# KSL-Guide: A Large-scale Korean Sign Language Dataset Including Interrogative Sentences for Guiding the Deaf and Hard-of-Hearing

Soomin Ham[1], Kibaek Park[1], YeongJun Jang[1], Youngtaek Oh[1],
Seokmin Yun[2], Sukwon Yoon[2], Chang Jo Kim[3], Han-Mu Park[3], and In So Kweon[1]
[1] Department of Electrical Engineering, KAIST, South Korea
[2] Testworks Inc., South Korea
[3] Korea Electronics Technology Institute (KETI), South Korea

*Abstract*— **Many advancements in computer vision and machine learning have shown potential for significantly improving the lives of people with disabilities. In particular, recent research has demonstrated that deep neural network models could be used to bridge the gap between the deaf who use sign language and hearing people. The major impediment to advancing such models is the lack of high-quality and large-scale training data. Moreover, previously released sign language datasets include few or no interrogative sentences compared to declarative sentences. In this paper, we introduce a new publicly available large-scale Korean Sign Language (KSL) dataset—*KSL-Guide*—that includes both declarative sentences and comparable interrogative sentences, which are required for a model to achieve high performance in real-world *interactive* tasks deployed on service applications. Our dataset contains a total of 121K sign language video samples featuring sentences and words spoken by native KSL speakers with extensive annotations (*e.g.*, gloss, translation, keypoints, and timestamps). We exploit a multi-camera system to produce 3D human pose keypoints as well as 2D keypoints from multi-view RGB. Our experiments quantitatively demonstrate that the inclusion of interrogative sentences in training for sign language recognition and translation tasks greatly improves their performance. Furthermore, we empirically show the qualitative results by developing a prototype application using our dataset, providing an interactive guide service that helps to lower the communication barrier between sign language speakers and hearing people.**

## I. INTRODUCTION

Sign languages are the primary language of more than 450 million people in the world who are deaf or hard of hearing [34]. Contrary to popular belief, sign languages are completely independent languages with their own grammar and set of words. For example, American Sign Language (ASL) and English are totally different, and the same applies to Korean Sign Language (KSL) and Korean. Moreover, sign language is not universal but differs from country to country. Even British and American sign language bears no similarity because sign language is not a motion representation of spoken language. For this reason, communication between a hearing person and a Deaf[1] person is often very difficult, as much as people speaking different languages. Such difficulty has been working as a huge obstacle to the Deaf's access to public information or social engagement. Fortunately, recent advances in computer vision, natural language processing, and robotics have been lowering this language barrier between hearing people and those who are Deaf. Several studies aimed to enable automatic recognition of sign language [11], [21], and more recent work offered direct translation from a sign language to a spoken language [5], [6], [36].

TABLE I: **Examples of interrogative and declarative sentences from our dataset.** Korean Sign Language (KSL) is represented by a gloss sequence, and Korean is the translation of KSL into spoken language. The underline refers to non-manual markers.

| | Interrogative | Declarative |
|---|---|---|
| KSL<br>Kor | q (wh-question)<br>[강남] [가다] [목적] [지하철] [갈아타다] [몇호]<br>GANGNAM GO PURPOSE SUBWAY TRANSFER LINE?<br>강남으로 가려면 몇호선으로 갈아타야 하나요?<br>Which subway line should I transfer to go to Gangnam ? | [2호] [갈아타다] [목적]<br>LINE-2 TRANSFER PURPOSE.<br>2호선으로 환승하려고 합니다.<br>I'd like to transfer to Line 2. |
| KSL<br>Kor | q (yes/no-question)<br>[밤] [12시] [후] [돈] [추가]<br>NIGHT 12-O'_CLOCK AFTER MONEY EXTRA?<br>심야 할증 요금이 적용되나요?<br>Does the late-night surcharge apply? | [백화점] [다음] [곳] [저기] [도착] [내리다]<br>DEPARTMENT_STORE NEXT SPOT THERE ARRIVAL GET_OFF.<br>백화점 지나서 내려주세요.<br>Let me get off past the department store. |
| KSL<br>Kor | q (yes/no-question)<br>[도착] [시간] [딱]<br>ARRIVAL TIME EXACT?<br>제 시간에 도착하나요?<br>Are we arriving on time? | [접근] [도착] [가능]<br>APPROACH ARRIVAL POSSIBLE.<br>다와갑니다.<br>We are almost there. |
| KSL<br>Kor | q (wh-question)<br>[지하철] [가다] [빨리] [방법]<br>SUBWAY GO FAST WAY?<br>어떻게 가는 것이 빠를까요?<br>How can I get there fast? | [시간] [급하다]<br>TIME HURRY.<br>제가 많이 급합니다.<br>I'm in a hurry. |
| KSL<br>Kor | q (yes/no-question)<br>[버스] [곳] [가깝다]<br>BUS PLACE NEAR?<br>근처에 버스정류장이 있나요?<br>Is there a bus stop nearby? | [핸드폰] [실종] [잃어버리다]<br>CELL_PHONE MISSING LOST.<br>휴대폰을 잃어버렸습니다.<br>I lost my cell phone. |
| KSL<br>Kor | q (wh-question)<br>[지하철][정기권] [사다] [무엇] [곳]<br>SUBWAY REGULAR-TICKET BUY WHAT PLACE?<br>정기승차권은 어디에서 사나요?<br>Where can I buy regular tickets? | [지하철] [정기권] [사다] [원하다]<br>SUBWAY REGULAR-TICKET BUY WANT.<br>정기승차권을 사고 싶습니다.<br>I want to buy regular tickets. |

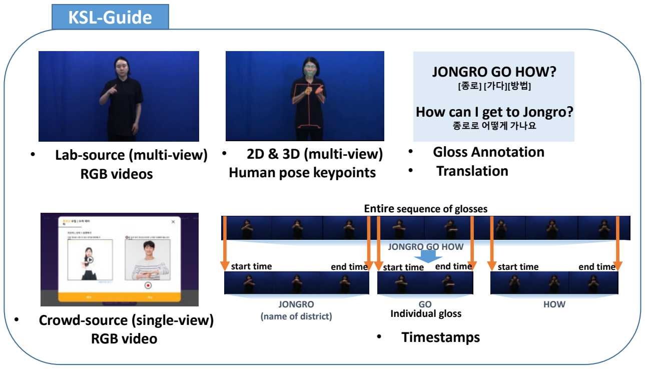


Fig. 1: **Illustration of the KSL-Guide dataset.** We provide RGB videos and various types of ground-truth annotations: gloss annotation, spoken language translation, 2D and 3D human pose keypoints, and timestamps.

