Information Processing and Management 61 (2024) 103774

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects

Sarah Alyami [a,b], Hamzah Luqman [a,c,*], Mohammad Hammoudeh [a]

[a] Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Saudi Arabia
[b] Computing Department, Applied College, Imam Abdulrahman Bin Faisal University, Saudi Arabia
[c] SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Saudi Arabia

ARTICLE INFO

Keywords:
Continuous sign language recognition
Sign language recognition
Sign language translation
Gesture recognition
Video understanding

ABSTRACT

Sign language is a form of visual communication employing hand gestures, body movements, and facial expressions. The growing prevalence of hearing impairment has driven the research community towards the domain of Continuous Sign Language Recognition (CSLR), which involves identification of successive signs in a video stream without prior knowledge of temporal boundaries. This survey article conducts a review of CSLR research, spanning the past 25 years, offering insights into the evolution of CSLR systems. A critical analysis of 126 studies is presented and organized into a taxonomy comprising seven critical dimensions: sign language, data acquisition, input modality, sign language cues, recognition techniques, utilized datasets, and overall performance. Additionally, the article investigated the classification of deep-learning CSLR models, categorizing them based on spatial, temporal, and alignment methods, while identifying their advantages and limitations. The article also explored various research aspects including CSLR challenges, the significance of non-manual features in CSLR systems, and identified gaps in existing literature. This literature taxonomy serves as a resource aiding researchers in the development and positioning of novel CSLR techniques. The study emphasizes the efficacy of multi-modal deep learning systems in capturing diverse sign language cues. However, the examination of existing research uncovers numerous limitations, calling for continued research and innovation within the CSLR domain. The findings not only contribute to the broader understanding of sign language recognition but also lay the foundations for future research initiatives aimed at addressing the persistent challenges within this emerging field.

1. Introduction

Hearing loss can significantly impact the quality of life of those affected. Deaf or speech-impaired individuals are more likely to suffer from depression and anxiety due to the difficulties faced in communicating their feelings and needs, leading to their withdrawal and isolation from their communities. According to the World Health Organization, the prevalence of hearing loss is on the rise, with estimates suggesting that by 2050, one in every 10 people will be affected (WHO, 2021). Deaf or speech-impaired people primarily rely on sign language for communication. Remarkably, there are over 300 sign languages in use globally, as reported by the World Federation of the Deaf (WHO, 2021). While sign language interpreters can play a crucial role in bridging the communication gap between hearing-impaired people and the wider community by translating sign language into speech and vice versa. However, this solution is impractical due to the scarcity of sign language interpreters and the ever-increasing number of

* Corresponding author at: Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Saudi Arabia.
E-mail addresses: snalyami@iau.edu.sa (S. Alyami), hluqman@kfupm.edu.sa (H. Luqman), mohammad.hammoudeh@kfupm.edu.sa (M. Hammoudeh).

https://doi.org/10.1016/j.ipm.2024.103774
Received 5 January 2024; Received in revised form 27 March 2024; Accepted 8 May 2024
Available online 24 May 2024

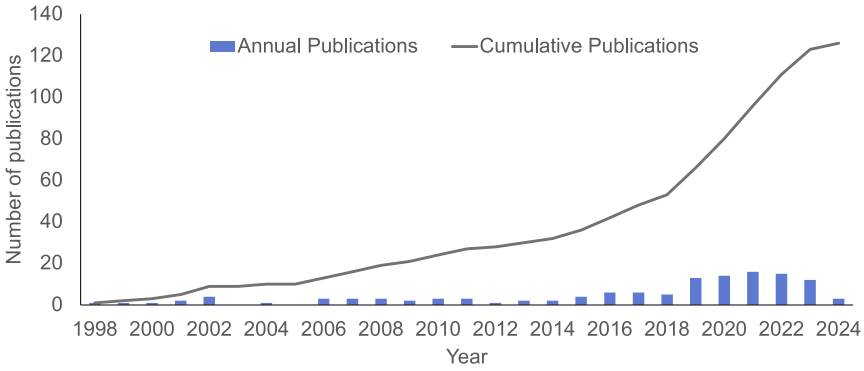

Fig. 1. Number of CSLR publications by year.

deaf and speech-impaired individuals worldwide. Consequently, technology-based sign language interpreting systems offer a scalable and practical solution that can bridge the communication divide between the deaf and hearing community.

Sign language is a visual language that uses a combination of hand gestures, body movements, and facial expressions for communication. Manual gestures performed using the signer's hands are characterized by four key features: hand shape, palm orientation, location, and movement. Meanings expressed using manual hand gestures are often reinforced by employing non-manual features such as body posture, head nodes, facial expressions, and lip patterns (Aloysius & Geetha, 2020). Despite the common misconception that sign languages are not real languages, they are, in fact, complete languages with their own grammar and syntax (Jachova, Kovacheva, & Karovska, 2008). Sign languages are linguistically independent of their related spoken language counterparts and do not necessarily share the same grammatical features. There are hundreds of distinct sign languages, with each region typically having its own unique sign language. Even countries that share a common spoken language may have independent sign languages, such as British Sign Language (BSL) and American Sign Language (ASL). This diversity arises because sign languages were developed by local deaf communities and, therefore, are heavily influenced by geographic location and local cultural customs (Jachova et al., 2008).

Sign language recognition (SLR) involves the task of recognizing and interpreting gestures typically conveyed in videos or images. SLR systems map sign gestures to their corresponding spoken-language words. The term "gloss" is used to describe the word label of the sign. SLR usually involves employing pattern recognition, computer vision, and Natural Language Processing (NLP) techniques. SLR systems can be categorized based on the recognized signs into finger-spelled, isolated, and continuous SLR systems. Finger-spelled signs are employed for alphabets, digits, and spelling names that lack equivalent signs. These signs are generally static with no motion and can be represented using still images. Conversely, isolated signs are equivalent to spoken words in natural languages. These signs can be static or dynamic, requiring video clips to represent dynamic signs. The type of SLR systems that target these signs are called isolated SLR systems. Continuous SLR (CSLR) involves recognizing multiple consecutive signs in a video stream. While isolated SLR can assist in sign language interpretation, it is limited to interpreting signs in isolation and cannot accurately translate sign language sentences and dialogues, as opposed to CSLR. CSLR is more complex, compared to isolated SLR, because the signs in the videos are unsegmented, with no clear pauses between them. Therefore, predicting the correct sequence of signs necessitates identifying the beginning and end of each sign within a series of successive frames (Aloysius & Geetha, 2020). CSLR has the potential to be a valuable tool for deaf and hard-of-hearing people by enhancing communication access in a variety of settings, including classrooms, workplaces, and healthcare. CSLR models can be utilized in several AI-powered applications, such as automatic video captioning, interpretation services, virtual reality games, and robotics (Wadhawan & Kumar, 2021). Due to the rapid advancements in deep learning (DL), CSLR had garnered significant attention from researcher in the past decade. Notably, approximately 70% of research studies in this field have been published within the last seven years, as illustrated in Fig. 1. However, despite the substantial research efforts in the CSLR, there is still ample room for improvement. In comparison to speech recognition, CSLR systems are far from reaching their full potential and become commercialized products, as current CSLR systems have limited vocabulary and are not designed for real-time applications. As such, this study seeks to provide a comprehensive literature review that showcases the advancements of CSLR and also identifies existing research gaps. The main contributions of this study can be summarized as follows:

- Providing a deep understanding of the task of CSLR, including its problem description and challenges.
- Presenting a detailed overview of CSLR datasets and methods proposed over the past two decades.
- Discussing various aspects of CSLR including, data capturing methods, input modalities, features, and techniques.
- Proposing a fine-grained taxonomy to classify the reviewed literature based on several aspects.
- Identifying open research gaps and outlining potential future research directions for enhancing the robustness of CSLR approaches.

The remainder of this review is organized as follows: Section 2 provides a review of related surveys. Section 3 offers an overview of the CSLR task, encompassing discussions on its challenges, datasets, acquisition devices, modalities, and features. Section 4

Table 1
A comparative analysis of existing SLR surveys.

Reference	Period	Features						
		F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
Aloysius and Geetha (2020)	1997–2019	x	x	✓	✓	x	✓	✓
Wadhawan and Kumar (2021)	2007–2017	✓	✓	x	x	x	x	x
Rastgoo et al. (2021)	2015–2020	x	✓	x	✓	✓	✓	✓
Adeyanju et al. (2021)	2001–2021	✓	✓	x	✓	x	x	x
Al-Qurishi et al. (2021)	2014–2021	✓	✓	x	✓	✓	✓	✓
Papastratis, Chatzikonstantinou, et al. (2021)	2018–2021	✓	✓	x	✓	x	✓	x
El-Alfy and Luqman (2022)	1998–2019	✓	✓	x	✓	x	✓	✓
Our survey	1997–2023	✓	✓	✓	✓	✓	✓	✓

F1: Reviewed vision and sensor based CSLR, F2: Discussed data acquisition methods, F3: Identified CSLR challenges, F4: Listed CSLR datasets, F5: Discussed the use of manual and non-manual features, F6: Identified CSLR open issues, and F7: Presented a taxonomy on SLR.

presents a detailed review of CSLR frameworks. Section 5 summarizes and analyzes the findings of this survey. Finally, Section 6 concludes this study, highlighting open research issues and charting directions for future research.

2. Related surveys

Numerous reviews on SLR have been previously presented. The paper (Aloysius & Geetha, 2020) offered an insightful overview of CSLR, discussing its challenges and methods. However, they primarily focused on vision-based methods only and did not cover sensor-based CSLR techniques. Moreover, DL-based CSLR was not discussed deeply, as the majority of the reviewed CSLR techniques adopted traditional approaches to CSLR. A comprehensive literature review on SLR research was presented in El-Alfy and Luqman (2022), which reviewed both isolated and continuous SLR methods. Nonetheless, it did not discuss CSLR in depth, including its challenges and approaches. Similarly, the authors (Rastgoo, Kiani, & Escalera, 2021) surveyed DL-based SLR approaches presented in the past five years. Their review concentrated solely on vision-based approaches and omitted sensor-based SLR techniques. The survey also focused on deep learning SLR methods and did not address classical (non-deep learning) approaches. Papastratis, Chatzikonstantinou, Konstantinidis, Dimitropoulos, and Daras (2021) presented a systematic literature review on the techniques of SLR, translation, and production. Nevertheless, they only included recent studies published after 2018; hence, the presented review was not comprehensive and lacked a clear overview of the research progress since the early years. Another systematic literature review on SLR was conducted in Wadhawan and Kumar (2021), but considered only research between 2007 and 2017, which are evidently outdated. The authors in Al-Qurishi, Khalid, and Souissi (2021) reviewed SLR approaches proposed between 2014 and 2021. However, the survey focused more on isolated SLR, while giving limited attention to CSLR approaches and challenges. The review paper (Adeyanju, Bello, & Adegboye, 2021) explored machine learning-based SLR approaches from 2001 and 2021, though CSLR methods received relatively little coverage.

Table 1 presents a comparative analysis of existing reviews of CSLR. Contrary to previous surveys, this survey paper primarily focuses on CSLR by providing a comprehensive review of CSLR research conducted over the past 25 years. It also presents a taxonomy (Fig. 3) of various research directions related to different sign languages, acquisition devices, input modalities, resources, recognition techniques, and challenges. In addition, it highlights the current gaps in the literature and suggests potential research directions to enhance the performance of CSLR. The following databases were used as data sources for our survey: Google Scholar, ProQuest, Scopus, and IEEE Explore. The initial research yielded 306 papers, which were subsequently reduced to 126 after excluding isolated SLR papers, studies that do not focus on the SL recognition, such as SL translation and signs spotting, and studies not written in English. We thoroughly reviewed all relevant publications until March 2024.

3. Continuous sign language recognition

CSLR refers to the process of recognizing and interpreting sign language gestures performed continually without pauses between signs. CSLR has become a crucial area of research owing to the increasing number of individuals with hearing impairments who use sign language as their primary mode of communication. CSLR is a sequence-to-sequence problem, where the goal is to map a set of frames $X = \{x_1, x_2, x_n\}$ to a set of glosses $Y = \{y_1, y_2, y_m\}$, where $n \neq m$. The produced glosses are order-consistent with the signs in the videos. Therefore, the problem lies in predicting the correct gloss sequence from a video featuring a sequence of signs. This process involves two tasks: establishing temporal boundaries from weakly annotated video sequences and recognizing the demonstrated signs.

Generally, a DL-based CSLR framework comprises three key modules: spatial, temporal, and alignment, as illustrated in Fig. 2. The spatial module extracts spatial features from the video-based input. The temporal module is responsible for modeling the temporal aspects of the sign sequence. Finally, the alignment module controls learning the alignment between the sequences and gloss labels to produce an ordered sequence of glosses.

The performance of CSLR is often measured using the Word Error Rate (WER) metric. This metric counts the required number of deletion, insertion, and substitution operations to transform the predicted label sequence into the actual label sequence. A lower WER indicates that the system is more accurate. In earlier studies, performance was reported in terms of accuracy, defined as the

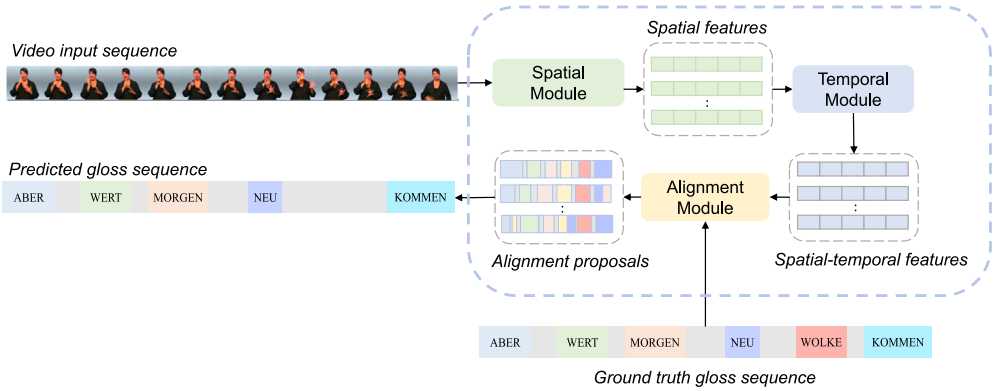

Fig. 2. General CSLR framework.

number of correctly identified signs over the total number of signs in the sentence. The Bilingual Evaluation Understudy (BLEU) metric can also be employed to assess CSLR systems. It involves performing n-gram comparisons between the predicted sentence and the ground truth sentence. However, the BLEU metric is more popular in translation tasks.

CSLR and Sign Language Translation (SLT) are two closely related tasks within the field of computer vision and NLP. CSLR systems produce a sequence of sign labels in the same order as they were signed in the video. This order is usually different than the structure of the natural language sentences (Aloysius & Geetha, 2020). SLT is an NLP task that extends CSLR by translating the recognized signs by CSLR into a spoken or written natural language, such as English or Arabic (Luqman & Mahmoud, 2019). SLT involves understanding the sign language's context, grammar, and semantics to generate coherent and contextually appropriate translations. This task is inherently more complex than CSLR as it requires understanding the syntax and grammar of the sign language and target language. Two approaches can be used to integrate the SLT with the CSLR: sign2gloss2text (Camgöz, Koller, Hadfield, & Bowden, 2020; Zhang, Müller, & Sennrich, 2023) and sign2text (Chen et al., 2022). In the first approach, a CSLR model is leveraged to produce intermediate glosses and the SLT system processes these glosses to generate coherent and grammatically correct text in the target language. On the other hand, the sign2text approach bypasses the intermediate step of glossing and directly translates sign language utterances into natural language text.

3.1. CSLR challenges

SLR is a complex problem entailing several challenges. For instance, one significant challenge involves the capture of features from multiple aspects. Sign language cues from different body parts are captured simultaneously, such as hand movements and facial expressions. Hence, an accurate SLR system must be capable of capturing and effectively joining these manual and non-manual cues. Another formidable problem is the issue of signer independence. Signers exhibit different appearances, such as skin color, body type, and height. Additionally, signers may have their unique ways of performing the same sign, influenced by factors such as hand dominance, signing speed, and sign language proficiency. These signer attributes can introduce considerable sign variations, which leads to difficulties in generalizing the SLR system to unseen signers when deployed in real-time applications (Elakkiya & Selvamani, 2019; Rastgoo et al., 2021).

Compared to finger-spelled and isolated SLR systems, CSLR is a more complex task that poses several additional difficulties, both from computational and linguistic perspectives. Beyond the aforementioned challenges, we discuss some key challenges of CSLR in the following points:

- Signs boundary detection and temporal segmentation: Identifying the start and end points of individual signs within a continuous sequence is a challenging task. This can be attributed to the lack of clear boundaries and pauses between signs in continuous sign language. Numerous approaches have been introduced for sign boundary detection to perform temporal segmentation prior to the recognition (Koulierakis, Siolas, Efthimiou, Fotinea, & Stafylopatis, 2021; Wei, Zhao, Zhou, & Li, 2021). The segmented signs are then recognized using an isolated SLR system. However, the performance of the recognition systems depends on the accuracy of the segmentation system, which is usually low. To address this issue, other techniques, such as Hidden Markov Models (HMM) and CTC, have been utilized in the literature to perform automatic alignment without the need for prior segmentation.
- Movement epenthesis: Movement epenthesis is a non-sign sequence that appears in sign language sentences when transitioning from one sign to another. CSLR models need to distinguish signs from movement epenthesis. Some studies explicitly defined procedures for movement epenthesis detection (Choudhury, Talukdar, Bhuyan, & Sarma, 2017; Yang, Tao, & Ye, 2016), while the majority of proposed methods implicitly handled movement epenthesis (Boháček & Hrúz, 2022; Zhang & Zhang, 2021).
- Co-articulation effect: Co-articulation in sign language refers to the changes in the beginning and ending of a sign based on the previous and the next signs. That is, the appearance of signs may be affected by neighboring signs. This increases intra-class variation and poses a challenge in sign spotting and recognition (Athira, Sruthi, & Lijiya, 2019).

