines denote using glosses as intermediate representations. The sample is from Phoenix-2014T [11], and we translate German into English.

2. Research Experience

My research starts from developing well-performed sign language understanding systems, with a focus on backbone network design and auxiliary training techniques [1–5]. After that, we move to a more realistic setting: online sign language recognition and translation [6]. Finally, we build

[1]A gloss is the written form of a sign, which is usually represented by a word or phrase.

a simple yet effective baseline for sign language generation with 3D avatars to complete the two-way communication loop [8].

2.1. Backbone Networks and Training Techniques

Two-Stream Network with Keypoint Modeling. Most existing SLR and SLT works explore to directly model RGB videos to understand sign languages. However, RGB videos are raw signals with substantial visual redundancy such as background and signer appearance, which may lead the model to overlook key information for sign language understanding. In [3], we propose to involve keypoints in sign language modeling, and introduce a two-stream network, TwoStream-SLR, to jointly model RGB videos and keypoint sequences for SLR. Furthermore, TwoStream-SLR can be easily extended to handle SLT by appending an additional translation network [12].

We represent keypoints as a sequence of heatmaps such that each stream can share the identical architecture (S3D [13]) without any ad-hoc design for the keypoint stream. A series of techniques are also proposed to better train the two-stream network: 1) Bidirectional lateral connections are introduced for inter-stream information exchange; 2) Sign pyramid networks improve the model robustness on sign duration variation; 3) Auxiliary CTC losses are added to enable shallow layers to learn meaningful features; 4) Frame-level self-distillation provides fine-grained supervision signals. Experimental results show that our TwoStream-SLR/SLT outperforms previous works by a large margin.

ISLR and Natural Language Priors. We move back to the fundamental task of sign language understanding: isolated sign language recognition (ISLR), aiming at recognizing the only gloss for a sign video. In [2], we first identify the existence of visually indistinguishable signs (VISigns) in sign languages and further categorize them into: VISigns with similar semantic meanings (*e.g.*, "Cold" and "Winter") and VISigns with distinct semantic meanings (*e.g.*, "Table" and "Afternoon"). For the former, we propose language-aware label smoothing, an improvement over vanilla label smoothing considering the semantic similarities among glosses. For the latter, we introduce inter-modality mixup, which leverages gloss embeddings to drive the model towards maximizing signs' separability in the feature space. Sufficient ablation studies verify that each technique can significantly boost the model performance over the corresponding VISigns. Utilizing an improved version of the two-stream network as the model backbone, the overall model achieves a new state-of-the-art performance on a series of popular ISLR benchmarks.

2.2. From ISLR to Online CSLR and SLT

As stated above, numerous CSLR works utilize the well-established CTC loss for model training. In the inference stage, these models typically process the entire sign video as input to make predictions. This inference scheme is called offline recognition. In contrast to mature speech recognition systems, which efficiently recognize spoken words on the fly, the field of CSLR still lags behind due to the absence of practical solutions for online recognition. Although CTC-based approaches can be adapted for online CSLR using a sliding window technique, we empirically find the discrepancy between training (entire and untrimmed sign videos) and inference (trimmed and short sign clips) lead to sub-optimal performance.

To mitigate the discrepancy problem, instead of training offline CSLR models with the CTC loss as the usual practice, we introduce an innovative online CSLR framework that slides an ISLR model over a sign video stream [6]. The ISLR model is trained using classification losses on a sign dictionary. We also propose a series of techniques such as sign augmentation and saliency loss to improve model training. Once the ISLR model is well optimized, online inference is achieved in a sliding-window manner. Finally, a rule-based post-processing algorithm is developed to clean predictions.

Experiments on three popular benchmarks show that our online solution can significantly outperforms the adaptation of CTC-based methods. Furthermore, our online framework can be extended to boost offline CSLR models and to support online SLT by appending a gloss-to-text network.

2.3. Sign Language Generation with 3D Avatars

Research on sign language understanding can help the hearing to understand the information conveyed by sign languages. As a reversed process, sign language generation (SLG, also known as sign language production or spoken-to-sign language translation) completes the communication loop between the deaf and hearing. Most existing SLG works are *incomplete* that only keypoints are generated, posing an understandable challenge for the deaf [14]. Although several studies further use the keypoints to animate a signer image and subsequently generate videos, the 2D video format is prone to blurriness and visual distortions.