The key enablers for such applied technologies are sign language datasets, which are used to train machine learning models. Compared to spoken languages with corresponding written languages, the number of existing datasets on sign

[1] The uppercase Deaf refers to people with difficulty in hearing who primarily communicate in sign language and share its culture, while the lowercase deaf is the term used to refer to people with difficulty in hearing who do not necessarily communicate in sign languages [25].

languages is extremely limited due to a variety of difficulties in generating large-scale, high-quality datasets of sign languages with extensive annotations.

Our work releases the first large-scale publicly available dataset[2] on Korean Sign Language (KSL), which has been recently recognized as *an official language* in the Republic of Korea. This KSL dataset consists of three major sets of content: 2,000 continuous sentences (**KSL-Guide-Sentence**), 3,000 isolated words (**KSL-Guide-Word**), and 1,000 fingerspelled words (**KSL-Guide-Fingerspelling**) performed by 20, 20 and 52 native signers, respectively. Our dataset includes a large number of interrogative sentences (*i.e.*, questions) as well as declarative sentences (*i.e.*, statements), which also distinguishes this dataset from the previous large-scale sign language datasets. The corpus of this dataset mainly deals with finding a place or getting directions and using public transportation. The aim of our work is to provide guide services at a subway station, at the airport, or in a taxi.

Dataset 구성.

Sign languages are made by manual and non-manual signals (or features) simultaneously. Non-manual signals (NMS) refer to movements and positions of the body, head, shoulders, and face, including the mouth, eyes/eyebrows, and facial expressions, that are not performed with the hands or fingers [18]. Non-manual features in sign languages are used to distinguish sentence types, such as **interrogative sentences** (wh-questions, yes-no questions), **declarative sentences** (negative, affirmative, neutral), conditionals, etc. [10], in ways similar to how intonation works in spoken language. This is because the non-manual expressions represent important information (*i.e.*, grammatical markers) that makes up the sentences in sign language, which is a visual language. Therefore, communication becomes possible only when we see the manual and non-manual signals together. In KSL, signers raise eyebrows, open eyes wide, and make eye contact with the other person to indicate interrogative sentences. Table I provides sentence examples from our dataset.

Our dataset exploits a calibrated multi-camera system to capture five multiple RGB views in a lab environment for every recording of sentence and word datasets (*i.e.*, a total of 100K×5 video samples). The multiple views enabled 3D pose reconstruction and we examined the estimation manually by human annotators to correct errors occurring due to the occlusions in the 2D images. Moreover, we utilized a crowdsourcing platform to collect fingerspelled words in the wild (*i.e.*, 21K video samples). We also provide various types of annotations for all the sign language videos, such as gloss annotation, spoken language translation, 2D and 3D human pose keypoints, and timestamps, as shown in Fig. 1. All the RGB videos and annotations were collaboratively produced and examined by native signers, sign language professionals, and interpreters.

Our experiments demonstrated that the inclusion of interrogative sentences in training data greatly improves the performance of sign language recognition and translation tasks. Furthermore, we deployed our dataset and the translation task in a prototype application to provide an interactive guide service that can be used to assist Deaf individuals in navigating a place or using public transportation. We envision that our dataset will become helpful in the community's effort to lower the barrier between the Deaf and the hearing majority.

## II. RELATED WORK

Sign language (SL) is a visual language that can be transcribed into textual form using a written notation called *gloss*. Gloss is not a translation but is the label for a unit of sign that has the closest meaning to the word in the spoken language [26]. A word usually corresponds to a single gloss, and a sentence is represented by *a sequence of glosses*. The actual translation of a sentence is different from the glosses, as is shown in Table I. Using neural networks trained on the appropriate dataset, we can automatically recognize signs and represent them by glosses, as well as translate signs or glosses into spoken language.

### A. Challenging Tasks in Sign Language

**Sign Language Recognition.** Sign language recognition is a task that predicts the corresponding gloss or the sequence of glosses from a given input video. Specifically, it is further divided into three subtasks: static sign language recognition (SSLR), isolated sign language recognition (ISLR), and continuous sign language recognition (CSLR). SSLR is the recognition of static gestures of fingerspelling, and each gesture corresponds to a letter of the alphabet or a numeral. In ISLR, an input video only contains a single sign, and the task is simply to identify the corresponding gloss. On the other hand, in CSLR, an input video includes more than one sign and the goal is to recognize a sequence of glosses in the video. CSLR is substantially more challenging than ISLR because it additionally requires the model to segment each sign and incorporate non-manual features from a continuous video. With the advent of deep learning, recent models exploit various architectures such as RNN, LSTM, and CNN [11], [12], [35]. More recently, methods of transmitting additional information to the model through an optical flow or analyzing multiple cues using keypoints have also been studied [12], [37].

**Sign Language Translation.** Sign language translation (SLT) requires the translation of sign languages into spoken language, and learning not only one-to-one correspondence but one-to-many or many-to-one mapping between the two languages. Compared to sign language recognition tasks, SLT is known to be more challenging because this translation needs to account for the differences between two completely different languages. In 2018, Camgoz et al. [5] released the first model to handle an SLT task, along with its dataset. Subsequently, Ko et al. [20] released a KSL dataset and introduced a translation model exploiting 2D human pose keypoints. More recent works have improved the performance of SLT by using transformer based models [6], [36].