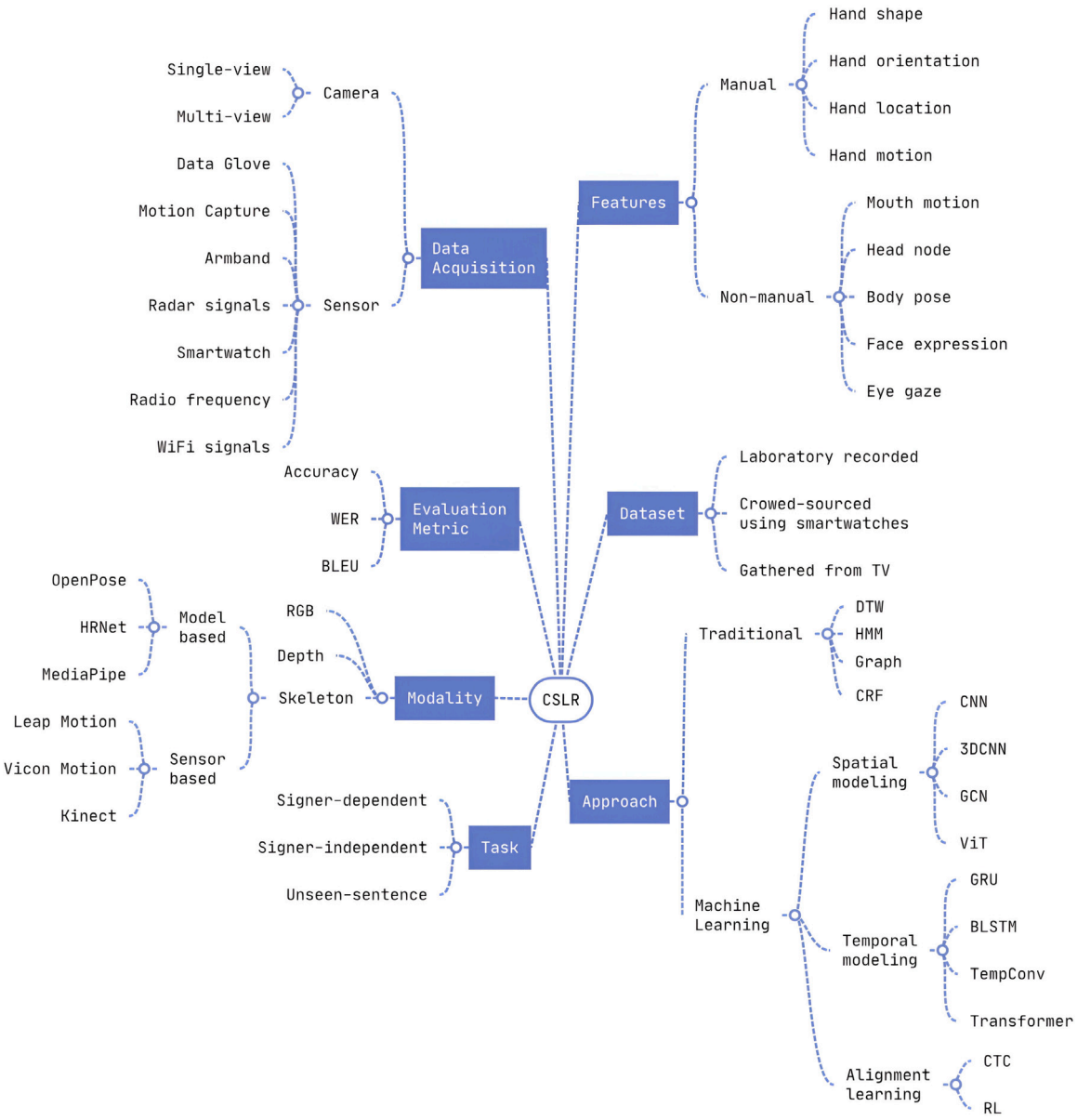

Fig. 3. CSLR taxonomy.

- Sign dependencies: Unlike isolated SLR, CSLR needs to model dependencies among the sequence of signs. Some models assume sign independence (for example, HMM-based models) and predict each sign independently, and then concatenate them to form the gloss sequence. However, these systems do not capture the contextual information and may provide inaccurate recognition. Therefore, there is a need for models that capture long-distance temporal relationships with high-level semantic information.
- Using finger-spelled signs within a sign language sentence: In the real world, sign language conversations may include finger-spelled signs. These signs are short and motionless as opposed to dynamic signs. A challenging aspect is to develop models that can jointly recognize finger-spelled and dynamic signs. In addition, motions may be added to static finger-spelled signs due to movement epenthesis. This will make recognizing these signs more challenging. Moreover, some isolated signs use fingers within the sign which adds another difficulty to the CSLR systems.

3.2. CSLR datasets

Building an intelligent SLR system relies mainly on having access to annotated datasets. SLR datasets can be categorized into isolated and continuous sign language datasets. Isolated signs datasets contain a collection of videos or images that capture individual

signs. These datasets focus on recording discrete signs, typically without a continuous conversation or context. This type of dataset is often employed for training and evaluating isolated SLR systems. In contrast, continuous sign language datasets contain signs performed continuously without pauses to form sentences and are utilized for developing CSLR systems. Isolated sign datasets are more common than continuous sign language datasets. This is because performing isolated signs is easier and does not require significant sign language experience. Conversely, sign language sentences are more difficult to perform, and an experienced signer is needed to form the sentence. Due to the lack of CSLR datasets, researchers often resort to building their own small datasets to train and evaluate their proposed recognition approaches. This makes it difficult to use these private datasets for cross-model performance comparisons. Several aspects are important to consider regarding SLR datasets. One important aspect is vocabulary size, which is the number of unique signs. The dataset should cover a wide range of vocabulary to obtain useful SLR systems capable of being incorporated in different domains. Another important aspect to consider is the number of signers in the dataset. The dataset should include different signers to create signer-independent CSLR systems. The variation amongst the signers in terms of appearance, signing speed, and being right-handed or left-handed can help in creating robust SLR models capable of generalizing to unseen signers when adopted in the real world. The number of samples per sign or sentence is also important, especially when the number of signers is limited. There should be a sufficient number of instances to increase variability and prevent over-fitting. Moreover, the dataset should include dynamic signs that require motion as opposed to static signs often used for alphabet letters or out-of-vocabulary signs.

In addition to the aforementioned criteria, CSLR datasets should encompass a large number of sentences that cover a wide range of vocabulary. The sentences should be performed in a realistic manner with no long pauses between the signs or any indication to segment the sentence. This is because the model would be trained to perform the segmentation and alignment process. The dataset should also contain sentences of varying lengths with different types of signs (dynamic, static, and finger-spelled signs). Moreover, gloss annotation should be provided with continuous sign language datasets to train the model to recognize sign language sentences. It is also recommended to have the natural language translation of each sentence. This helps in studying the linguistic characteristics of sign language and developing machine translation systems capable of translating sign language sentences into spoken language.

Table 2 describes the publicly available datasets for CSLR. The datasets are compared in terms of sign language, vocabulary size, number of sentences, number of signers, modality, and domain. The available datasets are for six sign languages, namely, German, Chinese, American, Greek, Russian, and Arabic. These sign languages were the subject of the majority of CSLR research, as revealed in Fig. 5, which displays the distribution of CSLR research per sign language. In terms of vocabulary size, the RWTH-PHOENIX-Weather-201 (Phoenix2014) (Koller, Forster, & Ney, 2015) dataset for GSL has the largest vocabulary size, with around 2048 signs. Conversely, the continuous Chinese Sign Language (CSL) (Huang, Zhou, Zhang, Li, & Li, 2018) and FluentSigners-50 (Mukushev et al., 2022) datasets have the largest number of signers, with 50 signers in each. The number of video samples shows an increase over the years, with the recent FluentSigners-50 (Mukushev et al., 2022) having the largest number of video samples, with approximately 43,250 videos. The RWTH-PHOENIX-Weather-2014 (Koller, Forster, & Ney, 2015) (Phoenix2014), SIGNUM (Von Agris & Kraiss, 2007), and CSL (Huang et al., 2018) datasets are the most popular benchmarking datasets for CSLR. The SIGNUM dataset (Von Agris & Kraiss, 2007) was created in 2007 in a controlled environment and contains around 19,000 GSL sentences in general domains. Forster et al. (2012), in 2012, released the largest publicly available CSLR dataset, RWTH-PHOENIX-Weather-2012 (Phoenix2012), for GSL with a vocabulary of around 1389 signs and 6841 sentences. The authors extended the dataset in 2014 (Phoenix2014) (Koller, Forster, & Ney, 2015), doubling the vocabulary size. This dataset is challenging as it is comprised of real-life recordings of weather forecasts. Also, 30% of the vocabulary occurred only once in the training data, having 0.54% out-of-vocabulary (OOV) rate, which are signs occurring in the test data and not present in the training data. The authors also provide a signer-independent split, leaving out signer-5 for testing (Phoenix2014SI). The annotation and sentence segmentation of the Phoenix2014 dataset was refined in another version named RWTH-PHOENIX-Weather-2014-T (Phoenix2014T). The dataset was also extended to include gloss and German annotations for both CSLR and SLT.

The CSL dataset (Huang et al., 2018) was published in 2018 and employed in numerous recent studies. The CSL dataset was also created in a laboratory environment and contains around 100 sentences recorded by 50 signers. There are two versions of the dataset, CSL split I is employed for signer-independent evaluation, where videos of 40 signers are in the training set and videos of 10 signers are kept for testing. The second split (CSL Split II) is utilized for unseen sentences testing, where 6% of the sentences are not seen during training. Even though the CSL dataset has numerous signers, it contains only 100 unique sentences with 178 signs. The authors addressed this issue in another newly released dataset, named CSL-Daily (Zhou, Zhou, Zhou, & Li, 2021), containing 2000 signs and 6598 sentences signed by 10 signers in a laboratory setting.

In contrast to the other datasets, FluentSigners-50 (Mukushev et al., 2022) was not created in a controlled environment. The dataset was crowd-sourced to obtain videos in different environment settings and recorded using various devices, such as webcams and smartphones. This diversity in recording conditions can enhance the generalization of the CSLR systems, leading to elevated performance in recognizing sign gestures in real life (Mukushev et al., 2022). The Scene-PHOENIX (Jang et al., 2022) also addressed realistic CSLR by augmenting the Phoenix2014 dataset with various artificial backgrounds. Their methodology can be applied to other lab-recorded CSLR datasets to create more realistic datasets. offer depth and skeleton information in addition to the RGB data (Duarte et al., 2021; Huang et al., 2018; Luqman, 2023). Multi-modal data sources provide a richer set of features for models to learn from. This enables extracting more discriminative features that capture the subtle details of sign language, which can lead to a better recognition performance. Samples from various types of datasets: real-world (Phoenix2014 dataset), lab created (CSL dataset), and crowd-sourced (FluentSigners-50) are displayed in Fig. 4.

Table 2
Publicly available CSLR datasets.

Dataset	Year	Sign language	Signs	Unique sentences	Signers	Samples	Modality	Domain
Purdue RVL-SLLL ASL (Martínez, Wilbur, Shay, & Kak, 2002)	2002	American	104	600	14	2,576	RGB	Disasters
RWTH-BOSTON-104 (Dreuw, Stein, & Ney, 2009)	2007	American	104	201	3	843	RGB	General
SIGNUM (Von Agris & Kraiss, 2007)	2008	German	450	780	25	19,500	RGB	General
Assaleh, Shanableh, Fanaswala, Amin, and Bajaj (2010)	2010	Arabic	80	40	1	760	RGB	General
RWTH-PHOENIX-Weather (Forster et al., 2012)	2012	German	1,081	2,640	7	2,640	RGB	Weather
RWTH-PHOENIX-Weather-2014 (Koller, Forster, & Ney, 2015)	2014	German	2,048	6,841	9	6,841	RGB	Weather
RWTH-PHOENIX-Weather-2014-T (Camgoz, Hadfield, Koller, Ney, & Bowden, 2018)	2018	German	1,066	8,257	9	8,257	RGB	Weather
CSL (Huang et al., 2018)	2019	Chinese	178	100	50	25,000	RGB, Depth, Skeleton	General
TheRuSLan (Kagirov, Ivanko, Ryumin, Axyonov, & Karpov, 2020)	2020	Russian	164	NA	13	NA	RGB, Depth	Supermarket
ArSL for Deaf Drivers (Abbas, Al-Barhamtoshy, & Alotaibi, 2021)	2021	Arabic	NA	215	3	NA	RGB	Deaf Drivers
LMSLR (Wang & Zhang, 2021)	2021	Chinese	298	100	10	10,000	Skeleton	General
Continuous GrSL (Adaloglou, Chatzis, Papastratis, Stergioulas, & Th, 2021)	2021	Greek	310	331	7	10,295	RGB, Depth	Public services
CSL-Daily (Zhou, Zhou, Qi, Pu, & Li, 2021)	2021	Chinese	2,000	6,598	10	21,000	RGB	General
ArabSign (Luqman, 2023)	2022	Arabic	95	50	6	9,335	RGB, Depth, Skeleton	General
FluentSigners-50 (Mukushev et al., 2022)	2022	Kazakh-Russian	278	173	50	43,250	RGB	General
ASL-Homework (Hassan et al., 2022)	2022	American	NA	NA	45	935	RGB, Depth	General
Scene-PHOENIX (Jang et al., 2022)	2022	German	2,048	6,841	9	6,841	RGB	Weather

Fig. 4. Samples from CSLR datasets. Phoenix2014 (Top), CSL (middle), and FluentSigners-50 (bottom).

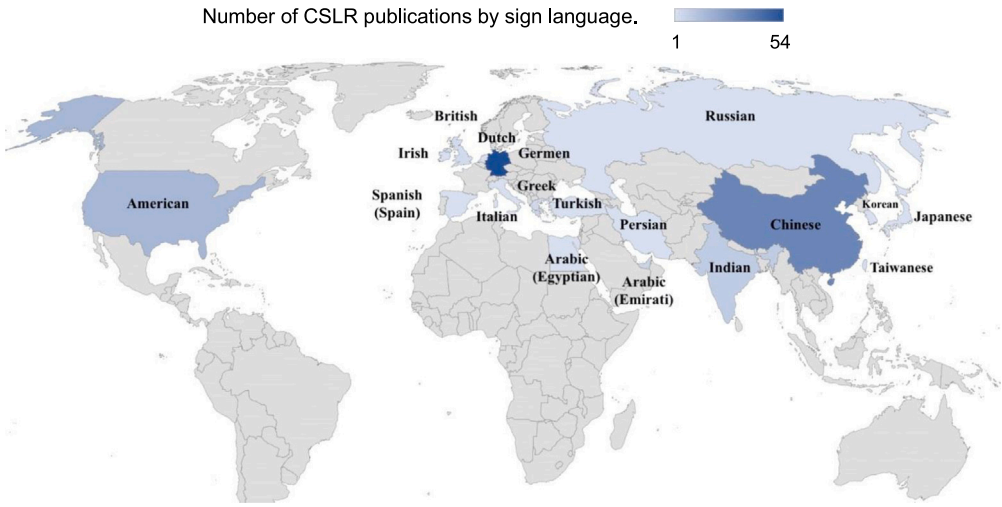

Fig. 5. Heatmap of CSLR research by sign language.