In [8], we present a simple yet effective SLG baseline whose outcomes are represented by 3D avatars. The baseline consists of three steps: 1) dictionary construction; 2) 3D sign estimation with a novel SMPL-X [15] fitting method, SMPLSign-X; 3) fulfill SLG in a retrieve-then-connect paradigm. The dedicated SMPLSign-X considers unique characteristics of sign languages, *e.g.*, upright upper body during signing and temporal consistency between frames, leading to superior estimation quality than the vanilla SMPLify-X [15]. The overall baseline sets a new state-of-the-art performance in back-translation evaluation. Besides, we also show that by-products of the 3D signs can boost keypoint-based sign language understanding systems.

3. Future Research

As described above, my research focuses on sign language understanding and generation. However, there is still a long way to go before developing a commercial-grade two-way communication system between the deaf and hearing. Below are several meaningful research directions in which I plan to invest my efforts.

3.1. Better Understanding and Generation Systems

Large-Scale Sign Language Datasets with High-Quality Dictionaries. Current sign language understanding systems usually suffer from data scarcity. Popular sign language datasets, *e.g.*, Phoenix-2014T [11] and CSL-Daily [16], only record tens of hours of video data, which are much smaller than widely-adopted datasets in the field of action recognition [4]. Although the auxiliary training techniques can somehow relieve the issue, large-scale sign language datasets are still essential to develop data-driven approaches. During the collection process, both automatic annotation pipelines [17, 18] and manual checking [7] should be considered. Besides, high-quality sign dictionaries may also become an important property of future sign language datasets. Our works [6, 8] have shown the value of sign dictionaries for online SLR and SLG, and thus improving the quality of the dictionaries is a straightforward way to boost the performance of these methods.

Lightweight Sign Encoders. Although our TwoStream Network [3] shows superior performance in both SLR and SLT, the two-stream architecture is inherently heavy, hindering its applications in real practice. Representing keypoints as heatmaps can avoid ad-hoc network design, but also leads to additional computational costs. More efficient keypoint representations such as skeleton graphs [19] and keypoint-only sign encoders [20] should be paid with more attention.

Gloss-Free Understanding and Generation Systems. Gloss annotations usually need efforts of sign language experts, rendering the annotation process expensive. Gloss-free sign language understanding and generation systems require the models to directly output spoken or sign languages without using glosses as intermediate representations. From my perspective, large-scale datasets are the key to develop gloss-free systems. Overlooking gloss annotations enables researchers to easily collect sign video-text paired data from Internet [18] or TV programs [17]. I believe that a model trained over large-scale paired data is capable to learn the alignments between videos and texts.

3.2. Including Sign Languages in LVLMs

In the past year, large language models (LLMs) have gained a lot of attention in the AI community. Several recent works [21–23] further extend LLMs to understand vision modalities, *i.e.*, images or videos. These extended LLMs are referred to as large vision-language models (LVLMs). However, existing LVLMs primarily focus on natural images and videos, overlooking the need of large models for sign language understanding and generation. In contrast to common videos in the natural world, sign videos are inherently *information-intensive* due to their linguistic nature. Thus, I believe it is feasible to include sign languages in LVLMs, and I think building a successful sign LVLM should consider the following aspects:

Visual Encoder Improvement. Current LVLMs usually use an off-the-shelf visual encoder, *e.g.*, CLIP [24] and LanguageBind [25], to tokenize vision inputs. However, these visual encoders are also pretrained on real-world data, lacking a focus on sign languages. To better exploit the spatial-temporal information in sign videos, it is essential to fine-tune the visual encoder, and there are several widely-adopted fine-tuning techniques [26, 27] that can be employed. Besides, similar to tokenizing a word into sub-words in machine translation, a sign can also be split into smaller action units based on universal sign language notation systems, *e.g.*, HamNoSys [28] and SignWriting [29]. Nevertheless, widely-adopted visual encoders are mostly based on ViT [30], which simply splits vision inputs into fixed-size chunks. This operation neglects the linguistic nature of sign languages, highlighting a room for improvement in future sign LVLMs.

Comparable Performance on Traditional Tasks. Although fine-tuning a domain-specific (sign-language-specific) LVLM is straightforward, a more fascinating direction is to endow current LVLMs with the capability of sign language understanding and maintain their performance on traditional tasks, *e.g.*, visual question answering. This requirement involves the concept of catastrophic forgetting [31], and techniques such as continual learning [27] and parameter-efficient fine-tuning [26] can be studied at the beginning.