Despite these recent developments, the domains of corpora for sign language datasets remain very limited (*e.g.*, weather,

[2] Our dataset is publicly available for download at https://www.aihub.or.kr/aidata/7965.

TABLE II: **Overview of *publicly available* sentence-level *continuous* sign language datasets.** Our dataset KSL-Guide-Sentence provides extensive annotations, including 2D/3D human pose keypoints and spoken language translations as well as gloss annotations. To the best of our knowledge, we are the first to include a considerable number of interrogative sentences and the largest number of KSL videos.

| Dataset | Year | Sentences | Signers | Samples | Data Type | Annotation | | | | Interrogative Sentence | Source | Sign Language |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Gloss[1] Sequence | Translation | Keypoints[2] 2D | Keypoints[2] 3D | | | |
| Purdue RVL-SLLL [23] | 2002 | 184 | 14 | 2.58K | RGB | ✓ | ✓ | ✗ | ✗ | ✗ | Lab | American |
| RWTH-BOSTON-104 [13] | 2007 | 201 | 3 | 201 | RGB | ✓ | ✓ | ✗ | ✗ | ✓ | Lab | American |
| SIGNUM [33] | 2008 | 780 | 25 | 21.1K | RGB | ✓ | ✓ | ✗ | ✗ | ✗ | Lab | German |
| ATIS [4] | 2008 | 595 | 6 | 5.87K | RGB | ✓ | ✓ | ✗ | ✗ | ✓ | Lab | Irish, German, South African |
| RWTH-PHOENIX-Weather 2012 [15] | 2012 | 1,980 | 7 | 1.98K | RGB | ✓ | ✓ | ✗ | ✗ | ✗ | TV | German |
| RWTH-PHOENIX-Weather 2014 [16] | 2014 | 6.861 | 9 | 6.86K | RGB | ✓ | ✓ | ✗ | ✗ | ✗ | TV | German |
| S-pot [32] | 2014 | 4,328 | 5 | 4.33K | RGB | ✗ | ✓ | ✗ | ✗ | ✓ | Lab | Finnish |
| Video-based CSL [19] | 2018 | 100 | 50 | 25K | RGB+D | ✗ | ✓ | ✗ | ✓ | ✗ | Lab | Chinese |
| RWTH-PHOENIX-Weather 2014T [5] | 2018 | 8,257 | 9 | 8.26K | RGB | ✓ | ✓ | ✗ | ✗ | ✗ | TV | German |
| KETI [20] | 2019 | 105 | 10 | 1.05K[3] | RGB | ✗ | ✓ | ✗ | ✗ | ✗ | Lab | Korean |
| Corpus NGT (release-v4) [9] | 2020 | N/A | 92 | 2.38K | RGB | ✓ | ✓ | ✗ | ✗ | ✗ | Lab | Dutch |
| How2Sign [14] | 2020 | 35,191 | 11 | 35.2K[4] | RGB+D | ✓ | ✓ | ✓ | ✗ | ✗ | Lab | American |
| GSL [1] | 2020 | 331 | 7 | 10.3K | RGB+D | ✓ | ✓ | ✗ | ✗ | ✓ | Lab | Greek |
| **KSL-Guide-Sentence (Ours)** | 2021 | 2,000 | 20 | 40K[5] | RGB | ✓ | ✓ | ✓ | ✓ | ✓ | Lab | Korean |

[1] Gloss is a transcribed version of a single sign into textual form.

[2] Human pose keypoints refer to joint positions for the body, face, and hands (*i.e.*, skeletons).

[3] 2 views (frontal and side) are captured simultaneously at one recording (*i.e.*, 1,050×2 videos in total).

[4] 2 views are captured simultaneously at one recording (*i.e.*, 35.2K×2 videos in total).

[5] 5 views are captured simultaneously at one recording in a multi-camera system (*i.e.*, 200K videos in total), and they are utilized for 3D keypoints modeling.

emergencies), especially at the level of spoken language annotation (*i.e.*, translation); therefore, more datasets to train SLT models are still needed.

## B. *Sign Language Datasets*

To provide high-quality services in SL, it is necessary to secure a large amount of annotated data. However, obtaining such data is often very time-consuming and expensive as it requires considerable effort from native signers and professionals who are fluent in both SL and a spoken language. For example, a Korean sentence can be expressed in multiple KSL sentences by variable ordering and combinations of glosses because KSL and Korean have independent structures and vocabularies from each other. Therefore, it is necessarily a demanding task because native signers should speak sentences in a tightly directed way (*i.e.*, follow the defined order and given glosses accurately). Table II summarizes the characteristics of publicly released sentence-level (*i.e.*, continuous) sign language datasets of languages to date.

The sentence-level sign language datasets are used for the task of CSLR, and each of them usually has a specific domain of discourse, such as weather, jobs, emergencies, health, or travel information. A few datasets are captured with RGB and depth sensors or provide human pose keypoints as well. Note that some datasets have extracted 2D or 3D keypoints (in small quantities) and while these are utilized for experiments, they are not released to the public (*e.g.*, [14]).

Compared to the previously available datasets, the KSL-Guide has several extra features that make it more useful than the available ones. First, the KSL-Guide is a unique dataset in that it includes a considerable number of interrogative sentences, which can be used to train the model for interactive tasks such as being a sign language chatbot. For example, a guide service can help Deaf people better by interactively asking and providing specific answers to their questions, rather than simply providing a large set of information without context. Second, the KSL-Guide provides extensive publicly available annotations. The translation annotations (*i.e.*, translation from SL to the spoken language of a sentence) enable the datasets to be used for SLT tasks, and human pose keypoint annotations greatly ease the development of SL recognition or translation tasks by allowing researchers to *i)* focus on recognition/translation tasks without the extra effort to extract keypoints, and *ii)* evaluate their 2D and 3D keypoint extraction mechanisms. Third, the KSL-Guide provides the raw data for the multiple viewpoints, which can be utilized to develop view-invariant SL recognition/translation models. Finally, the KSL-Guide is the first large-scale KSL dataset, including about 20× more sentences and 40× more samples compared to the previously available KSL dataset [20], which is relatively small (see Table II).