3.3. Acquisition devices

Based on the data acquisition device, SLR approaches can be classified as vision-based (Adaloglou et al., 2021; Albanie et al., 2020; Ananthanarayana et al., 2021; Assaleh et al., 2010; Bauer, Hienz, & Kraiss, 2000; Bauer & Kraiss, 2002; Camgoz, Hadfield, Koller, & Bowden, 2017; Camgoz et al., 2018; Camgöz et al., 2020; Cheng, Yang, Chen, & Tai, 2020; Cortés, García, Benítez, & Segura, 2006; Cui, Liu, & Zhang, 2017, 2019; Cui, Zhang, Li, & Wang, 2023; Dreuw, Rybach, Deselaers, Zahedi, & Ney, 2007; Elakkiya, 2020; Elakkiya & Selvamani, 2019; Forster, Oberdörfer, Koller, & Ney, 2013; Forster et al., 2012; Gao et al., 2021; Gweth, Plahl, & Ney, 2012; Hao, Min, & Chen, 2021; Hienz, Bauer, & Kraiss, 1999; Hu, Gao, Liu, & Feng, 2022, 2023a, 2023b; Huang & Ye, 2021; Huang et al., 2018; Infantino, Rizzo, & Gaglio, 2007; Kelly, Reilly Delannoy, Mc Donald, & Markham, 2009; Koishybay, Mukushev, & Sandygulova, 2020; Koller, Bowden, & Ney, 2016; Koller, Camgoz, Ney, & Bowden, 2020; Koller, Forster, & Ney, 2015; Koller, Zargaran, & Ney, 2017; Koller, Zargaran, Ney, & Bowden, 2016; Min, Hao, Chai, & Chen, 2021; Niu & Mak, 2020; Papastratis, Dimitropoulos, Konstantinidis, & Daras, 2020; Pei, Guo, & Zhao, 2019; Pu, Zhou, Hu, & Li, 2020; Pu, Zhou, & Li, 2018, 2019; Rao, Kishore, Kumar, & Sastry, 2017; Rekha, Bhattacharya, & Majumder, 2012; Roussos, Theodorakis, Pitsikalis, & Maragos, 2010; Sarkar, Loeding, Yang, Nayak, & Parashar, 2011; Slimane & Bouguessa, 2020; Tolba, Samir, & Aboul-Ela, 2013; Tripathi & Nandi, 2015; Vassilia & Konstantinos, 2006; Vogler & Metaxas, 2001; Von Agris, Blömer, & Kraiss, 2008; Von Agris, Knorr, & Kraiss, 2008; Wei et al., 2021; Wei, Zhou, Pu, & Li, 2019; Xiao, Qin, & Yin, 2020; Xie et al., 2023; Xie, Zhao, & Hu, 2021; Yang & Lee, 2011; Yang, Sarkar, & Loeding, 2007, 2010; Yang, Shi, Shen, & Tai, 2019; Yu, Huang, Hsu, Lin, & Wang, 2011; Yuan, Geo, Yao, & Wang, 2002; Zaboli, Serov, Mestetskiy, & Nagendraswamy, 2021; Zadghorban & Nahvi, 2018; Zhang, Pu, Zhuang, Zhou, & Li, 2019; Zhou, Zhou, & Li, 2019; Zhu, Li, Yuan, & Gan, 2023) or sensor-based (Ekiz et al., 2017; Fang, Gao, Chen, Wang, & Ma, 2002; Gao, Fang, Zhao, & Chen, 2004; Guilin, Hongxun, Xin, & Feng, 2006; Hassan, Assaleh, & Shanableh, 2017, 2019; Kong & Ranganath, 2014; Li, Zhou, & Lee, 2016; Liang & Ouhyoung, 1998; Meng et al., 2019; Sharma, Gupta, & Kumar, 2021; Suri & Gupta, 2019; Tubaiz, Shanableh, & Assaleh, 2015; Tuffaha, Shanableh, & Assaleh, 2015; Wang, Gao, & Xuan, 2001; Zhang, Zhang, & Zheng, 2020). In vision-based approaches, a camera is used to capture signs and represent them as videos or images. In sensor-based approaches, sensors, such as data gloves and armbands, are utilized to track and collect sign data. Sensor-based SLR approaches eliminate the need for image pre-processing and, therefore, require less computation. The majority of earlier studies (before 2007) incorporated sensors for data acquisition due to difficulties associated with vision-based recognition, where extensive measures need to be taken to capture the needed features. In addition, sensor-based SLR avoids challenges associated with computer vision recognition, such as occlusions, illumination differences, background complexity and different points of view. However, the main obstacle to this approach is their high-costs and impracticality, given that the signer must wear the sensors during signing, which limits their usage for real-world applications (Rastgoo et al., 2021). Conversely, vision-based approaches are cost-effective and user-friendly, as camera devices are more likely to be available in every household. Based on the conducted survey, the majority of CSLR studies (82%) utilized a vision approach, while 18% of the CSLR studies employed sensors for sign language acquisition.

3.3.1. Vision-based data acquisition

Vision-based CSLR relies on video data as its primary input. These videos capture the signing gestures performed by individuals in a sign language. High-quality cameras capable of capturing a large number of frames per second (fps) are needed to capture the fast sign gestures accurately. Webcams and smartphone cameras also provide a convenient method for sign language capturing since they are available in most households. New smartphones are equipped with high-resolution cameras with 4K resolution, which encouraged researchers to adopt them for sign language capturing. Recently, Mukushev et al. (2022) utilized the front camera of the smartphone to capture sign language gestures. A multi-modality Microsoft Kinect camera is also a popular option for capturing sign data. The Kinect is equipped with two cameras and one infrared (IR) sensor that provide RGB, depth, and skeleton data. Kinect was often employed in the literature to create multi-modal CSLR datasets (Huang et al., 2018; Jebali, Dakhli, & Jemni, 2021; Kagirov et al., 2020; Yang et al., 2016; Zhang, Zhou, & Li, 2014).

3.3.2. Sensor-based data acquisition

Several sensor-based methods have been utilized in the literature to capture sign data, such as data gloves, armbands, and smartwatches. Samples of the sensor devices employed in the literature for capturing sign data are depicted in Fig. 6. We categorize sensor-based CSLR studies according to the used sensor in Table 3. This table reveals that earlier studies (Fang et al., 2002; Gao et al., 2004; Guilin et al., 2006; Hassan et al., 2017, 2019; Kong & Ranganath, 2014; Li et al., 2016; Liang & Ouhyoung, 1998; Tubaiz et al., 2015; Tuffaha et al., 2015; Wang et al., 2001) relied on sensor gloves for capturing sign data. Nowadays, researchers interested in sensor-based CSLR are shifting their attention to less intrusive methods, such as using WiFi signals (Zhang et al., 2020). This section extensively discusses and analyzes sensor-based devices employed for CSLR.

- **Data Gloves.** Glove-embedded sensors were employed to track the motion of hands and fingers for CSLR (Fang et al., 2002; Gao et al., 2004; Guilin et al., 2006; Hassan et al., 2017, 2019; Kong & Ranganath, 2014; Li et al., 2016; Liang & Ouhyoung, 1998; Tubaiz et al., 2015; Tuffaha et al., 2015; Wang et al., 2001). Information provided from the gloves includes, the movement, rotation, and position of the hands and fingers. The DataGlove is a commercial sensor glove developed in the 90s. It is equipped with two sensors on each finger that computes joint flexion. A tracker is placed on the palm to detect position and orientation. The accuracy of the data is highly dependent on how well the glove fits over the finger joints. A sensor glove was first employed for CSLR in 1998 by Liang and Ouhyoung (1998). They utilized one DataGlove worn on the right hand by the signer. Seeing that most signs require the use of both hands, later studies integrated two sensor gloves to

adequately capture the signs. The CyberGlove (Fig. 6(a)) is another type of sensor-embedded glove employed for CSLR (Gao et al., 2004; Guilin et al., 2006; Kong & Ranganath, 2014; Wang et al., 2001). It is more sophisticated compared with DataGlove, as the embedded sensors capture the hand motion patterns more precisely. Furthermore, it displays greater flexibility and more comfort. However, it more costly, with its price ranging up to $18,000 for a single glove. The DG5 Vhand glove is a cheaper alternative, containing five bend sensors and a three-dimensional (3D) accelerometer used for capturing both hand orientation and movements. The DG5 Vhand was employed in several CSRL systems (Hassan et al., 2017, 2019; Tubaiz et al., 2015; Tuffaha et al., 2015). Li et al. (Li et al., 2016) argued that the commercial gloves (such as DataGlove, CyberGlove, and Vhand) were too expensive for real- world applications and opted to create their own custom-made sensor gloves. The prototype glove contained cheaper sensors placed in the center of each finger bone. The cost of a pair of gloves was estimated to be around 150$, which is considerably lower than the commercial gloves. However, this is at the expense of having less accurate data.

- **Armbands.** Sensor-equipped armbands are another popular option in sensor-based CSLR (Sharma et al., 2021; Suri & Gupta, 2019; Tateno, Liu, & Ou, 2020). Electromyography (EMG) sensors are placed on the armband (Fig. 6(b)) to capture muscle movement when performing the corresponding gestures. Tateno et al. (2020) utilized the Myo armband, which has eight EMG sensors for CSLR. The armband was placed under the elbow of the right hand to capture several one-handed sign language gestures. Although this method is less invasive than sensor gloves, EMG sensors are prone to capturing noise when sensing activities from adjacent muscles. In addition, several factors affect the accuracy of the data, such as body temperature, muscle dehydration, and body mass index. Alternatively, some studies suggested using inertial measurement units (IMUs), which contain a gyroscope and an accelerometer to provide motion data, to overcome the limitations of EMG-based armbands. Sharma et al. (2021) employed three IMUs mounted on each forearm using adhesive. Suri and Gupta (2019) used a GY-80 multi-board equipped with five sensors. The board was strapped on the forearm of the signer's dominant hand. However, the armband in their setup was very large, heavily restricting the movement of the signer's hands.
- **Motion Capture Devices.** Wearable sensors (for example, data gloves and armbands) may cause inconvenience since the signer is obligated to wear them during signing. Another alternative is motion capture systems that can capture the signer's hand motion without the need to wear cumbersome sensors. Leap Motion Controller (LMC) is a popular motion-capturing device that provides spatial information of fingers and hands (Fig. 6(c)). The LMC sensor is equipped with two IR cameras capable of sensing a distance of up to one meter. Two studies (Fang, Co, & Zhang, 2017; Mittal, Kumar, Roy, Balasubramanian, & Chaudhuri, 2019) employed LMC for CSLR. The signer performs the signs in front of the LMC, and the skeleton representation of the hands was captured. However, LMC fails to capture occluded hands, which is a problem for SLR since many signs involve hands occlusion during signing. LMC also requires the signer to be very close to the device. In addition, the data provided by the LMC are often noisy and require additional pre-processing.
- **Smartwatches.** Smartwatches are utilized by numerous people to track their health and provide timely notifications (Ekiz et al., 2017). They contain a 3D accelerometer and 3D gyroscope that can help capture hand motion. Ekiz et al. (2017) strapped two Samsung Gear S2 smartwatches on each signer's wrist to capture the Turkish sign language (TuSL). However, the system struggled to provide accurate recognition when tested on unseen signers. This is because of the high variation in the acceleration and gyroscope data amongst the different signers that limits the use of these sensors for CSLR.
- **Signal-based methods.** Signals have been utilized in the literature for sign language capturing. Gestures can be recognized by analyzing their effect on the surrounding WiFi signals. Specifically, Channel State Information (CSI) from the WiFi signals can help identify different actions. Zhang et al. (2020) gathered 100 ASL sentences performed by 30 signers using CSI. The setup includes a laptop (with an antenna as the transmitter) and another laptop (with three antennas as the receiver). The signer stands between the transmitter and the receiver and performs the signed sentences. The experiments revealed acceptable recognition accuracy with an average of 69% accuracy. Nonetheless, the experiments showed that the signer should be at most three meters away from the receiver antenna. Meng et al. (2019) argued that wearable devices are intrusive and WiFi-based data-capturing devices are susceptible to interference. Thus, they advocated utilizing radio frequency signals to capture sign data using a directional antenna, radio frequency identification (RFID) reader, and RFID tag. In this setup, the signer stands between the antenna and the tag and performs the signed sentences. The system was evaluated on nine sentences in CSL. However, the generalization of the proposed system was not evaluated in a signer-independent mode. Another approach was followed in Ye, Lan, Zhang, and Zhang (2020), where a Doppler radar was also utilized for collecting sign language sentences. Velocity data is computed by allowing a signal to bounce off the targeted object and assessing changes in the frequency of the bounced-back signal. Doppler-based radars are capable of detecting movement while overcoming background clutter. The CSLR system (Ye et al., 2020) employed micro-Doppler signatures of the hand motion for CSLR, where the signer gestured in front of the radar within a distance of 10 cm. Notably, only a single hand was used for data collection, which limits the system to single-hand signs.

3.4. Input modality

Three input modalities have been utilized by researchers for CSLR, RGB, depth, and skeleton information, as revealed in Table 4. RGB is the most commonly utilized input for CSLR since it provides a high-resolution detailed visual description of the sign gestures. Besides, the advances in computer vision and DL-based techniques have made it possible to easily extract highly informative features from RGB images. While RGB data is the primary modality utilized for CSLR, other modalities, such as depth and skeleton, can complement the RGB modality and enhance the CSLR performance. Skeleton or human pose data is the second most utilized modality

Table 3
Sensor-based CSLR studies categorized by the sensing method.

Data Acq.	References
Data glove	Fang et al. (2002), Gao et al. (2004), Guilin et al. (2006), Hassan et al. (2017, 2019), Kong and Ranganath (2014), Li et al. (2016), Liang and Ouhyoung (1998), Tubaiz et al. (2015), Tuffaha et al. (2015), Wang et al. (2001)
Armband	Sharma et al. (2021), Suri and Gupta (2019), Tateno et al. (2020)
Smartwatch	Ekiz et al. (2017)
Motion Capture	Fang et al. (2017), Jebali et al. (2021), Kumar, Sastry, Kishore, and Kumar (2018), Mittal et al. (2019), Wang and Zhang (2021)
Radio Frequency	Ye et al. (2020)
Radar Signal	Meng et al. (2019)
WiFi Signal	Zhang et al. (2020)

(a) (b) (c)

Fig. 6. Samples of sensor devices used for sign language capturing (a) CyberGlove (Wang et al., 2001), (b) Myo armband (Tateno et al., 2020), and (c) LMC (El-Alfy & Luqman, 2022).

in the literature for CSLR (Aditya et al., 2022; Brock, Farag, & Nakadai, 2020; Fang et al., 2017; Jebali et al., 2021; Ko, Kim, Jung, & Cho, 2019; Li & Meng, 2022; Mittal et al., 2019; Mocialov, Turner, Lohan, & Hastie, 2017; Wang & Zhang, 2021; Yang et al., 2016; Zhang, Tian, & Huenerfauth, 2016; Zhang et al., 2014; Zhang, Zhou, & Li, 2015; Zhou, Tam, & Lam, 2021, 2022; Zhou, Zhou, Zhou, & Li, 2021; Zuo & Mak, 2022a). Skeletal data encodes joint sequences to represent an abstract skeleton figure of the signer. This eliminates the need for several pre-processing tasks associated with RGB images, such as background removal and hand tracking. In addition, pose features can overcome other computer vision problems, such as occlusions and clutter in RGB images. Skeleton information can be captured using dedicated sensors or extracted from RGB images. Kinect and LMC are the most common devices utilized by researchers to obtain skeleton data. Recently, high-accuracy pose estimation models have been presented, such as MediaPipe (Lugaresi et al., 2019), OpenPose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2019), and MMpose (MMPose Contributors, 2020). These models can extract more pose features compared with sensor-based devices. Moreover, the models are light-weight and can be employed in mobile applications (Lugaresi et al., 2019). OpenPose is the most popular pose estimation method employed in CSLR-related literature, as shown in Table 5. Pose data has been leveraged as the single input for the CSLR frameworks (Fang et al., 2017; Jiao et al., 2023; Ko et al., 2019; Mittal et al., 2019; Wang & Zhang, 2021; Yang et al., 2016; Zhang et al., 2014, 2015). However, other researchers combined it with RGB to obtain more informative features (Aditya et al., 2022; Brock et al., 2020; Chen et al., 2022; Jebali et al., 2021; Li & Meng, 2022; Mocialov et al., 2017; Zhang et al., 2016, 2014; Zhou, Tam, & Lam, 2021; Zhou et al., 2022; Zhou, Zhou, Zhou, & Li, 2021; Zuo & Mak, 2022a).

CSLR researchers made use of the estimated human pose keypoints in different ways. Wang and Zhang (2021) modeled the keypoints as a graph, fed into a Graph Convolutional Network (GCN), whereas, others (Aditya et al., 2022; Mittal et al., 2019) fed 3D pose coordinates as raw data directly to the model. Another approach was to represent the keypoints as heatmap images to eliminate noisy features (Chen et al., 2022). Some researchers did not use the pose data directly; instead, they employed the pose data to identify the face and hand regions, and automatically create cropped images fed into the vision-based recognition system (Zhou, Tam, & Lam, 2021; Zhou et al., 2022; Zhou, Zhou, Zhou, & Li, 2021). Depth information conveys the distance between an object in an image and the capturing device. We observe in the literature that the depth data was always used along with the RGB data, as depth images alone do not portray the details of the sign gestures (Dreuw, Steingrube, Deselaers, & Ney, 2009; Jebali et al., 2021; Ye, Tian, Huenerfauth, & Liu, 2018; Zhang et al., 2016). Notably, in recent years, researchers appear less inclined to incorporate depth data, as it was mostly utilized in studies before 2018. This may be due to the increased use of the Kinect camera at that time, which provided RGB, depth, and pose data. Additional information from the gesture frames has been incorporated to improve the performance of CSLR, such as optical flow. Several studies leveraged optical flow to describe patterns of motion in consecutive frames (Cortés et al., 2006; Cui et al., 2019; Ye et al., 2018). Cui et al. (2019) employed DeepFlow to compute the optical flow, which was fed along with the full frame to a CNN-BLSTM model. The study showed that using both optical flow and RGB data led to around a 2% decrease in WER compared to only using the full-frame RGB data.

3.5. Sign representation

Both manual and non-manual features play a critical role in recognizing sign language gestures. Manual features in sign language refer to the movement of the hands and fingers that form the basis of sign language gestures. These features are characterized by

Table 4
Classification of CSLR studies based on input modality.