Two-Way Communication. The communication gap between the deaf and hearing are bidirectional. It is also a challenging topic to enable future sign LVLMs to generate signs. A simple solution based on our SLG baseline [8] is to retrieve the 3D sign dictionary and stitch adjacent signs together. More advanced techniques such as text-conditioned motion generation [32] and auto-regressive video generation [33] can also be considered.

3.3. Beyond Sign Languages

As a hand-centric visual language, research on sign languages helps me gain a lot of experience in hand modeling and multi-modality learning. Below are two challenging but intriguing directions beyond sign languages.

General Vision-Language Models. Existing LVLMs have achieved promising results on a series of benchmarks. However, there are still a majority of down-stream tasks, *e.g.*, medical image processing, autonomous driving, and sign lan-

guage understanding, remaining under-explored. As stated in Section 3.2, better fine-tuning techniques that can avoid forgetting deserve more attention. Furthermore, how to build a universal model for both generation and understanding is also an interesting topic. DreamTeacher [34] reveals that a well-trained generative model can serve as distillation targets for an image backbone, suggesting implicit connections between generative and understanding models. Another exciting work, VideoPoet [33], demonstrates state-of-the-art capabilities in video generation using LLM-based frameworks. I am convinced that auto-regressive image/video generation will be the key to unify the vision and language modalities.

Human-Object Interaction. Building interactive AI assistants in the physical world is always an exciting topic. In real life, hands play a key role in interaction with objects, but how to reconstruct hands and co-articulated objects from an egocentric view poses a new challenge [35, 36]. According to my experience in modeling sign languages, the class information of gestures could provide prior knowledge of hand poses. Due to biological and physical constraints of human body, strange hand poses would never appear. Thus, predicting the gesture/action class in advance may simplify the reconstruction process.

References

[1] **Ronglai Zuo** and Brian Mak. C2SLR: Consistency-enhanced continuous sign language recognition. In *CVPR*, pages 5131–5140, 2022. 1

[2] **Ronglai Zuo**, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *CVPR*, pages 14890–14900, 2023. 2

[3] Yutong Chen*, **Ronglai Zuo***, Fangyun Wei*, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. In *NeurIPS*, 2022, (*equal contribution). 2, 3

[4] **Ronglai Zuo** and Brian Mak. Improving continuous sign language recognition with consistency constraints and signer removal. *ACM TOMM*, 2024. 3

[5] **Ronglai Zuo** and Brian Mak. Local context-aware self-attention for continuous sign language recognition. In *Interspeech*, pages 4810–4814, 2022. 1

[6] **Ronglai Zuo**, Fangyun Wei, and Brian Mak. Towards online sign language recognition and translation. *arXiv, Submitted to ECCV*, 2024. 1, 2, 3

[7] Zhe Niu*, **Ronglai Zuo***, Brian Mak, and Fangyun Wei. A Hong Kong sign language corpus collected from sign-interpreted TV news. In *LREC-COLING*, 2024 (*equal contribution). 1, 3

[8] **Ronglai Zuo***, Fangyun Wei*, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. *arXiv, Submitted to ECCV*, 2024 (*equal contribution). 1, 2, 3

[9] Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. Automatic dense annotation of large-vocabulary sign language videos. In *ECCV*, pages 671–690, 2022. 1

[10] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *CVPR*, 2023. 1

[11] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, 2018. 1, 3

[12] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742, 2020. 2

[13] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018. 2

[14] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. Sign language production: A review. In *CVPRW*, pages 3451–3461, 2021. 2

[15] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2

[16] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *CVPR*, 2021. 3

[17] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. Bbc-oxford british sign language dataset, 2021. 3

[18] David Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus, 2023. 3

[19] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 3

[20] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *ICCV*, pages 20676–20686, October 2023. 3

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2023.

[23] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2023. 3

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[25] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li,

Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023. 3

[26] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 3

[27] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, pages 7959–7971, 2022. 3

[28] Wikipedia contributors. Hamburg notation system — Wikipedia, the free encyclopedia, 2023. [Online; accessed 12-December-2023]. 3

[29] Wikipedia contributors. Signwriting — Wikipedia, the free encyclopedia, 2023. [Online; accessed 12-December-2023]. 3

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 3

[31] Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models, 2023. 3

[32] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 3

[33] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3, 4

[34] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models. In *ICCV*, pages 16698–16708, 2023. 4

[35] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, pages 12943–12954, 2023. 4

[36] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *ICCV*, pages 20270–20281, 2023. 4