TABLE III: **Statistics of KSL-Guide dataset.** KSL-Guide can be divided into three datasets: KSL-Guide-Sentence, KSL-Guide-Word, and KSL-Guide-Fingerspelling according to its content.

| Name | Total Samples | Contents | Signers | Camera Angles | Acquisition | Annotation |
|---|---|---|---|---|---|---|
| KSL-Guide-Sentence | 40K | 2,000 Sentences | 20 | 5 | Multi-camera (Lab) | {gloss sequence, timestamps, translation, keypoints} |
| KSL-Guide-Word | 60K | 3,000 Words | 20 | 5 | Muti-camera (Lab) | {gloss, keypoints} |
| KSL-Guide-Fingerspelling | 21K | 800 Proper Nouns<br>200 Numbers | 52 | 1 | Crowd-sourced (in the wild) | {word, keypoints} |
| Total | 521K | - | 72 | - | - | - |

# III. KSL-GUIDE DATASET

## A. *Overview*

The KSL-Guide is a large-scale Korean Sign Language dataset containing videos and annotations for sign language sentences, words, and fingerspelled words and numbers frequently used for dialogues on transportation and navigation. The main goal of this dataset is to promote research or service development with deep learning approaches for sign languages, such as continuous sign language recognition (CSLR) and sign language translation (SLT) tasks. Specifically, the dataset consists of three parts: KSL-Guide-Sentence, KSL-Guide-Words, and KSL-Guide-Fingerspelling, as shown in Table III. **KSL-Guide-Sentence** contains 2,000 different sentences in KSL, where each of 20 signers records videos for all sentences in a multi-view camera system having five cameras (*i.e.*, total of 40K samples from views). Among 2,000 sentences, our dataset includes 966 interrogative sentences, which distinguishes it from many previously released datasets. This dataset can be used for CSLR and SLT tasks as illustrated in Fig. 2. **KSL-Guide-Word** includes 60K videos of 3,000 unique

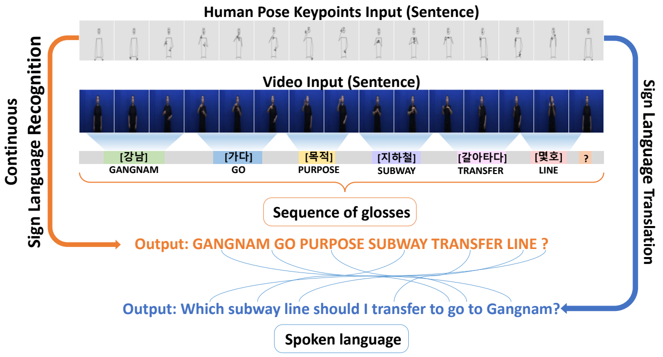

Human Pose Keypoints Input (Sentence)
Continuous Sign Language Recognition
Sign Language Translation
Video Input (Sentence)
[강남]
GANGNAM
[가다]
GO
[목적]
PURPOSE
[지하철]
SUBWAY
[갈아타다]
TRANSFER
[몇호]
LINE
?
Sequence of glosses
Output: GANGNAM GO PURPOSE SUBWAY TRANSFER LINE ?
Output: Which subway line should I transfer to go to Gangnam?
Spoken language

Fig. 2: **Illustration of CSLR and SLT tasks on KSL-Guide-Sentence.** CSLR takes as input RGB video frames and outputs a sequence of glosses while SLT takes as input 2D human pose keypoints and outputs spoken language translations.

words in KSL, recorded under the same conditions as the KSL-Guide-Sentence dataset. This dataset can be used for isolated sign language recognition tasks (top diagram of Fig. 3). Finally, **KSL-Guide-Fingerspelling** includes a total of 21K videos crowd-sourced by fifty-two signers. In each video, the signer represents a word (*i.e.*, a proper noun or number) as a sequence of fingerspelled letters and numerals. KSL-Guide-Fingerspelling is for the static sign language recognition task (bottom diagram, Fig. 3), which recognizes a single alphabet from representative frames in the video. All the videos in these three datasets are recorded in HD (1920×1080) resolution at a rate of 30 frames per second.

Each video in the dataset comes with extensive annotations. First, all videos come with word-level or sentence-level ground-truth annotations. Specifically, for each video in **KSL-Guide-Sentence**, a sequence of glosses corresponding to the video is provided. Moreover, our dataset also includes timestamp annotations, where we mark the exact start and end time for each gloss in a video sequence. In addition, we provide spoken Korean sentences (*i.e.*, translations) corresponding to the videos. Note that translation is different from the sequence of glosses since Korean (spoken language) and KSL have very different linguistic structures (*e.g.*, the order of words, lengths of sentences). For **KSL-Guide-Word** and **KSL-Guide-Fingerspelling**, word-level annotations are provided. All annotations here are carefully generated by multiple native signers as well as experts who have comprehensive understanding in both Korean and KSL. Along with word/sentence-level annotations, the KSL-Guide also provides keypoint annotations for all videos. A total of 137 human pose keypoints in 2D and 3D coordinates are provided respectively for each frame of the videos.