Modality	References
RGB	Adaloglou et al. (2021), Albanie et al. (2020), Ananthanarayana et al. (2021), Assaleh et al. (2010), Bauer et al. (2000), Bauer and Kraiss (2002), Camgoz et al. (2017, 2018), Camgöz et al. (2020), Cheng et al. (2020), Cortés et al. (2006), Cui et al. (2017, 2019, 2023), Dreuw et al. (2007), Elakkiya (2020), Elakkiya and Selvamani (2019), Forster et al. (2013, 2012), Gao et al. (2021), Guo et al. (2023), Gweth et al. (2012), Hao et al. (2021), Hienz et al. (1999), Hu et al. (2022, 2023a, 2023b), Hu, Gao, Liu, Pun, and Feng (2023), Hu, Pu, Zhou, Fang, and Li (2024), Hu, Pu, Zhou, and Li (2023), Huang and Ye (2021), Huang et al. (2018), Infantino et al. (2007), Jang et al. (2022, 2023), Kelly et al. (2009), Koishybay et al. (2020), Koller, Bowden, and Ney (2016), Koller et al. (2020), Koller, Forster, and Ney (2015), Koller et al. (2017), Koller, Zargaran, et al. (2016), Min et al. (2021, 2022), Niu and Mak (2020), Papastratis et al. (2020), Pei et al. (2019), Pu et al. (2020, 2018, 2019), Rao et al. (2017), Rekha et al. (2012), Roussos et al. (2010), Sarkar et al. (2011), Slimane and Bouguessa (2020), Tolba et al. (2013), Tripathi and Nandi (2015), Vassilia and Konstantinos (2006), Vogler and Metaxas (2001), Von Agris, Blömer, and Kraiss (2008), Von Agris, Knorr, and Kraiss (2008), Wei et al. (2021, 2019), Xiao et al. (2020), Xie et al. (2023, 2021), Yang and Lee (2011), Yang et al. (2007, 2010, 2019), Yu et al. (2011), Yuan et al. (2002), Zaboli et al. (2021), Zadghorban and Nahvi (2018), Zhang et al. (2019), Zheng et al. (2023), Zhou et al. (2019), Zhu et al. (2023), Zuo and Mak (2022b)
Skeleton	Fang et al. (2017), Jiao et al. (2023), Ko et al. (2019), Mittal et al. (2019), Wang and Zhang (2021), Yang et al. (2016), Zhang et al. (2014, 2015)
RGB + Depth	Dreuw, Steingrube, et al. (2009), Jebali et al. (2021), Zhang et al. (2016)
RGB + Skeleton	Aditya et al. (2022), Brock et al. (2020), Chen et al. (2022), Jebali et al. (2021), Li and Meng (2022), Mocialov et al. (2017), Zhang et al. (2016, 2014), Zhou, Tam, and Lam (2021), Zhou et al. (2022), Zhou, Zhou, Zhou, and Li (2021), Zuo and Mak (2022a, 2022a, 2024)

Table 5
Classification of CSLR studies based on the pose method.

Pose method	References
Kinect	Jebali et al. (2021), Yang et al. (2016), Ye et al. (2018), Zhang et al. (2014, 2015)
Leap Motion	Fang et al. (2017), Jebali et al. (2021), Mittal et al. (2019), Wang and Zhang (2021)
Vicon Motion Capture	Suri and Gupta (2019)
MMPose	Aditya et al. (2022), Chen et al. (2022), Jiao et al. (2023), Wei and Chen (2023), Zhou, Zhou, Zhou, and Li (2021), Zuo and Mak (2022a, 2024)
OpenPose	Ananthanarayana et al. (2021), Brock et al. (2020), Ko et al. (2019), Li and Meng (2022), Mocialov et al. (2017), Wang and Zhang (2021), Zhou, Tam, and Lam (2021), Zhou et al. (2022)

various parameters such as hand shape, hand orientation, movement direction, and movement speed. The accurate recognition of these features is essential for CSLR systems to accurately identify and interpret sign language gestures. Non-manual features in sign language refer to facial expressions, body poses, and head movements that accompany manual signs to convey meaning. Non-manual gestures are also utilized for conveying some linguistic features that cannot be expressed by manual gestures, such as negation and emphasis. These gestures play a critical role in the interpretation of sign language and are often used to disambiguate between similar manual signs (Luqman & El-Alfy, 2021). Earlier approaches of CSLR employed only manual features (Assaleh et al., 2010; Bauer et al., 2000; Bauer & Kraiss, 2002; Cortés et al., 2006; Dreuw et al., 2007; Dreuw, Stein, & Ney, 2009; Elakkiya & Selvamani, 2019; Fang et al., 2017, 2002; Gao et al., 2004; Guilin et al., 2006; Hassan et al., 2017, 2019; Hienz et al., 1999; Infantino et al., 2007; Kelly et al., 2009; Koller, Forster, & Ney, 2015; Kong & Ranganath, 2014; Kumar et al., 2018; Liang & Ouhyoung, 1998; Mittal et al., 2019; Rekha et al., 2012; Roussos et al., 2010; Sarkar et al., 2011; Suri & Gupta, 2019; Tolba et al., 2013; Tripathi & Nandi, 2015; Tubaiz et al., 2015; Tuffaha et al., 2015; Vassilia & Konstantinos, 2006; Vogler & Metaxas, 2001; Von Agris, Blömer, & Kraiss, 2008; Von Agris, Knorr, & Kraiss, 2008; von Agris, Zieren, Canzler, Bauer, & Kraiss, 2008; Wang et al., 2001; Yang & Lee, 2011; Yang et al., 2007, 2010, 2016; Yu et al., 2011; Yuan et al., 2002; Zhang et al., 2014, 2014). In vision-based CSLR, a some frameworks (Koller, Bowden, & Ney, 2016; Koller, Zargaran, et al., 2016; Slimane & Bouguessa, 2020) relied on cropped hand images for only the right hand, assuming that it as the dominant hand. Since most signs require both hands, later studies (Huang et al., 2018; Zhou, Zhou, Zhou, & Li, 2021) incorporated cropped images for both right and left hands. Nowadays, the current trend is to use the full frame image fed into convolutional neural networks (CNN) models for feature extraction. This approach was spearheaded by Koller et al. (2017) in 2017 and was followed by most subsequent CSLR works (Adaloglou et al., 2021; Aditya et al., 2022; Camgöz et al., 2020; Chen et al., 2022; Cheng et al., 2020; Cui et al., 2019, 2019, 2023; Elakkiya, Vijayakumar, & Kumar, 2021; Hao et al., 2021; Hu et al., 2022, 2023a, 2023b; Huang et al., 2018; Koishybay et al., 2020; Koller et al., 2020, 2020; Li & Meng, 2022; Min et al., 2021; Niu & Mak, 2020; Papastratis, Dimitropoulos, & Daras, 2021; Papastratis et al., 2020; Pei et al., 2019; Pu et al., 2020, 2018, 2019; Rao et al., 2017; Slimane & Bouguessa, 2020; Suliman, Deriche, Luqman, & Mohandes, 2021; Wei et al., 2021; Xiao et al., 2020; Xie et al., 2023, 2021; Yang et al., 2019; Zhang et al., 2019; Zhou, Lui, Tam, & Lam, 2020; Zhou, Tam, & Lam, 2021; Zhou et al., 2022, 2019; Zhou, Zhou, Zhou, & Li, 2021; Zhu, Li, Yuan, & Gan, 2022; Zhu et al., 2023).

Several studies emphasized the manual gestures by including cropped hand images along with the full frame in multi-channel models (Camgoz et al., 2017; Cui et al., 2019; Forster et al., 2013; Huang et al., 2018; Koller et al., 2020; Koller, Zargaran, et al.,

2016; Slimane & Bouguessa, 2020; Zhou, Tam, & Lam, 2021; Zhou et al., 2022; Zhou, Zhou, Zhou, & Li, 2021). This approach was proven to increase the recognition accuracy, as seen in the current SOTA (Zhou et al., 2022), where the CA-SignBERT model accepts cropped hand with full frame images. Facial expressions are one of the most important non-manual gestures in sign language, as they help express feelings and emotions. They are also used to express grammar moods, such as raising eyebrows to express wishful thoughts or hypothetical situations in GSL (von Agris et al., 2008). Several studies incorporated facial expressions for CSLR (Forster et al., 2013; Hu et al., 2023a; Koller, Forster, & Ney, 2015; Sarkar et al., 2011; Von Agris, Blömer, & Kraiss, 2008; Von Agris, Knorr, & Kraiss, 2008; von Agris et al., 2008; Yang & Lee, 2011; Zhang et al., 2016; Zhou, Zhou, Zhou, & Li, 2021; Zuo & Mak, 2022a). In earlier studies, statistical models such as Active Appearance Models (AAM) aided in identifying the face region and extract facial features (Forster et al., 2013; Koller, Forster, & Ney, 2015; Von Agris, Blömer, & Kraiss, 2008; Von Agris, Knorr, & Kraiss, 2008; von Agris et al., 2008; Yang & Lee, 2011). Yang and Lee (2011) defined six facial expressions and utilized the extracted face keypoints to classify them using Support Vector Machine (SVM). The identified expression along with the manual features assisted with CSLR. Von Agris, Knorr, and Kraiss (2008) reported that combining manual and face features resulted in approximately a 6% increase in the recognition accuracy. Later studies (Zhou, Zhou, Zhou, & Li, 2021; Zuo & Mak, 2022a) resorted to DL-based pose estimation methods, such as HRnet, to extract face keypoints, guiding the model to focus on the hand and face parts. Other investigations incorporated sensor-equipped devices, such as Kinect camera, to extract the face keypoints (Zhang et al., 2016). Hu et al. (2023a) introduced a correlation network that emphasizes the face and hand regions. Mouthing is an essential part of sign language that is typically used to differentiate between signs with the same manual gestures. For instance, the signs for TODAY and NOW in BSL are identical and can only be differentiated by lip motion and shape (von Agris et al., 2008). This aspect has been studied in multiple studies (Infantino et al., 2007; Koller et al., 2020; Von Agris, Blömer, & Kraiss, 2008; Von Agris, Knorr, & Kraiss, 2008; von Agris et al., 2008). Infantino et al. (2007) incorporated a region-growing algorithm to localize the signer's hand and lip regions. The authors also identified lip pauses, which were utilized for sentence segmentation. Von Agris, Blömer, and Kraiss (2008), Von Agris, Knorr, and Kraiss (2008), von Agris et al. (2008) employed the AAM keypoints to define 11 features describing the lip outline. Koller et al. (2020) also used AAM to localize the mouth region and create cropped mouth images. Three CNN-LSTM models were trained individually for gloss, mouth shape, and hand shape recognition. The extracted features by these models were fed into a multi-channel HMM to produce the final gloss sequence.

Other non-manual features, such as head movement and body pose, were employed by researchers to boost the accuracy of CSLR systems. Head movement typically helps convey grammatical information in sign language, such as negation and questions (Jebali et al., 2021; Sarkar et al., 2011; von Agris et al., 2008). Sarkar et al. (2011) tracked head motion in the video frames to detect negation in sign language. von Agris et al. (2008) proposed a method to estimate the head pose from the AAM extracted face landmarks along with optical flow. Body pose is also employed in interpreting sign language, as the body of the signer acts as a reference in the signing space and can convey additional meanings. For instance, leaning the body motion forward or backward helps in distinguishing between ENTICE and REJECT signs in GSL (von Agris et al., 2008). Several studies employed body pose estimation methods to describe the signer's pose and used it in addition to the full frame to enhance the CSLR system (Aditya et al., 2022; Chen et al., 2022; Li & Meng, 2022; Ye et al., 2018; Zuo & Mak, 2022a). Additionally, line of eye or gaze is also an informative parameter in sign language. Sudden changes in eye focus during sign language conversations can express indirect speech, such as referring to another absent person (von Agris et al., 2008). To our knowledge, von Agris et al. (2008), is the only study that incorporated eye gaze for CSLR, where iris localization was performed, then the line of sight was identified by analyzing the intensities surrounding the pupil. Table 6 summarizes the usage of different combinations of sign language cues for CSLR. Evidently, researchers currently prefer using full image frames to obtain a global representation of the signs. Nonetheless, some sign language cues are subtle and may not be clearly visible, especially in low-resolution images. Consequently, emphasizing the finer details of sign language through cropped hand images (Cui et al., 2019; Forster et al., 2013; Gweth et al., 2012; Huang et al., 2018; Koller, Bowden, & Ney, 2016; Koller et al., 2020; Koller, Zargaran, et al., 2016; Slimane & Bouguessa, 2020; Zhou, Tam, & Lam, 2021; Zhou et al., 2022) or skeleton description (Aditya et al., 2022; Chen et al., 2022; Li & Meng, 2022; Ye et al., 2018; Zuo & Mak, 2022a) has been proven to enhance the recognition capability of CSLR systems.

3.6. Sequence alignment methods

Frame-gloss alignment techniques aim to map a frame or a sequence of video frames to its corresponding gloss. This is a crucial step for recognizing a continuous stream of signs (sign language sentences) since CSLR is a weakly supervised task, and the gloss labels lack the exact temporal location. Hence, frame-gloss alignment is necessary to identify the boundaries between signs and understand the temporal dependencies between them. HMMs have been traditionally utilized for frame-gloss alignment, where each state in the HMM represents a part of the sign sentence, and the model learns the probabilities of transitioning between states. However, recent CSLR approaches use other techniques for frame-gloss alignment, such as CTC loss and reinforcement learning (Aloysius & Geetha, 2020).

CTC aligns input sequences with target sequences without requiring explicit alignment information during training. It is particularly useful for tasks like speech recognition and CSLR where the alignment between the input and output sequences is not known prior. CTC loss is computed by considering all possible alignments and summing up the loss scores. A blank label {—} is utilized to identify labels that are not categorized during the decoding process. Specifically, the blank label acts as a placeholder for any gesture in the input video clip that is not part of the sign language vocabulary. This helps in aligning the input and output sequences, and dynamic programming techniques are employed for the decoding process (Graves, Fernández, Gomez, & Schmidhuber, 2006), as illustrated in Fig. 7. Koller et al. (2017) and Cui et al. (2017) were the first to utilize CTC for CSLR and

Table 6
Classification of CSLR studies based on utilized sign cues.

Cues	References
Full frame	Adaloglou et al. (2021), Aditya et al. (2022), Camgoz et al. (2017), Camgöz et al. (2020), Chen et al. (2022), Cheng et al. (2020), Cui et al. (2019, 2019, 2023), Elakkiya et al. (2021), Hao et al. (2021), Hu et al. (2022, 2023a, 2023b), Hu, Gao, Liu, and Feng (2024), Hu, Pu, et al. (2024, 2023), Huang et al. (2018), Jang et al. (2022), Koishybay et al. (2020), Koller et al. (2020, 2020, 2017), Li and Meng (2022), Min et al. (2021, 2022), Niu and Mak (2020), Papastratis, Dimitropoulos, and Daras (2021), Papastratis et al. (2020), Pei et al. (2019), Pu et al. (2020, 2018, 2019), Rao et al. (2017), Slimane and Bouguessa (2020), Suliman et al. (2021), Wei et al. (2021), Xiao et al. (2020), Xie et al. (2023, 2021), Yang et al. (2019), Zhang et al. (2019), Zhou, Lui, et al. (2020), Zhou, Tam, and Lam (2021), Zhou et al. (2019), Zhou, Zhou, Zhou, and Li (2021, 2021), Zhu et al. (2022, 2023), Zuo and Mak (2022b)
Hand(s)	Assaleh et al. (2010), Bauer et al. (2000), Bauer and Kraiss (2002), Cortés et al. (2006), Dreuw et al. (2007), Dreuw, Stein, and Ney (2009), Elakkiya and Selvamani (2019), Fang et al. (2017, 2002), Gao et al. (2004), Guilin et al. (2006), Hassan et al. (2017, 2019), Hienz et al. (1999), Infantino et al. (2007), Kelly et al. (2009), Koller, Forster, and Ney (2015), Kong and Ranganath (2014), Kumar et al. (2018), Liang and Ouhyoung (1998), Mittal et al. (2019), Rekha et al. (2012), Roussos et al. (2010), Sarkar et al. (2011), Suri and Gupta (2019), Tolba et al. (2013), Tripathi and Nandi (2015), Tubaiz et al. (2015), Tuffaha et al. (2015), Vassilia and Konstantinos (2006), Vogler and Metaxas (2001), Von Agris, Blömer, and Kraiss (2008), Von Agris, Knorr, and Kraiss (2008), von Agris et al. (2008), Wang et al. (2001), Yang and Lee (2011), Yang et al. (2007, 2010, 2016), Yu et al. (2011), Yuan et al. (2002), Zhang et al. (2014, 2014)
Upper Body Pose	Brock et al. (2020), Jiao et al. (2023), Ko et al. (2019), Mocialov et al. (2017), Wang and Zhang (2021), Zhang et al. (2015)
Full frame + Hand(s)	Cui et al. (2019), Forster et al. (2013), Gweth et al. (2012), Huang et al. (2018), Koller, Bowden, and Ney (2016), Koller et al. (2020), Koller, Zargaran, et al. (2016), Slimane and Bouguessa (2020), Zhou, Tam, and Lam (2021), Zhou et al. (2022)
Full frame + Body Pose	Aditya et al. (2022), Chen et al. (2022), Li and Meng (2022), Wei and Chen (2023), Ye et al. (2018), Zuo and Mak (2022a, 2024)
Hands + Face	Forster et al. (2013), Koller, Forster, and Ney (2015), Sarkar et al. (2011), Von Agris, Blömer, and Kraiss (2008), Yang and Lee (2011)
Hands + Mouth	Infantino et al. (2007), Koller et al. (2020), Von Agris, Blömer, and Kraiss (2008), Von Agris, Knorr, and Kraiss (2008)
Hands + Face + Mouth	Jiao et al. (2023), Von Agris, Knorr, and Kraiss (2008)
Hands + Face + Body Pose	Zhang et al. (2016)
Hands + Head Pose	Jebali et al. (2021)
Hands + Gaze + Mouth + Pose	von Agris et al. (2008)
Full frame + Hands + Face + Body Pose	Zhou, Zhou, Zhou, and Li (2021)

most subsequent CSLR models adopt CTC for alignment (Cheng et al., 2020; Cui et al., 2019; Min et al., 2021; Niu & Mak, 2020; Pu et al., 2020, 2019). Nonetheless, CTC suffers from some limitations, such as conditional independence assumption and overfitting. To overcome these limitations, Adaloglou et al. (2021) proposed two modified CTC loss functions, Entropy Regularization CTC (EnCTC) and Stimulated CTC (StimCTC). The Entropy Regularization CTC introduces an entropy regularization factor to prevent overfitting. The StimulatedCTC uses an auxiliary uni-directional RNN, which encodes the sentence's history, thus overcoming the independence assumption. The two losses were combined, composing an Entropy Stimulated CTC loss (EnStimCTC).