### *B. Data Acquisition*

**Sentence/Word Dataset.** For the KSL-Guide-Sentence/Word dataset, sign language experts carefully selected 2,000 sentences and 3,000 words that are frequently used for dialogues on public transportation (e.g., Seoul Metro, Taxi) and navigation. Then, we recruited twenty native signers who primarily use sign language as their first language. For each signer, we recorded 2,000 sentences and 3,000 words in a multi-camera system with five cameras. Specifically, a total of five synchronized cameras were placed as shown in Fig. 4: one on the front, two at 30 degrees to the left/right of the front camera, and the other two at 30 degrees to the top/bottom of the front camera. We placed a blue screen behind the signer, and we also provided enough light to secure a fast shutter speed (*i.e.*, 1/2000 sec) so as not to have motion blur. Finally, we also calibrated the intrinsic and extrinsic parameters of each camera. We set the front camera as the reference (*i.e.*, world coordinate) so that we could properly obtain 3D coordinates from each camera view. For each signer, five cameras recorded videos of 2,000 sentences and 3,000 words, resulting in 10K videos for the sentence dataset, and 15K videos for the word dataset. Thus, KSL-Guide-Sentence and KSL-Guide-Word include a total of 200K and 300K videos, respectively, from twenty signers.

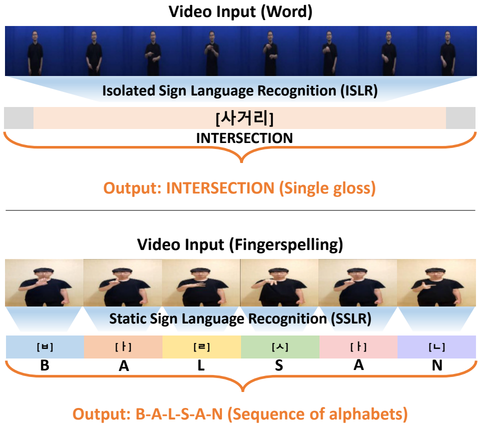

Video Input (Word)
Isolated Sign Language Recognition (ISLR)
[사거리]
INTERSECTION
Output: INTERSECTION (Single gloss)
Video Input (Fingerspelling)
Static Sign Language Recognition (SSLR)
[ㅂ]
B
[ㅏ]
A
[ㄹ]
L
[ㅅ]
S
[ㅏ]
A
[ㄴ]
N
Output: B-A-L-S-A-N (Sequence of alphabets)

Fig. 3: **Illustrations of recognition tasks: ISLR and SSLR. Top:** Isolated sign language recognition on KSL-Guide-Word; **Bottom:** Static sign language recognition on KSL-Guide-Fingerspell.

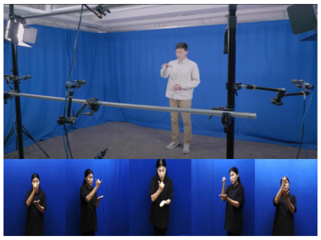


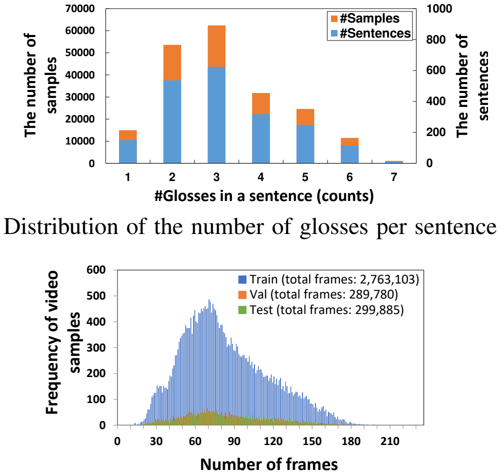
Fig. 4: **Lab setting for data acquisition of KSL-Guide-Sentence and KSL-Guide-Word**. **Top:** Our multi-camera system consists of five synchronized cameras. **Bottom:** (from left to right) Front, Left, Top, Right, and Bottom views.

**Fingerspelling Dataset.** For KSL-Guide-Fingerspelling, we recruited fifty-two native signers through a Korean online crowd-sourcing platform (similar to Amazon Mechanical Turk) named aiWorks. Fingerspelled signs are sign combinations that correspond one-to-one to the letters of the alphabet or numerals in sequential order (see Fig. 3, bottom). This dataset contains a total of 21K front view videos recorded by seventeen male and thirty-five female signers. Among those videos, 16.8K videos include fingerspelled names of places in Korea (*e.g.*, Gangnam Station, N Seoul Tower), and the remaining 4.2K videos include fingerspelled numbers. All crowd-sourced videos were recorded in the wild and exhaustively verified by professionals to ensure that each video did indeed contain the proper fingerspelling of the provided nouns and numbers.

### C. Annotations

**Gloss and Spoken Language Annotations.** Since we collected our videos in the laboratory by requesting each signer to perform assigned sentences or words, our dataset includes ground-truth gloss annotations for both KSL-Guide-Sentence and KSL-Guide-Word datasets. The ground-truth gloss annotations were generated by interpreting the Sentence/Word dataset into glosses by the experts. As depicted in Fig. 3, a single gloss label is provided for each word input video because it is an isolated sign, while a gloss sequence label is provided for the sentence because it is a continuous sign (see Fig. 2). In addition, the KSL-Guide-Sentence dataset also includes a ground-truth spoken language translation for each video. Overall, we annotated 3,000 unique signs in KSL-Guide-Word dataset, and 1,000 unique signs in KSL-Guide-Fingerspelling dataset. Moreover, in the KSL-Guide-Sentence dataset, there are 40K sentence-level annotations comprising 6,363 gloss instances, 319 vocabulary entries (unique glosses), 2,000 sentences, and corresponding spoken language sentences. The number of glosses in a sentence for 2,000 sentences ranged from 1 to 7, and the distribution is shown in Fig. 5 (a). The histogram of the number of frames for all the video samples from KSL-Guide-Sentence shows a similar distribution, as plotted in Fig. 5 (b). The unique aspect of our KSL-Guide-Sentence dataset is that we additionally provide *timestamp* data, which segments each sign (*i.e.*, gloss) consisting of a sentence (see Fig. 1). For this purpose, three Deaf oracles who speak sign language as their first language manually marked the start and end time of each gloss in the video. They checked the section for each gloss segment and modified it according to the majority rule. We envision that this timing annotation could potentially be used to train deep learning models for gloss segmentation tasks within continuous video, such as for sign spotting [2], [24], [28], [32].