Several other methods were introduced to alleviate the overfitting problem associated with CTC. Niu and Mak (2020) presented a gradient stopping scheme and a stochastic frame dropping method to mitigate the CTC overfitting effect. Conversely, Zhu et al. (2022) investigated a multi-level CTC loss, computed in several stages within the network, specifically, after extracting the frame-wise features and the temporal features.

Alternatively, Reinforcement Learning (RL) can be applied to enhance the frame-gloss alignment. RL is a trial-and-error training technique where feedback is given to the learning agent based on its own actions in a certain task. Wei et al. (2021) presented RL for frame-gloss alignment, where the REINFORCE algorithm was employed for detecting the semantic boundary of each sign. The WER inverse was also utilized to reward the agent for estimating the gloss timestamps. Their approach showed promising results, which provided an alternative to using supervised learning for CSLR. However, in the proposed approach, the action search space must be carefully estimated, as wrongfully estimating the pooling size will significantly diminish the system's performance.

4. CSLR approaches

Several techniques have been explored in CSLR-related literature. These approaches can be classified based on the feature extraction and learning techniques into two categories: traditional and DL-based approaches. Traditional approaches depend on hand-crafted features extracted from sign gesture video or images. In contrast, DL-based approaches implicitly learn the needed features for CSLR.

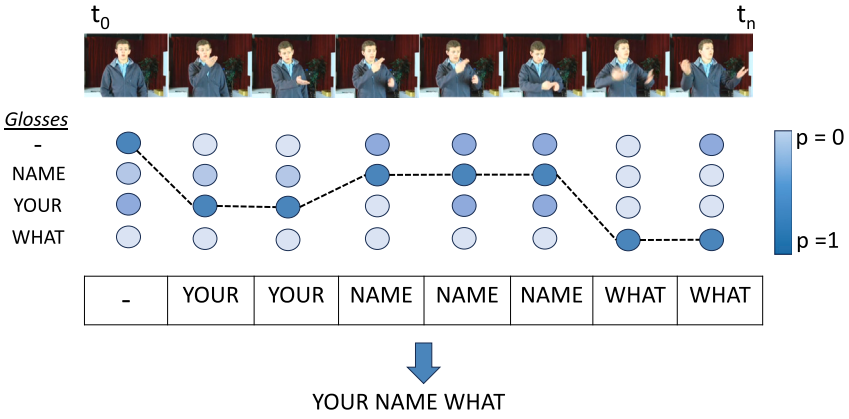

Fig. 7. Simplified overview on CTC alignment. The predicted gloss sequence is computed by taking the most likely gloss per time-step (a darker circle indicates higher probability). Repeated glosses and blanks {—} are removed to produce the gloss sequence. The dashed line represents the predicted gloss path.

4.1. Traditional approaches

Most CSLR-based research before 2015 adopted traditional approaches for CSLR (Assaleh et al., 2010; Bauer et al., 2000; Bauer & Kraiss, 2002; Cortés et al., 2006; Fang et al., 2002; Gao et al., 2004; Guilin et al., 2006; Hienz et al., 1999; Liang & Ouhyoung, 1998; Vassilia & Konstantinos, 2006; Vogler & Metaxas, 2001; Von Agris, Knorr, & Kraiss, 2008; Wang et al., 2001; Yang et al., 2007, 2010; Yu et al., 2011; Yuan et al., 2002; Zhang et al., 2014), as summarized in Table 7. Handcrafted CSLR features were extensively used in these studies, such as Histogram of Oriented Gradients (HOG) (Buehler, Everingham, & Zisserman, 2009), frequency domain features (AL-Rousan, Assaleh, & Tala'a, 2009), and Fourier descriptors (Chan-Wah & Surendra, 2002; Pan et al., 2016). To model the sequence of signs, HMM (Bauer et al., 2000; Hienz et al., 1999; Liang & Ouhyoung, 1998), Dynamic Time Warping (DTW) (Gao et al., 2004; Tripathi & Nandi, 2015; Yang et al., 2010) and Conditional Random Field (CRF) (Choudhury et al., 2017; Kong & Ranganath, 2014) were primarily adopted.

4.1.1. Hidden Markov model (HMM)

The success of HMM in speech recognition encouraged researchers to adopt it for CSLR (Aloysius & Geetha, 2020). HMM is a probabilistic model employed for CSLR to find the optimal sequence amongst a chain of signs corresponding to a sentence. HMM was first utilized for CSLR in Liang and Ouhyoung (1998), which is the earliest study on CSLR, dating back to 1998. It was employed to recognize 196 sentences in Taiwanese sign language (TSL) collected using DataGlove. Multiple studies (Fang et al., 2002; Guilin et al., 2006; Wang et al., 2001) adopted the same approach using HMM to recognize sign data obtained from sensor-equipped gloves. Alternatively, Hienz et al. (1999) introduced a vision-based approach using HMM to recognize GSL signs, and expanded their research in Bauer et al. (2000) using colored gloves worn by the signer to facilitate hand detection and tracking.

In light of the success of HMMs with the vision-based approach, several HMM-based CSLR systems were developed for various sign languages, such as Greek (GrSL) (Vassilia & Konstantinos, 2006), Spanish (SSL) (Cortés et al., 2006), Arabic (ArSL) (Assaleh et al., 2010), TSL (Yu et al., 2011), Italian sign language (ItSL) (Infantino et al., 2007), and Irish (IrSL) (Kelly et al., 2009). However, most of the HMM models were trained to recognize a limited number of signs using a private collection of data. Dreuw et al. (2007) presented the first publicly available CSLR dataset named RWTH-BOSTON-104 with a vocabulary of 104 signs in ASL. A WER of 17.9% was achieved by combining two HMM models trained independently on hand trajectory and hand velocity. The authors included depth data in another study (Dreuw, Steingrube, et al., 2009) and reported a WER of 19.6%. Subsequently, two GSL public datasets were released, SIGNUM and Phoenix2014 datasets. Von Agris, Blömer, and Kraiss (2008), Von Agris, Knorr, and Kraiss (2008), von Agris et al. (2008) presented the SIGNUM dataset that contained 450 signs in GSL. The authors focused on signer-independent CSLR using HMM and eigen-voice adaptation. In the following years, Koller, Forster, and Ney (2015) presented the Phoenix2014 dataset for GSL. A statistical approach was proposed using HMM and maximum likelihood linear regression, setting a 53% WER on Phoenix2014. Furthermore, HMMs have been adapted to model manual and non-manual parameters like body pose and facial expressions in a multi-stream HMM (Forster et al., 2013, 2013; Vogler & Metaxas, 2001), enhancing recognition accuracy. In contrast, using Kinect and LMC for feature extraction, Jebali et al. (2021) achieved a 95.1% accuracy in recognizing French Sign Language (FSL) sentences, though the dataset was limited to 33 signs. Some researchers (Bauer & Kraiss, 2002; Elakkiya & Selvamani, 2019; Infantino et al., 2007; Roussos et al., 2010; Yuan et al., 2002) argued that decomposing signs into subunits can lead to better recognition performance when scaling up the vocabulary size. Following related research on speech recognition, several HMM models were trained for subunit recognition using approach like K-means (Infantino et al., 2007; Roussos et al., 2010). Nonetheless, the independence assumption of HMM leads to difficulties in capturing complex sign language features and global context, impacting CSLR accuracy.

4.1.2. Dynamic time warping (DTW)

DTW is a dynamic programming (DP) technique that finds the similarity amongst sequences with different speeds and lengths (Aloysius & Geetha, 2020). DTW was employed in several CSLR studies (Ekiz et al., 2017; Gao et al., 2004; Tripathi & Nandi, 2015; Yang et al., 2010; Zhang et al., 2014). A CyberGlove was utilized for data collection in Gao et al. (2004), Tuffaha et al. (2015) and modeled using DTW. However, the study (Gao et al., 2004) made an invalid assumption that movement epenthesis is similar in different CSL sentences, which affected the framework's performance. Besides, DTW was leveraged in multiple vision-based CSLR systems. Yang et al. (2007) applied a level building (LB) algorithm using DP for sign segmentation in ASL. They enhanced the LB approach in Yang and Lee (2011) using nested DP with DTW. The authors in another study (Yang et al., 2010) integrated the Level Building algorithm with HMM instead of DTW and obtained a better performance. To leverage the strengths of both DTW and HMM, Zhang et al. (2014, 2015) introduced a DTW- HMM framework, where DWT was used for segmentation and HMM for classification. Alternatively, K-Nearest Neighbor (KNN) was employed to classify the DTW-segmented sentences in Hassan et al. (2017, 2019), Tubaiz et al. (2015) and Zadghorban and Nahvi (2018). Although DTW was successfully applied in multiple studies, DTW is sensitive to variations in sign speed. DTW also struggles when presented with a large vocabulary and unseen sentences. (Aloysius & Geetha, 2020).

Other traditional techniques have been presented in CSLR-related literature, such as graph modeling (Tolba et al., 2013) and CRF (Choudhury et al., 2017; Kong & Ranganath, 2014). Graph modeling was applied for recognizing sentences of ArSL (Tolba et al., 2013) and ISL (Kumar et al., 2018), where sentences are modeled as a connected graph, and a graph matching algorithm is applied to detect the signed sentences. Additionally, CRFs have been used in CSLR as a probabilistic modeling technique to capture the temporal dependencies in sign language gestures (Choudhury et al., 2017; Kong & Ranganath, 2014). CRFs provide a powerful alternative to HMM, especially given their ability to capture both local and global contexts in sign language. Nonetheless, CRF-based CSLR systems can face challenges in handling large sign vocabularies and require substantial amounts of annotated data (Aloysius & Geetha, 2020).

4.2. Deep learning approaches

After 2015, most CSLR researchers abandoned hand-crafted features and explored deep learning techniques for feature extraction. The breakthrough of CNNs encouraged researchers to adopt deep CNNs to effectively extract frame level features to enhance the performance of CSLR models. These features were fed into different temporal learning models, such as HMM and RNN. Recently, researchers started to explore other techniques for CSLR, such as GCN and Transformers. In the following subsections, we categorize and review the CSLR works based on the most used architectures, such as CNN-HMM (Koller, Bowden, & Ney, 2016; Koller et al., 2020; Koller, Ney, & Bowden, 2015; Koller, Zargaran, et al., 2016), CNN-RNN (Camgoz et al., 2017; Cui et al., 2017, 2019; Fang et al., 2017; Koller et al., 2017; Pu et al., 2020; Sharma et al., 2021), 3DCNN (Albanie et al., 2020; Pu et al., 2018, 2019; Yang et al., 2019), GCN (Wang & Zhang, 2021) and Transformer networks (Ko et al., 2019; Niu & Mak, 2020; Slimane & Bouguessa, 2020; Zhang et al., 2019). A summary of the reviewed DL-based CSLR approaches is presented in Table 8.

4.2.1. CNN and HMM

Given the success of HMM in past CSLR studies, researchers were still inclined to adopt HMMs to model the sign sequences. However, the increasing popularity of neural networks steered researchers into adopting learned features instead of using traditional hand-crafted features. Therefore, several studies (Brock et al., 2020; Gweth et al., 2012; Koller, Bowden, & Ney, 2016; Koller et al., 2020; Koller, Ney, & Bowden, 2015; Koller, Zargaran, et al., 2016) relied on neural networks for feature extraction, followed by HMM to model and recognize the sign sequence. Gweth et al. (2012) led the way, where a Multi-Layer Perceptron (MLP) was built to extract features from a sequence of sign language images. The extracted features were fed into a Gaussian HMM (GHMM) to obtain the sign labels. Nonetheless, subsequent studies mostly adopted CNNs as they are better suited for spatial feature extraction. Koller, Ney, and Bowden (2015) was the first to use CNNs to extract features for CSLR, where a CNN-HMM model was proposed for mouth shape learning. In another work (Koller, Zargaran, et al., 2016), the authors proposed a CNN-HMM model named "Deep-Sign" for CSLR. The authors extended their work in Koller, Bowden, and Ney (2016), where they used HamNoSys annotations to model cross-language subunits based on hand orientations. However, the system requires gloss, hand, and mouth shape annotations to train the multi-stream CNN-LSTM-HMM model.

4.2.2. CNN and RNN

The limitations of HMMs in capturing the global context in sign language sentences prompted a shift towards Recurrent Neural Networks (RNNs) in CSLR. RNNs, designed to handle long dependencies in sequences, brought new capabilities to CSLR. Long Short-Term Memory (LSTM), a type of RNN, was introduced to address issues like the vanishing gradient problem commonly found in traditional RNNs. One significant advancement was the adoption of Bidirectional RNNs, particularly Bidirectional LSTMs (BLSTMs), which allowed information to flow from both past and future observations in the sequence. This bidirectional information flow enabled the learning of more complex relationships in CSLR. BLSTM was first utilized for CSLR in Camgoz et al. (2017) and Koller et al. (2017), where a CNN-BLSTM model with CTC was proposed. Researchers explored various strategies to enhance the performance of CNN-BLSTM models in CSLR. For instance, Camgoz et al. introduced SubUNets, combining CNNs with BLSTMs and CTC (Camgoz et al., 2017). Their work highlighted the effectiveness of combining two SubUNets trained on cropped hand and full-frame images. To address shortcomings in feature extraction, an iterative training strategy was proposed (Cui et al., 2017; Koller et al., 2017). This strategy involved training the network using an expectation maximization-like method and fine-tuning the

Table 7
Summary of the surveyed traditional approaches for CSLR. Results are in accuracy, unless indicated otherwise.

Ref.	Year	Language	Data Acq.		Method	Dataset	Result
			Vision	Sensor			
Liang and Ouhyoung (1998)	1998	TSL		✓	HMM	196 sentences	80.4%
Yu et al. (2011)	2011		✓		HMM	40 signs, 3 sentences	67.0%
Hienz et al. (1999)	1999	GSL	✓		HMM	52 signs	95.0%
Bauer et al. (2000)	2000		✓		HMM	97 signs	91.7%
Bauer and Kraiss (2002)	2002		✓		HMM, K-means	12 signs	80.8%
Von Agris, Blömer, and Kraiss (2008)	2008		✓		HMM	SIGNUM	75.8%*
Von Agris, Knorr, and Kraiss (2008)	2008		✓		HMM	SIGNUM	87.4%
von Agris et al. (2008)	2008		✓		HMM	100 Sentences	87.7%
Forster et al. (2013)	2013		✓		Multi-stream HMM	SIGNUM	10.7% WER
Koller, Forster, and Ney (2015)	2015		✓		HMM	Phoenix2014	53.0% WER
Koller, Forster, and Ney (2015)	2015		✓		HMM	SIGNUM	16.4% WER
Elakkiya and Selvamani (2019)	2019		✓		HMM	Phoenix2014	88.1%
Vogler and Metaxas (2001)	2001	ASL	✓		Parallel HMMs	22 signs	84.8%
Dreuw et al. (2007)	2007		✓		HMM	RWTH-BOSTON-104	17.9% WER
Yang et al. (2007)	2007		✓		DP	25 sentences	83.0%
Dreuw, Steingrube, et al. (2009)	2009		✓		HMM, PCA	RWTH-BOSTON-104	19.6% WER
Yang et al. (2010)	2010		✓		DTW	Perdue (10 sentences)	80.0%
Roussos et al. (2010)	2010		✓		K-means	400 signs, 843 sentences	82.0%
Yang and Lee (2011)	2011		✓		CRF-SVM	98 sentences	NA
Sarkar et al. (2011)	2011		✓		HMM	25 sentences	19.0% WER
Kong and Ranganath (2014)	2014			✓	CRF-SVM	107 signs, 74 sentences	86.6%
Zhang et al. (2016)	2016		✓		Bag-of-Words, K-means, SVM	27 signs	36.0%
Wang et al. (2001)	2001	CSL		✓	HMM	100 sentences	90.0%
Yuan et al. (2002)	2002		✓		HMM	40 sentences	70.0%
Fang et al. (2002)	2002			✓	SRN-HMM	100 sentences	85.0%
Gao et al. (2004)	2004			✓	DTW	1500 sentences	90.8%
Guilin et al. (2006)	2006			✓	HMM	543 sentences	70.2%
Zhang et al. (2014)	2014		✓		DTW-HMM	180 sentences	82.2%
Zhang et al. (2015)	2015		✓		DTW	180 sentences	85.2%
Yang et al. (2016)	2016		✓		HMM	20 sentences	12.2% WER
Li et al. (2016)	2016			✓	HMM	510 signs, 1,024 sentences	87.4%
Meng et al. (2019)	2019			✓	RF	9 sentences	98.1%
Vassilia and Konstantinos (2006)	2006	GrSL	✓		HMM	71 signs	89.0%
Cortés et al. (2006)	2006	SSL	✓		HMM	33 signs	99.5%
Infantino et al. (2007)	2007	ItSL	✓		SMO	80 sentences	82.5%
Kelly et al. (2009)	2009	IrSL	✓		Multi-channel HMM	160 sentences	95.0%
Assaleh et al. (2010)	2010	ArSL	✓		HMM	80 signs, 40 sentences	94.0%
Tolba et al. (2013)	2013		✓		Graph Matching	100 signs, 30 sentences	80.0%
Tuffaha et al. (2015)	2015		✓		Polynomial classifier	80 signs, 40 sentences	85.0%
Tubaiz et al. (2015)	2015			✓	KNN	80 signs, 40 sentences	98.9%
Hassan et al. (2017)	2017			✓	KNN	80 signs, 40 sentences	97%
Hassan et al. (2019)	2019			✓	KNN	80 signs, 40 sentences	97.7%
Rekha et al. (2012)	2011	ISL	✓		SVM	NA	93.2%
Tripathi and Nandi (2015)	2015		✓		DTW	11 sentences	90.0%
Kumar et al. (2018)	2018			✓	Graph Matching	200 signs	98.3%
Zadghorban and Nahvi (2018)	2018	PSL	✓		HMM, KNN-DTW	300 sentences	93.0%
Ekiz et al. (2017)	2017	TuSL		✓	DTW, LR	13 sentences	97.6%
Jebali et al. (2021)	2021	FSL		✓	HMM	33 signs	93.8%

*Result using signer-independent evaluation.

feature extractor with pseudo labels produced by the network. The resulting model, named ReSign (Koller et al., 2017), combined CNN-BLSTM with HMM and significantly reduced the WER by 12% compared to their earlier model, Deep-Sign (Koller, Zargaran, et al., 2016), on the Phoenix2014 dataset. Camgoz et al. (2017) presented SubUNets, which consists of 2DCNN-BLSTM. The study revealed that combining two SubUNets trained with cropped hand images and the full frame resulted in the best performance. The authors (Pu et al., 2020) presented a Cross Modal Augmentation (CMA) method, where pseudo pairs of text-video were composed based on delete/substitute/add operations done on both the text label and corresponding frames in the video. However, this approach required additional efforts for the manual creation of the pseudo labels. The framework was enhanced using Prior Aware CMA (PA-CMA) (Hu, Pu, et al., 2024) by leveraging a language model to suggest the pseudo labels. To further improve feature extraction, Cui et al. (2019) suggested training the feature extractor using gloss-level alignment proposals. This approach led to a 4% reduction in WER on the Phoenix2014 dataset, compared to the ReSign model (Koller et al., 2017). While many CSLR models focused on

GSL or CSL, some researchers utilized CNNs and RNNs to recognize other sign languages with smaller private datasets, such as Dutch Sign Language (DSL) (Mocialov et al., 2017) and Japanese Sign Language (JSL) (Brock et al., 2020). Besides vision-based approaches, CNN-RNNs were employed to model sensor data gathered from LMC (Fang et al., 2017) and sensor gloves (Sharma et al., 2021). Fang et al. (2017) proposed a hierarchical bidirectional RNN (HB-BRNN) to model the skeleton data. While Sharma et al. (2021) investigated building CSLR models pre-training on isolated SLR data. This approach, however, was only evaluated on a small private set of 40 sentences in ISL.