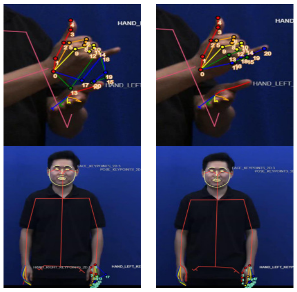
**Human Pose Keypoint Annotations.** For generation of keypoint annotations in KSL-Guide-Sentence and KSL-Guide-Word, we first extracted 2D pose keypoints from two viewpoints (*i.e.*, front and left) using OpenPose [7]. With OpenPose, we extracted 25 body keypoints, 21 hand keypoints from each hand, and 70 facial keypoints, resulting in a total of 137 keypoints for each frame. These keypoints were then reviewed and manually corrected by human annotators to make the annotated keypoints more accurate (see Fig. 6). With these keypoints from two viewpoints, we used triangulation [17] to obtain 3D pose keypoints and then reprojected these 3D keypoints onto three 2D spaces. Each one corresponded to the viewpoints of the three remaining cameras (*i.e.*, three cameras on the left, top, and bottom of the front camera). Again, the resulting 2D keypoints were reviewed by annotators to improve the accuracy. We find that our triangulation and 3D modeling-based mechanism allowed us to obtain 2D keypoints more accurately compared to simply utilizing the 2D keypoints detected by OpenPose on all five viewpoints. This is because OpenPose sometimes fails to extract keypoints from some viewpoints due to occlusions resulting from hand movements (*i.e.*, hand-hand overlapping or face-hand occlusion). This mechanism also allowed us to reduce the effort of the annotators in reviewing the 2D keypoints. The sequences of pose keypoints were used to train the SLT model (*i.e.*, translating from pose sequences to spoken language), and could also be used in Sign Language Production (SLP) of deep learning approaches. The SLP model [29], [31] generates continuous sign language videos directly from spoken language using 2D pose keypoints, and [30] produces continuous 3D sign pose sequences from gloss or spoken language. In this manner, we further expect that both 2D and 3D human pose keypoints we annotated in sign language could be exploited in various tasks.



(a) Distribution of the number of glosses per sentence

(b) Histogram of frame numbers

Fig. 5: **Statistics of the KSL-Guide-Sentence dataset.** (a) Each sentence from KSL-Guide-Sentence consists of 1 to 7 glosses in a sequence; (b) Each video from KSL-Guide-Sentence consists of 14 to 236 frames.



(a) Before correction

(b) After correction

Fig. 6: **Example of annotated human pose keypoints**. (a) example of wrong estimated keypoints by OpenPose due to self-occlusion; (b) corrected keypoints

## IV. EXPERIMENTAL ANALYSIS

### A. Methodology

We evaluate our KSL-Guide-Sentence dataset on two different tasks: continuous sign language recognition (CSLR) and sign language translation (SLT). Specifically, our goal is to identify whether including *interrogative sentences* in the training and validation sets leads to improved model performance on a test set with interrogative sentences. Such a test set closely models real-world use cases for service applications where question answering is critical.

**Dataset.** For these experiments, we use a subset of the KSL-Guide-Sentence dataset. Specifically, we only utilize frontal view videos and do not employ timestamp annotations to follow the real-world use cases in which multiple cameras are typically unavailable and input signs are continuous. Among the 2,000 sentences recorded by twenty signers, 966 are interrogative sentences, and the other 1,034 are declarative sentences. The interrogative and declarative sentence compositions are mostly pairs of questions and statements or answers about the same content. (*e.g.*, "How can I get to the opposite side of the platform?" or "I want to go to the opposite side of the platform." "What's the balance on my transportation card?" or "You have enough balance.").

**Experimental Configurations.** From the dataset, we randomly select a portion (*i.e.*, around 9%) of declarative sentences as a validation set and another portion as a test set. Similarly, interrogative sentences are randomly selected as a validation and test set, respectively. The remaining videos constitute a training set. We design four different experiments based on this split and summarized the configurations in Table IV (top row). Experiment 1 is the case where only declarative sentences are included in the training/validation/test sets. Experiment 2 is the case where the training and validation sets only include declarative sentences while the test set includes both interrogative and declarative sentences. Experiment 3 is the case where all training, validation, and test sets include interrogative sentences as well as declarative sentences. This experiment uses all front view videos in KSL-Guide-Sentence dataset. Finally, the last configuration, Experiment 4 is where we configured the size of the training set to be similar to that of Experiment 1. This configuration helps us identify whether the potential improvement in the model performance was from increase in the size of the training set or from the inclusion of interrogative sentences.

### *B. Continuous Sign Language Recognition*

**Task Description.** Continuous sign language recognition (CSLR) is the task of identifying a sequence of glosses from a given input video. To measure the model performance, we utilize a commonly used metric named the word error rate (WER). This is essentially the length-normalized word-level edit distance (*i.e.*, the minimum operation required for the predicted sequence to be matched to the ground truth). A lower WER value indicates that the recognition result is close to the reference, while a high WER value indicates that the recognition result does not resemble the original sequence of glosses.

$$WER = \frac{\text{\#substitutions} + \text{\#deletions} + \text{\#insertions}}{\text{\#words in reference}} \qquad (1)$$

**Model.** We use one of the state-of-the-art CSLR models called a fully convolutional end-to-end network (FCN) [8] as a baseline, and which concurrently learns spatial and temporal features from weakly annotated videos with sentence-level annotations (*i.e.*, does not include timestamps for each gloss). This model only takes RGB video frames as inputs and demonstrates high performance on previously released datasets such as PHOENIX14 [16] and CSL [19]. We choose FCN for the evaluation model because it is end-to-end trainable without pre-training and capable of online recognition that is appropriate for real-world applications. We could get 26.9% WER on PHOENIX14T dataset [5] from our implementation.