4.2.3. Cnn, temporal convolutions and RNN

Temporal Convolutions (TempConvs) can capture sequential information in a time series using 1DCNNs and dilations. Initially introduced by Cui et al. (2017), TempConvs were integrated with a pretrained VGG model and a BLSTM layer for enhanced temporal feature encoding of sign sequences, obtaining a WER of 38.7% on the Phoenix2014 dataset. The approach, however, was limited by its reliance on images of the signer's right hand only, overlooking the involvement of both hands in sign language communication. Subsequent studies (Cui et al., 2019; Hao et al., 2021; Hu et al., 2022, 2023a, 2023b; Jang et al., 2023; Min et al., 2021, 2022; Papastratis et al., 2020; Zhou, Zhou, Zhou, & Li, 2021) built on this foundation, employing a 2DCNN-1DCNN framework for spatio-temporal encoding. Papastratis et al. (2020) proposed integrating an RNN language model to process gloss sequences alongside a 2DCNN-TempConv-BLSTM for video frames. Min et al. (2021) presented Visual Alignment Constraint (VAC) loss as an alternative to previous staged optimization methods, which reduced overfitting the sequence model with less training time. Further, RadialCTC (Min et al., 2022) was introduced to mitigate CTC's peaky behavior, achieving further performance improvements.

Efforts to refine VAC (Min et al., 2021) included Self-Mutual Knowledge Distillation (SMKD) (Hao et al., 2021) for simultaneous visual and temporal module training and a Temporal Lift Pooling (TLP) (Hu et al., 2022), which replaced the max pooling layers in VAC, improving the WER by 2% on Phoenix2014. Several studies aimed to enhance spatial attention towards informative regions in the frames (Hu et al., 2023a, 2023b; Jang et al., 2023). The Self-Emphasizing Network (SEN) (Hu et al., 2023b) utilized a modified ResNet to generate attention maps towards informative regions, and Divide and Focus Convolution (DFConv) (Jang et al., 2023) was presented to process the frame's top and bottom parts separately, thus forcing the model to pay attention to both manual and non-manual cues. Correlation Network (CorrNet) (Hu et al., 2023a) computed correlation maps of body trajectories between adjacent frames, achieving SOTA WERs across multiple datasets using only RGB data. Alternatively, Spatial-Temporal Multi-Cue (STMC) (Zhou, Zhou, Zhou, & Li, 2021) presented a multi-modal approach, which integrated multiple features (pose, face, hands, and full frame). The framework was jointly trained for CSLR and translation tasks. However, the multi-modal framework delivered modest behavior compared to the RGB based CorrNet (Hu et al., 2023a). Jointly training CSLR frameworks for several signs languages in a collaborative manner was investigated in Hu, Pu, et al. (2023), where a shared visual backbone was utilized along with separate paths to model each sign language. Limited WER improvement was obtained on Phoenix2014 with 20.9% WER, despite training the network on Phoenix2014, CSL and GrSL-SD datasets. Recently, Few efforts were directed to optimize model efficiency without significantly compromising performance, such as AdaBrowse (Hu, Gao, et al., 2023), AdaSize (Hu, Gao, et al., 2024), and Temporal Super Resolution (TSRNet) (Zhu et al., 2023). These models illustrated the potential of dynamic resolution adjustments in real-time applications.

4.2.4. 3DCNN

3DCNNs have emerged as a powerful tool for spatio-temporal feature extraction, making them particularly well-suited for applications like CSLR. Pu et al. (2018) introduced a 3DResNet-TempConv model trained using an iterative optimization approach. The authors enhanced the framework by including an attention window and achieved a WER reduction of around 1% (Pu et al., 2019). A similar training approach using pseudo labels for iterative optimization was proposed in Pei et al. (2019). The method was employed to train a 3DResNet-BGRU model. Zhou et al. (2019) also utilized 3DCNN-BGRU trained iteratively with pseudo labels. An I3D model was used instead of 3DResNet, which resulted in a significant reduction of 6% in WER compared to Pei et al. (2019). Wei et al. (2019) suggested using an n-gram and a word-based classifier with 3DResNet-BLSTM, yielding a WER of 50.9% for unseen sentence recognition using the CSL dataset. The study (Yang et al., 2019) presented the Structured Feature Network (SF-Net), where features were extracted at three levels: frame, gloss, and sentence, using 2D/3D convolutions-LSTM-BLSTM. Huang and Ye (2021) argued that the performance of the CSLR models drastically degrades when encountering long sequences. To overcome this issue, the study proposed Boundary-Adaptive encoder (BAE), showing significant improvement on the CSL dataset. Adaloglou et al. (2021) conducted a comparative study on several prominent CLSR models including I3D-BLSTM and 3DResNet-BLSTM (Pu et al., 2019). The models were evaluated on three datasets Phoenix2014, CSL, and a newly presented dataset for GrSL, concluding that pre-training the CSLR models using an isolated sign language dataset enhanced the recognition results. Multi-stream 3DCNN networks were explored in Chen et al. (2022) and Huang et al. (2018). Huang et al. (2018) proposed using a two-stream 3DCNN (C3D) with Hierarchical Attention Network and Latent Space (LS-HAN) fed with hand and full frame images. More recently, Chen et al. (2022) presented a two-stream network that takes the full frame and pose heatmaps as inputs encoded separately by a S3D model (Xie, Sun, Huang, Tu, & Murphy, 2018), achieving SOTA results on several datasets. Wei and Chen (2023) presented a cross-lingual CSLR framework by using similar signs from various sign languages, thus improving training with additional data. The two-stream model (Chen et al., 2022) was adopted to test the framework, achieving new SOTA WERs of 16.7% and 18.6% on Phoenix2014 and Phoenix2014T, respectively. Despite its effectiveness, the method's complexity lies in creating sign dictionaries and training isolated SLR models to map signs between languages towards obtaining extra labeled data.

4.2.5. Fully convolutional network

Fully Convolutional Networks (FCN) are based solely on convolution layers and do not use dense layers, which results in faster training. Cheng et al. (2020) argued that CSLR systems that employ RNN are inefficient for online CSLR recognition, as they require reading the full frame sequence to output the equivalent gloss sequence. An FCN with a CTC decoder was introduced overcome these challenges. The FCN yielded 23.9 and 3% WERS on the Phoenix2014 and CSL datasets, respectively. The same approach was followed in Zhou, Ng, Cai, and Cheung (2020). However, an FCN followed by a self-attention network was proposed. Although the model employed a self-attention layer to capture global features, it failed to identify the relevant features, which resulted in around an 8% increase in WER compared to the FCN (Cheng et al., 2020). Xie et al. (2023) built on the FCN model (Cheng et al., 2020) by proposing a Multi-scale Local-Temporal Similarity Fusion Network (mLTSF-Net). The proposed method was designed to tackle gloss length variations by adaptively fusing temporally similar features, and successfully enhanced the accuracy of the FCN (Cheng et al., 2020) by around 1%.

4.2.6. GCN

GCNs are neural networks designed to model graph-structured data. GCNs were utilized in Li and Meng (2022) and Wang and Zhang (2021) to learn latent connections between the skeleton joints in sign language. Wang and Zhang (2021) presented a Multi-Stream Spatial-Temporal GCN (Multi-Stream ST-GCNs) for CSL recognition. The proposed method utilized skeleton information from hand, pose, and face. in Li and Meng (2022), a similar approach was followed to model skeleton data. However, contrary to the 2D-GCN presented in Wang and Zhang (2021), a 3D-GCN was introduced to capture both spatial and temporal information. Compared to the 2D-GCN (Wang & Zhang, 2021), the 3D-GCN based model yielded slightly worse performance on the CSL dataset with around a 0.3% increase in WER. This questions the contribution of the 3D data for CSLR, especially with the increase in model complexity. More recently, ST-GCNs were utilized to model skeleton data from the face, hands, and body separately in the Co-Sign framework (Jiao et al., 2023). The proposed model outperformed the SMKD model (Hao et al., 2021) by 0.1% while leveraging only skeleton data.

4.2.7. Transformer network

Transformers (Vaswani et al., 2017) have revolutionized the field of machine learning by introducing the self-attention mechanism. Transformers were utilized for sequence learning in several CSLR frameworks (Camgöz et al., 2020; Guo et al., 2023; Niu & Mak, 2020; Papastratis, Dimitropoulos, & Daras, 2021; Slimane & Bouguessa, 2020; Xie et al., 2021; Zhang et al., 2019; Zhou, Tam, & Lam, 2021; Zhou et al., 2022; Zhu et al., 2022; Zuo & Mak, 2022a, 2022b, 2024). Zhang et al. (2019), integrated 3DResNet into a Transformer encoder decoder architecture, showcasing the robustness of Transformers as sequence learners. Niu and Mak (2020) proposed Stochastic Fine-Grained Labeling (SFL) to enhance sequence alignment in a ResNet-Transformer model. A modified self-attention method called Local Context-Aware Transformer Encoder (LCTE) was presented to enhance standard Transformers (Zuo & Mak, 2022b), showing noteworthy improvement over vanilla Transformer. Another study (Zhu et al., 2022) combined ResNet with a Multi-Scale Temporal Network (MSTNet), which utilized 1DCNNs with varied receptive fields followed by a Transformer encoder, resulting in a notable reduction in WER. Multi-modal Transformer-based models were investigated in Slimane and Bouguessa (2020) and Zuo and Mak (2022a, 2024). The study (Slimane & Bouguessa, 2020) utilized cropped hand and full frame images encoded by 2DCNN-Transformer. Alternatively, pose heatmaps were leveraged to enforce spatial attention in a VGG11-Transformer model named (C2SLR) (Zuo & Mak, 2022a). The framework was extended with the Signer Removal Method (SRM) to enhance signer-independent CSLR, achieving new SOTA results on the CSL signer independent dataset (Zuo & Mak, 2024). The multi-modal SignBERT (Zhou, Tam, & Lam, 2021) encoded frames along with cropped hands images using a ResNet-BERT model. This model was further enhanced through dynamic weighting schemes in CA-SignBERT (Zhou et al., 2022), leading to significant performance gains. The prospects of leveraging textual information to train CSLR models were investigated in Guo et al. (2023) and Zheng et al. (2023). A BERT language model trained using the gloss sequence was implemented in Guo et al. (2023) to improve the contextual BLSTM-based module. Similarly, Zheng et al. (2023) proposed to pre-train a language model comprised of self-attention and BLSMT using the gloss sequences of the training data to enhance contextual encoding and alignment. However, the language model contributed to limited performance gain with around 0.6 WER improvement. Moreover, Transformers were leveraged for joint CSLR-SLT systems using the sign2gloss2text approach (Camgöz et al., 2020; Papastratis, Dimitropoulos, & Daras, 2021; Xie et al., 2021), where the gloss predictions obtained from the CSLR module were fed into an SLT module to produce spoken language translations. SLRGAN (Papastratis, Dimitropoulos, & Daras, 2021) is composed of a GAN to produce glosses, which were translated into text using a Transformer model. Camgöz et al. (2020) introduced jointly trained Transformers for CSLR and SLT. The framework was refined by modifying the self-attention mechanism and used relative position encoding, slightly improving the performance (Xie et al., 2021).

4.2.8. Vision transformer

Vision Transformers (ViT) treat an image as a sequence of patches, positional encodings are added, and the obtained sequences are encoded by standard Transformer encoders (Li & Meng, 2022). Few CSLR investigated ViTs for visual features extraction (Cui et al., 2023; Li & Meng, 2022). A two-stream model using RGB and pose data was introduced in Li and Meng (2022), where a pretrained ViT was utilized to extract visual features from RGB frames, and an Attention-enhanced Multi-scale 3DGCN (AM3D-GCN) was designed to encode the OpenPose features. The model was evaluated on the CSL and Phoenix2014T datasets, reporting WERs of 1.9% and 22.8%, respectively. Contrary to the previous Transformer-based models, a fully Transformer-based framework, named Spatial Temporal Transformer (ST-Transformer), was presented in Cui et al. (2023). The ST-Transformer achieved new SOTA

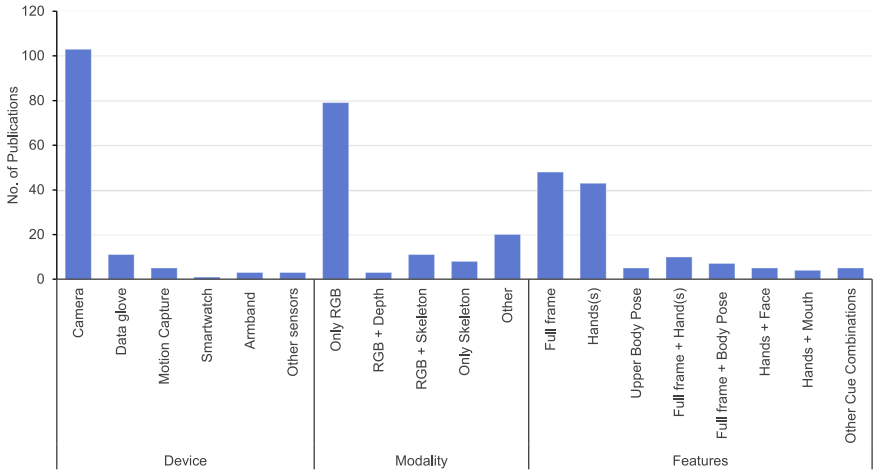

Fig. 8. Analysis of CSLR approaches in terms of input modality, data acquisition, and features.

result on the CSL dataset with 1.2% WER. Zhang, Guo, Yang, Liu, and Hu (2023) utilized Swin Transformer (SwinT), which is an enhanced ViT equipped with shifted window multi-head self attention. The authors proposed the Cross-modal Contextualized Sequence Transduction (C2ST) framework, which leveraged textual information from the gloss sequence using BERT language model. The framework outperformed previous SOTA methods, obtaining WERs of 17.7% and 18.9%, on the Phoenix2014 and Phoenix2014T datasets, respectfully.

4.2.9. Other methods

Some less-popular methods were presented in the literature for CSLR, such as Transducer (Gao et al., 2021), Capsule Network (CapsNet) (Suri & Gupta, 2019), and Deep Belief Network (DBN) (Zhang et al., 2020). RNN-Transducer is a sequence-to-sequence model that overcomes some limitations of CTC. Gao et al. (2021) pointed out that CTC does not consider inter-dependencies between the output sequence and proposed to use RNN-Transducer as an alternative. Evaluating this approach on the CSL dataset yielded a WER of 6.1%. Sensor-based approaches were presented in Suri and Gupta (2019) and Zhang et al. (2020). Suri and Gupta (2019) suggested to apply CapsNet as an alternative to CNN for CSLR, which outperformed CNN with around 7% accuracy. However, the model was tested on only 20 sentences in ISL. WiFi data were leveraged for CSLR in Zhang et al. (2020). A DBN was introduced to extract features from the WiFi data describing the motion of the signer. DBN is a graphical generative model that has shown good recognition performance, especially with non-vision data (Zhang et al., 2020). The features extracted by the DBN were passed to HMM for temporal learning and classification. The method was evaluated on 100 sentences in ASL and obtained an accuracy of 83%.

5. Discussion and analysis

From the comprehensive literature review, we analyze the reviewed literature in terms of sign language, data acquisition, input modality, model architecture, and performance. Fig. 8 depicts an analysis of CSLR frameworks in terms of input modality, data acquisition, and features.

Sign Language. Although there are hundreds of sign languages used across the globe, only a few sign languages were targeted in the literature by researchers for CSLR, as can be seen in Fig. 9. This can be attributed to the shortage of publicly available datasets for most sign languages. In addition, most of the sign languages lack a documented description of their vocabulary, grammar, and structure. Moreover, the number of people who use a specific sign language plays an important role in attracting researchers to target that language, such as GSL, CSL, and ASL. As shown in Fig. 9, only 16 sign languages were addressed by all CSLR studies. GSL is the most addressed sign language, targeted by 40% of all CSLR studies. This is because of the increased popularity of the Phoenix2014 dataset, which is one of the earliest publicly available CSLR datasets. The CSL is the second most popular sign language, addressed by 25% of the CSLR studies. Thirdly is ASL, used in 12% of the studies. Other sign languages such as Arabic, Indian, and British were addressed only a handful of times.

Data Acquisition. Earlier studies before 2005 focused on wearable sensor-based data acquisition. Afterward, researchers shifted their attention to non-intrusive cameras, as illustrated in Fig. 8. With the increased focus on camera capturing, researchers have aimed to create more realistic recordings. Furthermore, to add variation in the data and increase the CSLR system's generalization, some newly released datasets, such as the GrSL dataset (Adaloglou et al., 2021), captured the videos with slight variations in camera orientation and position. Another trend is using more than one camera to provide a multi-view perspective, such as the How2Sign (Duarte et al., 2021) for SLT, which was recorded using two cameras for front and side view. Also, the FluentSigners-50 dataset (Mukushev et al., 2022) was the only CSLR dataset recorded using smartphones and tablets. Following their approach, we

Table 8
Summary of the surveyed DL-based approaches for CSLR. Results are in WER unless indicated otherwise.

Method	Year	Sign Lang.	Data Acq.		Architecture	Dataset	Result (%)	
			Vision	Sensor			Dev	Test
MLP-HMM (Gweth et al., 2012)	2012		✓		MLP-HMM		–	84.0 Acc.
ReSign (Koller et al., 2017)	2017		✓		(CNN-BLSTM-HMM)	SIGNUM	–	4.8
IterativeTrain Cui et al. (2019)	2019		✓		CNN-BLSTM		–	**2.8**
ReSign (Koller et al., 2017)	2017		✓		(CNN-BLSTM-HMM)	Phoenix2014 SI	54.1	44.1
CMA (Pu et al., 2020)	2020		✓		(CNN-BLSTM)		34.8	34.3
RadialCTC (Min et al., 2022)	2022		✓		(2DCNN-1DCNN-BLSTM)		33.8	**32.2**
SRM (Zuo & Mak, 2024)	2024		✓		(CNN-Transformer)		33.1	32.7
HamNoSys (Koller, Bowden, & Ney, 2016)	2016		✓		CNN-HMM		49.6	45.1
DeepSign (Koller, Zargaran, et al., 2016)	2016		✓		CNN-HMM		38.3	38.8
SubUNets (Camgoz et al., 2017)	2017		✓		CNN-BLSTM		40.8	40.7
StagedOpt (Cui et al., 2017)	2017		✓		CNN-BLSTM		39.4	38.7
ReSign (Koller et al., 2017)	2017		✓		CNN-BLSTM-HMM		27.1	26.8
IterativeOpt (Pu et al., 2018)	2018		✓		3DCNN-TempConv		38.0	37.3
LS-HAN (Huang et al., 2018)	2018		✓		3DCNN-Atten-DTW		–	61.7 Acc.
PL (Pei et al., 2019)	2019		✓		3DCNN-BGRU		40.9	40.6
RL (Zhang et al., 2019)	2019		✓		3DCNN-Transformer		38.0	38.3
Align-iOpt (Pu et al., 2019)	2019		✓		3DCNN-BLSTM-Atten		37.1	36.7
SFNet (Yang et al., 2019)	2019		✓		(2D3DCNN-LSTM-BLSTM)		35.6	34.9
DPD (Zhou et al., 2019)	2019		✓		3DCNN-GRU		35.6	34.5
IterativeTrain (Cui et al., 2019)	2019		✓		CNN-BLSTM		23.1	22.8
IterativeST (Koishybay et al., 2020)	2020		✓		2DCNN-1DCNN-BLSTM		34.5	34.4
SAFI (Zhou, Ng, et al., 2020)	2020		✓		(2+1)D-CNN-Atten		31.7	31.3
SAN (Slimane & Bouguessa, 2020)	2020		✓		CNN-Transformer		29.0	29.7
ST-LSTM (Xiao et al., 2020)	2020		✓		CNN-BLSTM-Atten		–	76.1 Acc.
Multi-Stream (Koller et al., 2020)	2020		✓		CNN-BLSTM-HMM		26.0	26.0
Stochastic (Niu & Mak, 2020)	2020		✓		3DCNN-Transformer		24.9	25.3
CMAlign (Papastratis et al., 2020)	2020		✓		CNN-1DCNN-BLSTM		23.9	24.0
FCN (Cheng et al., 2020)	2020		✓		FCN		24.6	23.9
CMA (Pu et al., 2020)	2020		✓		CNN-BLSTM		21.3	21.9
EnStimCTC (Adaloglou et al., 2021)	2021		✓		CNN-1DCNN		28.8	29.1
SBD-RL (Wei et al., 2021)	2021		✓		CNN-RNN with RL		23.4	23.5
SLRGAN (Papastratis, Dimitropoulos, & Daras, 2021)	2021		✓		GAN		23.7	23.4
PiSLTRc (Xie et al., 2021)	2021	GSL	✓		CNN-Transformer		23.4	23.2
H-GAN (Elakkiya et al., 2021)	2021		✓		GAN		–	88.0 Acc.
VAC (Min et al., 2021)	2021		✓		CNN-1DCNN-BLSTM	Phoenix2014	21.2	22.3
SMKD (Hao et al., 2021)	2021		✓		CNN-1DCNN-BLSTM		20.8	21.0
STMC (Zhou, Zhou, Zhou, & Li, 2021)	2021		✓		CNN-TempConv-BLSTM		21.1	20.7
SignBERT (Zhou, Tam, & Lam, 2021)	2021		✓		(3+2+1)DCNN-BERT-LSTM		21.2	20.2
MSTN (Li & Meng, 2022)	2022		✓		(ViT+3DCNN)-Transformer		–	22.8
LCSA (Zuo & Mak, 2022b)	2022		✓		CNN-Transformer		21.4	21.9
STAMF (Aditya et al., 2022)	2022		✓		CNN-Atten-BLSTM		20.5	21.5
MSTNet (Zhu et al., 2022)	2022		✓		CNN-1DCNN-Transformer		20.3	21.4
TLP (Hu et al., 2022)	2022		✓		CNN-1DCNN-BLSTM		19.7	20.8
C2SLR (Zuo & Mak, 2022a)	2022		✓		CNN-Transformer		20.5	20.4
RadialCTC (Min et al., 2022)	2022		✓		CNN-1DCNN-BLSTM		19.4	20.2
CA-SignBERT (Zhou et al., 2022)	2022		✓		(3+2+1)DCNN-BERT-LSTM		18.3	18.6
TSRNet (Zhu et al., 2023)	2023		✓		(CNN-1DCNN-Transformer)		23.4	24.7
mLTSF-Net (Xie et al., 2023)	2023		✓		FCN		22.9	23.0
SEN (Hu et al., 2023b)	2023		✓		CNN-1DCNN-BLSTM		19.5	21.0
AdaBrowse (Hu, Gao, et al., 2023)	2023		✓		CNN-1DCNN-BLSTM		19.6	20.7
DFConv (Jang et al., 2023)	2023		✓		CNN-1DCNN-BLSTM		20.9	20.8
CTCA (Guo et al., 2023)	2023		✓		CNN-1DCNN-BLSTM		19.5	20.3
Multilingual (Hu, Pu, et al., 2023)	2023		✓		CNN-1DCNN-BLSTM		20.3	20.9
Co-Sign (Jiao et al., 2023)	2023		✓		GCN-1DCNN-BLSTM		19.7	20.1
CVT-SLR (Zheng et al., 2023)	2023		✓		CNN-Atten-BLSTM		19.8	20.1
ST-Transformer (Cui et al., 2023)	2023		✓		ViT-Transformer		19.9	19.9
CorrNet (Hu et al., 2023a)	2023		✓		CNN-1DCNN-BLST		18.8	19.4
Two-Stream-SLR (Chen et al., 2022)	2023		✓		3DCNN		18.4	18.8
C2ST (Zhang, Guo, et al., 2023)	2023		✓		ViT-1DCNN-BLSTM		17.5	17.7
Cross-Ling. Wei and Chen (2023)	2023		✓		3DCNN		15.7	**16.7**
PA-CMA (Hu, Pu, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		20.2	20.0
AdaSize (Hu, Gao, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		19.7	20.9
SLTR (Camgöz et al., 2020)	2020		✓		CNN-Transformer	Phoenix2014T	24.6	24.4
CMAlign (Papastratis et al., 2020)	2020		✓		CNN-1DCNN-BLSTM		24.1	24.3
Multi-Stream (Koller et al., 2020)	2020		✓		CNN-BLSTM-HMM		22.1	24.1
PiSLTRc (Xie et al., 2021)	2021		✓		CNN-Transformer		21.8	22.9
SMKD (Hao et al., 2021)	2021		✓		CNN-1DCNN-BLSTM		20.8	22.4
TLP (Hu et al., 2022)	2022		✓		CNN-1DCNN-BLSTM		19.4	21.2
C2SLR (Zuo & Mak, 2022a)	2022		✓		CNN-Transformer		20.2	20.4
SEN (Hu et al., 2023b)	2023		✓		CNN-1DCNN-BLSTM		19.3	20.7
CorrNet (Hu et al., 2023a)	2023		✓		CNN-1DCNN-BLSTM		18.9	20.5
Co-Sign (Jiao et al., 2023)	2023		✓		GCN-1DCNN-BLSTM		19.5	20.1

(*continued on next page*)

Table 8 (*continued*).

Method	Year	Sign Lang.	Data Acq.		Architecture	Dataset	Result (%)	
			Vision	Sensor			Dev	Test
Two-Stream-SLR (Chen et al., 2022)	2023		✓		3DCNN		17.7	19.3
C2ST (Zhang, Guo, et al., 2023)	2023		✓		ViT-1DCNN-BLSTM		17.3	18.9
Cross-Ling (Wei & Chen, 2023)	2023		✓		3DCNN		16.9	**18.5**
AdaSize (Hu, Gao, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		19.7	21.2
PA-CMA (Hu, Pu, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		18.8	20.0
VAC (Jang et al., 2022)	2022		✓		CNN-1DCNN-BLSTM	Scene-PHOENIX	23.8	24.1
BR-DAE (Jang et al., 2022)	2022		✓		CNN-1DCNN-BLSTM		22.5	**23.1**
OpenPose-LSTM (Mocialov et al., 2017)	2017	DSL	✓		OpenPose-LSTM	Private	–	80.7
LS-HAN (Huang et al., 2018)	2018		✓		3DCNN-Attention		–	82.7 Acc.
DPD (Zhou et al., 2019)	2019		✓		3DCNN-GRU		–	4.7
SF-Net (Yang et al., 2019)	2019		✓		3DCNN-BLSTM		–	3.8
FCN (Cheng et al., 2020)	2020		✓		FCN		–	3.0
CMAlign (Papastratis et al., 2020)	2020		✓		CNN-1DCNN-BLSTM		–	2.4
BAE (Huang & Ye, 2021)	2021		✓		3DCNN-BLSTM-Atten		–	7.4
Transducer (Gao et al., 2021)	2021		✓		RNN-Transducer		–	6.1
PiSLTRc (Xie et al., 2021)	2021		✓		CNN-Transformer		–	2.8
EnStimCTC (Adaloglou et al., 2021)	2021		✓		CNN-1DCNN		–	2.4
SLRGAN (Papastratis, Dimitropoulos, & Daras, 2021)	2021		✓		GAN		–	2.1
VAC (Min et al., 2021)	2021		✓		VAC CNN-1DCNN-BLSTM		–	1.6
SignBERT (Zhou, Tam, & Lam, 2021)	2021		✓		(3+2+1)DCNN-BERT-LSTM		–	1.5
ST-GCN (Wang & Zhang, 2021)	2021		✓		GCN	CSL Split I	–	1.3
MSTN (Li & Meng, 2022)	2022		✓		(ViT+3DCNN)-Transformer		–	1.9
LCSA (Zuo & Mak, 2022b)	2022	CSL	✓		CNN-Transformer		–	1.4
CA-SignBERT (Zhou et al., 2022)	2022		✓		(3+2+1)DCNN-BERT-LSTM		–	1.1
C2SLR (Zuo & Mak, 2022a)	2022		✓		CNN-Transformer		–	0.9
STAMF (Aditya et al., 2022)	2022		✓		CNN-Atten-BLSTM		–	0.7
MSTNet (Zhu et al., 2022)	2022		✓		CNN-1DCNN-Transformer		–	0.7
mLTSF-Net (Xie et al., 2023)	2023		✓		FCN		–	2.5
ST-Transformer (Cui et al., 2023)	2023		✓		ViT-Transformer		–	1.2
SEN (Hu et al., 2023b)	2023		✓		CNN-1DCNN-BLSTM		–	0.8
CorrNet (Hu et al., 2023a)	2023		✓		CNN-1DCNN-BLSTM		–	0.8
AdaSize (Hu, Gao, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		–	0.8
SRM (Zuo & Mak, 2024)	2024		✓		CNN-Transformer		–	**0.6**
WIC-NGC (Wei et al., 2019)	2019		✓		3DCNN-BLSTM with N-Grams		–	50.9
Align-iOpt (Pu et al., 2019)	2019		✓		3DCNN-BLSTM-Atten	CSL Split II	–	32.7
CMA (Pu et al., 2020)	2020		✓		CNN-BLSTM		–	24.0
SBD-RL (Wei et al., 2021)	2021		✓		CNN-RNN with RL		–	26.8
PA-CMA (Hu, Pu, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		–	**22.5**
CorrNet (Hu et al., 2023a)	2023		✓		CNN-1DCNN-BLSTM		30.6	30.1
C2ST (Zhang, Guo, et al., 2023)	2023		✓		ViT-1DCNN-BLSTM		25.9	25.8
Two-Stream-SLR (Chen et al., 2022)	2023		✓		3DCNN	CSL-Daily	25.4	**25.3**
Co-Sign (Jiao et al., 2023)	2023		✓		GCN-1DCNN-BLSTM		28.1	27.2
AdaSize (Hu, Gao, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		31.3	30.9
PA-CMA (Hu, Pu, et al., 2024)	2024		✓		CNN-1DCNN-BLSTM		29.4	28.7
EM-Sign (Ye et al., 2020)	2020			✓	YOLOv3-tiny	Private	–	23.5
EnStimCTC (Adaloglou et al., 2021)	2021		✓		I3D-BLSTM		6.6	6.1
SLRGAN (Papastratis, Dimitropoulos, & Daras, 2021)	2021		✓		GAN	GrSL-SI	2.8	2.2
CA-SignBERT (Zhou et al., 2022)	2022	GrSL	✓		(3+2+1)DCNN-BERT-LSTM		2.2	**2.2**
EnStimCTC (Adaloglou et al., 2021)	2021		✓		CNN-1DCNN		38.9	42.3
SLRGAN (Papastratis, Dimitropoulos, & Daras, 2021)	2021		✓		GAN	GrSL-SD	30.5	37.1
CA-SignBERT (Zhou et al., 2022)	2022		✓		(3+2+1)DCNN-BERT-LSTM			**31.1**
Multilingual (Hu, Pu, et al., 2023)	2023		✓		CNN-1DCNN-BLSTM		32.7	33.5
DeepASL (Fang et al., 2017)	2017			✓	BLSTM	Private	–	16.1
3DRCNN (Ye et al., 2018)	2018	ASL	✓		3DCNN-RNN	Private	–	69.2 Acc.
ArmBand (Tateno et al., 2020)	2020			✓	LSTM	Private	–	97.7 Acc.
DBN (Zhang et al., 2020)	2020			✓	DBN-HMM	Private	–	83.0 Acc.
CapsNet (Suri & Gupta, 2019)	2019			✓	CapsNet	Private	–	94.0 Acc.
LeapMotion (Mittal et al., 2019)	2019	ISL		✓	CNN-LSTM	Private	–	89.5 Acc.
IMUArm (Sharma et al., 2021)	2021			✓	CNN-BLSTM	Private	–	87.6 Acc.
Pose-SLT (Ko et al., 2019)	2019	KSL	✓		OpenPose-Transformer	Private	–	55.2 Acc.
Non-manual (Brock et al., 2020)	2020	JSL	✓		OpenPose-RF-CNN	Private	–	11.3

anticipate a shift from using sophisticated cameras and sensors to more accessible data-capturing methods, such as smartphones, tablets and webcams.