**Training details.** We re-implement the FCN network and mostly utilize the default parameters listed in [8]. Since this model takes video frames of which the size is 224 × 224, we resize the videos in our datasets. The mini-batch size is set to 2, and we train the model using a single Quadro RTX 8000 GPU. We train for 80 epochs using the Adam optimizer with momentum 0.9, and an initial learning rate of 0.0001. The learning rate is downscaled by a factor of two at epochs 40 and 60. Unless mentioned specifically, other options were kept the same as for the default training configurations.

**Results.** Table IV (bottom row) shows the CSLR model's WER on four experimental configurations explained above. First, the results from Experiment 1 and 2 show that the model only trained with declarative sentences fails to perform well on the test set with interrogative sentences. On the other hand, the results of Experiment 3 demonstrate that the same model can achieve good performance when trained with the dataset including both declarative and interrogative sentences. Finally, the results of Experiment 4 show that the model still achieves accuracy comparable with that in Experiment 3, despite the reduced size of the training set. Additionally, we further cut down the size of the training set to 8.25K (*i.e.*, half the size of Experiment 4) and test the result on the same test set as with the other experiments. The result is 12.97% of WER, which means even a much smaller training set can work effectively and comparably. This confirms that the inclusion of interrogative sentences is the primary source of the lower WER rather than the increase in the size of the training set. Overall, we find that the inclusion of interrogative sentences in the training dataset almost halve the WER. This shows that the use of our dataset–KSL-Guide, which includes interrogative sentences, can potentially result in noticeable performance improvements of many CSLR models for real-world applications in terms of understanding user questions.

### *C. Sign Language Translation*

**Task Description.** Sign language translation (SLT) is the task of obtaining spoken language sentences from sign language videos. SLT is fundamentally similar to other translation tasks; yet, it demands additional effort to accommodate its visual features. The majority of SLT has two parts: tokenization and neural machine translation (NMT). Here, we use a method in which human pose keypoints are used for the tokenization. Then tokens are directly converted to spoken language sentences in the NMT architecture without

TABLE IV: **Summary of dataset splits and the experiment results. Top row:** Experimental configurations; Experiment 3 is the case of using all data in KSL-Guide-Sentence, and the other experiments are using a subset from Experiment 3. **Bottom row:** Results are achieved by the four configurations above on CSLR and SLT tasks. The lower value for CSLR and the higher value for SLT indicate better performance.

| | Experiment 1 | | | Experiment 2 | | | Experiment 3 | | | Experiment 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sentence type | Train | Val | Test | Train | Val | Test | Train | Val | Test | Train | Val | Test |
| # Declarative | 17K | 1.7K | 1.8K | 17K | 1.7K | 1.8K | 17K | 1.7K | 1.8K | 7.9K | 1.7K | 1.8K |
| # Interrogative | – | – | – | – | – | 1.8K | 16K | 1.7K | 1.8K | 8.6K | 1.7K | 1.8K |
| Total samples | 17K | 1.7K | 1.8K | 17K | 1.7K | 3.6K | **33K** | 3.4K | 3.6K | 16.5K | 3.4K | 3.6K |

| | | | Result | | |
|---|---|---|---|---|---|
| Task | Metric | Test (Val) | Test (Val) | **Test** (Val) | Test (Val) |
| CSLR | WER ↓ | 6.76 (9.34) | 25.24 (9.34) | **8.88** (7.18) | 10.85 (11.20) |
| SLT | BLEU ↑ | 76.10 (76.68) | 39.53 (76.69) | **72.21** (76.06) | 71.08 (74.01) |
| | ROUGH-L ↑ | 85.27 (85.61) | 55.18 (85.64) | **81.10** (84.76) | 80.35 (83.66) |
| | METEOR ↑ | 60.88 (59.57) | 41.85 (59.60) | **58.72** (61.28) | 58.22 (60.26) |

explicit intermediates like glosses. To measure the model performance, we employ various metrics such as BLEU [27], ROUGE-L [22], and METEOR [3], which are frequently used to evaluate the quality of machine translations.

**Model.** For our experiments, we use as a baseline the recently released sign language translation model [20], which only takes keypoints as input. This property makes the model less vulnerable to the background and therefore has an advantage in real-world scenarios. This model takes as input 124 human pose keypoints (*i.e.*, 12 body, 21 for each hand, and 70 face), excluding lower body parts, which are not shown in the videos. The 2D coordinates of the pose keypoints we annotated in KSL-Guide-Sentences are provided as inputs to the model. Specifically, we only utilize keypoints obtained from the front camera when running this model. Finally, we also perform a data augmentation by random frame skip sampling, as specified in the original paper [20].

**Training details.** We train the model with a single NVIDIA V100 GPU using the mini-batch size of 128. The model was trained for a total of 100 epochs using the Adam optimizer, and the initial learning rate is set to 0.001. We adjust the learning rate every 20 epochs with an exponential decay factor of 0.5. We use a dropout probability of 0.5 and a gradient clipping threshold of five. Moreover, the object 2D normalization method and data augmentation by a factor of 10 were used. The other hyper-parameters followed the default configuration listed in the paper presenting the model [20].

**Results.** Table IV (bottom row) shows the results of running the SLT model across four different experimental configurations. The overall trend is similar to that of the CSLR task. Comparing the experimental results from Experiment 3 and 2 demonstrates that including interrogative sentences in the dataset substantially improves all three metrics by: 32.68% (BLEU), 25.29% (ROUGE-L), and 16.87% (METEOR). Also, the results from Experiment 3 and 4 indicate that the impact of increased training set size is not very significant when compared to the impact of including interrogative sentences. Overall, we find that simply including interrogative sentences can greatly boost performance without any change in the model. This implies that securing a proper set of data is as important as advancing the model architecture for both sign language recognition and translation tasks.

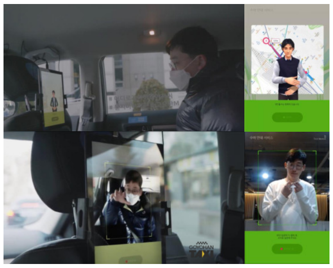


Fig. 7: **Prototype application.** Using the sign language translation model trained on our dataset, the application can facilitate the communication between Deaf and non-deaf people.

## V. APPLICATION SCENARIOS

We developed a prototype application exploiting the SLT model trained on KSL-Guide that could be used as an assistance system for Deaf and hard-of-hearing people. This prototype is intended to provide an interactive guide service with respect to transportation and navigation for Deaf people. Specifically, a signer can ask a question to this prototype application using Korean Sign language (KSL), and the application presents an answer to the question by displaying a 3D avatar speaking KSL (see Fig. 7).

For this application, we selected a total of 1,000 sentences from KSL-Guide-Sentence. Then, we prepared the answers for each interrogative sentence in the chosen set. For each answer, we prepared a video clip containing an animation of a 3D avatar speaking the answer in KSL. The video was generated by performing motion capture on a native signer speaking the answer. When a user asks questions using sign language, the neural network model [20] translates the sign language sentences into spoken language (Korean); and then the application answers the question using prepared correspondence. The demo for this application is attached as a supplementary video. This video shows the potential for Deaf people to communicate with non-sign speakers in their daily lives, such as at a subway station or in a taxi, through this KSL-Guide Service (see Fig. 7).

## VI. CONCLUSIONS AND FUTURE WORKS

We present a new large-scale Korean Sign Language (KSL) dataset for guiding Deaf and hard-of-hearing people. The KSL-Guide dataset includes extensive ground-truth annotations that can potentially be used to develop automatic sign language recognition and translation models, as well as to do other complex tasks such as sign spotting, human pose estimation, and multimodal-based sign generation. As a contribution, our work distinguishes itself from previous works by having a considerable number of interrogative sentences. Many real-world applications of sign language recognition and translation models require handling of interrogative sentences. We have demonstrated the advantage of having a considerable number of interrogative sentences in a dataset, which allows high performance of recognition and translation tasks, and facilitates interactive communication between Deaf and hearing people in the community. Future work will propose a baseline model specialized in recognizing non-manual features in sign language and tackle the potential gap between laboratory and real-life conditions.

## VII. ACKNOWLEDGMENTS

This work was supported by the AI Data Construction Project funded by the National Information Society Agency. We thank all the signers who contributed to this project, EQ4ALL Co., Ltd., Korea Nazarene University, and the Korea Association of the Deaf for their support. Special thanks to Dr. Koh from Korea Nazarene University for her valuable feedback and advice on Korean Sign Language.

## REFERENCES

[1] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2020.

[2] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*, 2020.

[3] S. Banerjee and A. Lavie. 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*, pages 65–72, 2005.

[4] J. Bungeroth, D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way, and L. van Zijl. The atis sign language corpus. 2008.

[5] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[6] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033, 2020.

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[8] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai. Fully convolutional networks for continuous sign language recognition. In *European Conference on Computer Vision*, pages 697–714. Springer, 2020.

[9] N. Corpus. An open access digital corpus of movies with annotations of sign language of the netherlands. *Nijmegen: Centre for Language Studies, Radboud University Nijmegen.*, 2008.

[10] O. Crasborn. Nonmanual structures in sign language. *Encyclopedia of Language and Linguistics*, 8:668–672, 12 2006.

[11] R. Cui, H. Liu, and C. Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7361–7369, 2017.

[12] R. Cui, H. Liu, and C. Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, 2019.

[13] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. Speech recognition techniques for a sign language recognition system. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[14] A. Duarte, S. Palaskar, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. *arXiv preprint arXiv:2008.08143*, 2020.

[15] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater, and H. Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789, 2012.

[16] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916, 2014.

[17] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003.

[18] A. Herrmann and M. Steinbach. *Nonmanuals in sign language*, volume 53. John Benjamins Publishing, 2013.

[19] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683, 2019.

[21] O. Koller, S. Zargaran, and H. Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4297–4305, 2017.

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

[23] A. M. Martínez, R. B. Wilbur, R. Shay, and A. C. Kak. Purdue rvl-slll asl database for automatic recognition of american sign language. In *International Conference on Multimodal Interfaces*, pages 167–172. IEEE, 2002.

[24] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[25] National Association of the Deaf. American sign language: Community and culture. https://www.nad.org/resources/american-sign-language/community-and-culture-frequently-asked-questions/. Online.

[26] S. C. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Computer Architecture Letters*, 27(06):873–891, 2005.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318, 2002.

[28] K. Renz, N. C. Stache, S. Albanie, and G. Varol. Sign language segmentation with temporal convolutional networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE, 2021.

[29] B. Saunders, N. C. Camgoz, and R. Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020.

[30] B. Saunders, N. C. Camgoz, and R. Bowden. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705. Springer, 2020.

[31] S. Stoll, N. C. Camgöz, S. Hadfield, and R. Bowden. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the British Machine Vision Conference*. BMVA, 2018.

[32] V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa, and J. Laaksonen. S-pot–a benchmark in spotting signs within continuous signing. In *Proceedings of the International Conference on Language Resources and Evaluation*. ELRA, 2014.

[33] U. Von Agris, M. Knorr, and K.-F. Kraiss. The significance of facial features for automatic sign language recognition. In *International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008.

[34] World Health Organization. Addressing the rising prevalence of hearing loss, 2018.

[35] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai. Sf-net: Structured feature network for continuous sign language recognition. *arXiv preprint arXiv:1908.01341*, 2019.

[36] K. Yin and J. Read. Better sign language translation with stmc-transformer. In *International Conference on Computational Linguistics*, pages 5975–5989, 2020.

[37] H. Zhou, W. Zhou, Y. Zhou, and H. Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 13009–13016, 2020.