Input Modality. Multi-modal CSLR has gained increased attention recently. Several SOTA CSLR frameworks (Chen et al., 2022; Zhou, Tam, & Lam, 2021; Zhou et al., 2022) made use of skeletal data to enhance the CSLR performance by providing more detailed information about the signers joints and body pose. The availability of off-the-shelf pose estimation methods such as OpenPose, has encouraged researchers to leverage skeletal data as the primary modality (Wang & Zhang, 2021) or in combination with RGB (Aditya

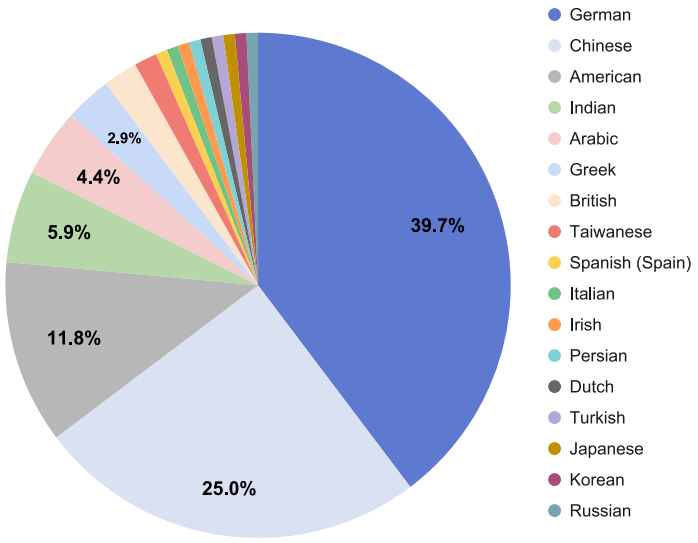

Fig. 9. Distribution of CSLR studies per sign language.

et al., 2022; Chen et al., 2022; Wei & Chen, 2023), as seen in Fig. 8. Combining RGB and pose data has resulted in boosting the CSLR performance, as observed in the SOTA TwoStream-SLR models (Aditya et al., 2022; Chen et al., 2022; Wei & Chen, 2023). Also, training the models on both video and depth data results in creating more robust models, less sensitive to environment changes, as seen in Dreuw, Steingrube, et al. (2009), Ye et al. (2018) and Zhang et al. (2016).

Sign Representation. Past studies mainly focused on manual features either obtained from sensors (e.g. data gloves) or using hand cropped images. With the emergence of DL, the full frame is predominantly used for feature extraction, as shown in Fig. 8. Although the full frame provides a global view of the sign gestures, some gesture movements are small and may not be very clear in the full frame, such as finger movements. Therefore, it is beneficial to emphasize these subtle gestures, as seen in the SOTA CA-SignBERT (Zhou et al., 2022) model. The authors leveraged the full frame and cropped hand images in their two-stream model. Moreover, studies have recently aimed to emphasize non-manual features in different ways, such as using cropped face images (Koller et al., 2020) or using the pose keypoints of the face and body pose (Aditya et al., 2022; Chen et al., 2022; Wei & Chen, 2023). Some studies argued that using additional modalities or channels involves extra computation and proposed alternative methods to dynamically identify and emphasize the non-manual cues (Hu et al., 2023a, 2023b).

Recognition Approaches. In the last decade, researchers abandoned the traditional approaches that depend on hand-crafted features for CSLR and utilized DL-based techniques for feature learning. The majority of the DL-based CSLR models adopted CNNs for spatial feature extraction, as seen in Table 9 and Fig. 10. A variety of pretrained CNN backbones were leveraged, such as CaffeNet (Camgoz et al., 2017), ResNet (Aditya et al., 2022; Gao et al., 2021; Hao et al., 2021; Hu et al., 2022, 2023a, 2023b; Min et al., 2021; Yang et al., 2019; Zhu et al., 2022, 2023), GoogLeNet (Cui et al., 2017, 2019; Koller et al., 2017; Pu et al., 2020), VGG (Cui et al., 2017; Zhou, Zhou, Zhou, & Li, 2021; Zuo & Mak, 2022a, 2024), and BN-Inception (Papastratis, Dimitropoulos, & Daras, 2021; Papastratis et al., 2020).

Alternatively, 3DCNNs were leveraged in some investigations for both spatial and temporal encoding. Amongst the DL models benchmarked on the Phoenix2014 dataset, 27% employed 3DCNNs. Various pretrained 3DCNNs were utilized, including 3DResNets (Pei et al., 2019; Pu et al., 2018, 2019; Zhang et al., 2019), I3D model (Adaloglou et al., 2021; Albanie et al., 2020), C3D (Huang et al., 2018; Ye et al., 2018) and S3D (Chen et al., 2022; Wei & Chen, 2023). Moreover, most of the utilized 3DCNNs were pretrained on action recognition data, such as the Kinetics-400 dataset (Adaloglou et al., 2021; Albanie et al., 2020; Chen et al., 2022; Koishybay et al., 2020), Sports1M (Ye et al., 2018), and UCF101 dataset (Pei et al., 2019) This can be attributed to the similarity between SLR and action recognition tasks. Conversely, few studies (Pu et al., 2018, 2019; Zhang et al., 2019) pretrained their 3DCNNs on isolated SLR datasets. The best performing model (Wei & Chen, 2023) on Phoenix2014, Phoenix2014T and CSL-Daily utilized an S3D backbone pretrained on Kinetics-400. Although 3DCNNs were successfully incorporated for CSLR, they are complex and often require excessive memory usage.

As for temporal modeling, BLSTM was the most frequent network employed previous research to learn time-wise features. This can be attributed to the capabilities of BLSTM to utilize past and future information while producing sign language sentences. Several studies empirically found that BLSTM performs better compared to the uni-directional LSTM for CSLR (Cui et al., 2017; Koller et al., 2017). However, some studies argued that RNN and its variants are unsuitable for online CSLR in the real world and adopted FCNs (Cheng et al., 2020; Xie et al., 2023), where the temporal features are learned using 1D TempConvs. Although several studies employed TempConvs for temporal encoding, most networks used fixed receptive fields to model signs of varying lengths, which limits the model's capability to grasp the full temporal aspects of different signs. Few studies addressed this issue

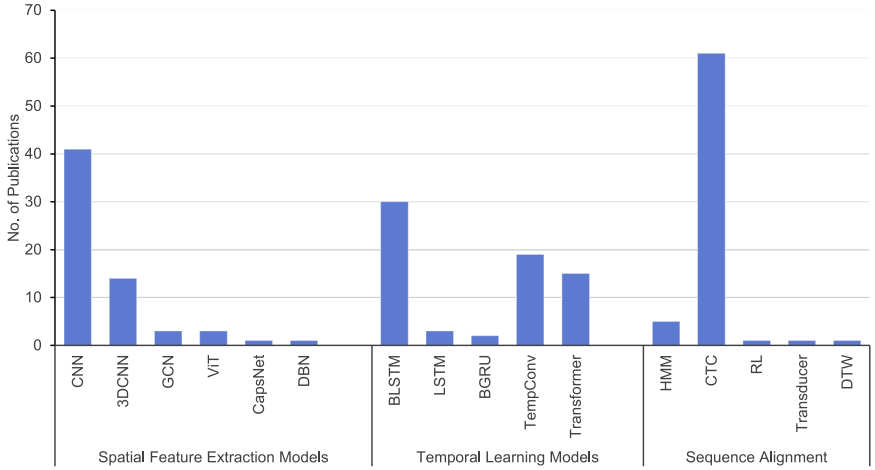

Fig. 10. Analysis of DL-based CSLR frameworks in terms of techniques used for spatial, temporal, and alignment modeling.

using TempConvs with multi-scale receptive fields (Wei et al., 2021; Xie et al., 2023; Zhu et al., 2022), achieving better recognition accuracy. Another alternative to RNNs employed in the literature was using self-attention to capture the sequential dependencies. Multiple studies (Camgöz et al., 2020; Cui et al., 2023; Ko et al., 2019; Li & Meng, 2022; Niu & Mak, 2020; Papastratis, Dimitropoulos, & Daras, 2021; Slimane & Bouguessa, 2020; Xie et al., 2021; Zhang et al., 2019; Zhou, Tam, & Lam, 2021; Zhou et al., 2022; Zhu et al., 2022) proposed Transformer-based frameworks for CSLR. However, most studies utilized the Transformer for temporal modeling only (Camgöz et al., 2020; Ko et al., 2019; Niu & Mak, 2020; Papastratis, Dimitropoulos, & Daras, 2021; Slimane & Bouguessa, 2020; Xie et al., 2021; Zhang et al., 2019; Zhou, Tam, & Lam, 2021; Zhou et al., 2022; Zhu et al., 2022), and few studies investigated ViTs for spatial feature extraction (Cui et al., 2023; Li & Meng, 2022; Zhang, Guo, et al., 2023). The C2ST (Zhang, Guo, et al., 2023) model, utilized Swin Transformer for spatial feature extraction and achieved SOTA results. Notably, only one study explored the prospects of a fully Transformer-based CSLR framework (SP-Transformer) (Cui et al., 2023). Regarding sequence alignment methods, almost all DL-based CSLR methods employed CTC to align the frames with the gloss sequence. Few DL-based CSLR studies explored alternatives to CTC, such as HMM (Koller, Bowden, & Ney, 2016; Koller et al., 2020, 2017; Koller, Zargaran, et al., 2016; Zhang et al., 2020), RL (Wei et al., 2021), Transducer models (Gao et al., 2021), and DTW (Huang et al., 2018).

Performance. In the following, we discuss the performance of CSLR systems with various public CSLR datasets, namely, Phoenix2014, Phoenix2014T, CSL, CSL-Daily and continuous GrSL datasets. The performance of CSLR models has significantly improved since the late of 90s'. Fig. 11 depicts the performance of CSLR models reported on the Phoenix2014 multi-signer dataset, which is the most commonly used CSLR benchmarking dataset. The figure emphasizes the best performing model in each year, since the date the dataset was released (2014) till the year of writing this survey (2023). The HMM baseline obtained 53% WER using hand-crafted features in 2014 (Koller, Forster, & Ney, 2015). In the following year, a significant WER reduction of 8% was achieved by adopting CNN learned features instead of the traditional features fed into an HMM model in the DeepSign network (Koller, Bowden, & Ney, 2016). Using RNNs and iterative re-alignment in the CNN-BLSTM-HMM model (ReSign) (Koller et al., 2017), resulted in a steep reduction of WER by 12%. Researchers subsequently abandoned HMM-based models and adopted a common BLSTM-CTC architecture, which further enhanced the performance by around 7% using also iterative training (Cui et al., 2017). Gradual improvements in the CSLR performance were achieved throughout the following years. As of the time this survey was conducted, the cross-ling (Wei & Chen, 2023) is the best performing model with a WER of 16.8% on the Phoenix2014 dataset using additional labeled data. The C2ST (Zhang, Guo, et al., 2023), however, is the best performing model trained using only the Phoenix2014 dataset with language modeling obtaining a WER of 17.7%. Similarly, the cross-ling (Wei & Chen, 2023) model achieved the best performance on Phoenix2014T with 18.6% WER, which highlights the significance of labeled data in improving the performance CSLR models, particularly with datasets having limited examples. In addition, the SOTA models utilized RGB and pose data which signifies the importance of multi-modality for CSLR systems, specially complementing videos with skeleton data that can thoroughly describe the subtle details of the signers hands, face and body.

As for the CSL dataset, which is the second most benchmarked dataset, the SRM (Zuo & Mak, 2024) model is the current best performing models with 0.6% WER on the CSL split I dataset for sign independent CSLR. This highlights the effectiveness of the proposed feature disengagement method for signer independent CSLR. On the other hand, PA-CMA (Hu, Pu, et al., 2024) achieved the best results with CSL split II for unseen sentences CSLR with 22.5% WER. However, their approach entailed creating video and label augmentations to enhance the recognition of new sentences. Moreover, the CSL dataset is relatively small and researchers have shifted to the newly released CSL-Daily dataset, which is a larger and more challenging dataset. The dataset has few benchmarks, with the two-stream 3DCNN model (Chen et al., 2022) having the best result, obtaining 25.3% WER. Regarding the GrSL dataset, which is also newly released, the CA-SignBERT (Zhou et al., 2022) obtained the best performance for both the signer-independent

Table 9
Classification of deep learning CSLR frameworks based on techniques used for spatial, temporal and alignment modeling.

	Method	References
Spatial	CNN	Aditya et al. (2022), Brock et al. (2020), Camgoz et al. (2017), Camgöz et al. (2020), Cui et al. (2017, 2019), Gao et al. (2021), Guo et al. (2023), Hao et al. (2021), Hu et al. (2022, 2023a, 2023b), Hu, Gao, et al. (2024, 2023), Hu, Pu, et al. (2024, 2023), Jang et al. (2022, 2023), Koishybay et al. (2020), Koller, Bowden, and Ney (2016), Koller et al. (2020, 2017), Koller, Zargaran, et al. (2016), Min et al. (2021, 2022), Mittal et al. (2019), Niu and Mak (2020), Papastratis, Dimitropoulos, and Daras (2021), Papastratis et al. (2020), Pu et al. (2020), Sharma et al. (2021, 2021), Wei et al. (2021), Xie et al. (2023), Ye et al. (2020), Zheng et al. (2023), Zhou, Tam, and Lam (2021), Zhou, Zhou, Zhou, and Li (2021), Zhu et al. (2022), Zuo and Mak (2022a, 2022b, 2024)
	3DCNN	Adaloglou et al. (2021), Chen et al. (2022), Huang and Ye (2021), Huang et al. (2018), Pei et al. (2019), Pu et al. (2018, 2019), Wei and Chen (2023), Wei et al. (2019), Yang et al. (2019), Ye et al. (2018), Zhang et al. (2019), Zhou et al. (2022, 2019)
	GCN	Jiao et al. (2023), Li and Meng (2022), Wang and Zhang (2021)
	ViT	Cui et al. (2023), Li and Meng (2022), Zhang, Guo, et al. (2023)
	CapsNet	Suri and Gupta (2019)
	DBN	Zhang et al. (2020)
Temporal	BLSTM	Aditya et al. (2022), Camgoz et al. (2017), Cui et al. (2017, 2019), Fang et al. (2017, 2017), Gao et al. (2021), Guo et al. (2023), Hao et al. (2021), Hu et al. (2022, 2023a), Hu, Gao, et al. (2024, 2023), Hu, Pu, et al. (2024, 2023), Jang et al. (2022, 2023), Jiao et al. (2023), Koishybay et al. (2020), Koller et al. (2020, 2017), Min et al. (2021, 2022), Papastratis, Dimitropoulos, and Daras (2021), Papastratis et al. (2020), Pu et al. (2020), Sharma et al. (2021, 2021), Wei et al. (2021), Zheng et al. (2023), Zhou, Zhou, Zhou, and Li (2021)
	LSTM	Mittal et al. (2019), Mocialov et al. (2017), Pu et al. (2019), Tateno et al. (2020)
	BGRU	Pei et al. (2019), Zhou et al. (2019)
	TempConv	Guo et al. (2023), Hao et al. (2021), Hu et al. (2022, 2023a, 2023b), Hu, Gao, et al. (2024, 2023), Hu, Pu, et al. (2024, 2023), Jang et al. (2022, 2023), Jiao et al. (2023), Min et al. (2021, 2022), Papastratis, Dimitropoulos, and Daras (2021), Papastratis et al. (2020), Pu et al. (2018), Xie et al. (2023), Zhou, Zhou, Zhou, and Li (2021), Zhu et al. (2022)
	Transformer	Camgöz et al. (2020), Cui et al. (2023), Ko et al. (2019), Li and Meng (2022), Niu and Mak (2020), Papastratis, Dimitropoulos, and Daras (2021), Slimane and Bouguessa (2020), Xie et al. (2021), Zhang et al. (2019), Zheng et al. (2023), Zhou, Tam, and Lam (2021), Zhou et al. (2022), Zhu et al. (2022), Zuo and Mak (2022a, 2022b, 2024)
Alignment	HMM	Koller, Bowden, and Ney (2016), Koller et al. (2020, 2017), Koller, Zargaran, et al. (2016), Zhang et al. (2020)
	CTC	Adaloglou et al. (2021), Aditya et al. (2022), Brock et al. (2020), Camgoz et al. (2017), Camgöz et al. (2020), Chen et al. (2022), Cheng et al. (2020), Cui et al. (2017, 2019, 2023), Fang et al. (2017), Gao et al. (2021), Guo et al. (2023), Hao et al. (2021), Hu et al. (2022, 2023a, 2023b), Hu, Gao, et al. (2024, 2023), Hu, Pu, et al. (2024, 2023), Huang and Ye (2021), Huang et al. (2018), Jang et al. (2022, 2023), Jiao et al. (2023), Ko et al. (2019), Koishybay et al. (2020), Koller, Bowden, and Ney (2016), Koller et al. (2020, 2017), Koller, Zargaran, et al. (2016), Li and Meng (2022), Min et al. (2021, 2022), Mocialov et al. (2017), Niu and Mak (2020), Papastratis, Dimitropoulos, and Daras (2021), Papastratis et al. (2020), Pei et al. (2019), Pu et al. (2020, 2018, 2019), Sharma et al. (2021), Slimane and Bouguessa (2020), Wang and Zhang (2021), Wei et al. (2021, 2019), Xie et al. (2023, 2021), Yang et al. (2019), Ye et al. (2018), Zhang et al. (2019), Zheng et al. (2023), Zhou, Tam, and Lam (2021), Zhou et al. (2022), Zhou, Zhou, Zhou, and Li (2021), Zhu et al. (2022), Zuo and Mak (2022a, 2022b, 2024)
	RL	Wei et al. (2021)
	Transducer	Gao et al. (2021)
	DTW	Huang et al. (2018)

and unseen sentences splits, yielding WERs of 2.2% and 31.15%, respectively. In summary, the performance of CSLR models can vary widely depending on several factors, including the dataset employed, modalities used, and the complexity of the sign language being modeled, and while CSLR models have shown significant improvements in accuracy over the years, they generally lag behind, especially in terms of robustness and real-time performance. We also note that the majority of CSLR research focused on reducing error rates and limited attention has been given to other metrics such as model complexity and inference time, which need to be considered for online CSLR. Only few studies (Hu, Gao, et al., 2023; Zhu et al., 2023) explored creating lightweight models.

6. Research gaps and future directions

With the advancements in DL-based approaches, CSLR systems have shown significant accuracy improvements over the years. However, there is still much room for improvement, as revealed in this literature review. In the following, we highlight the observed

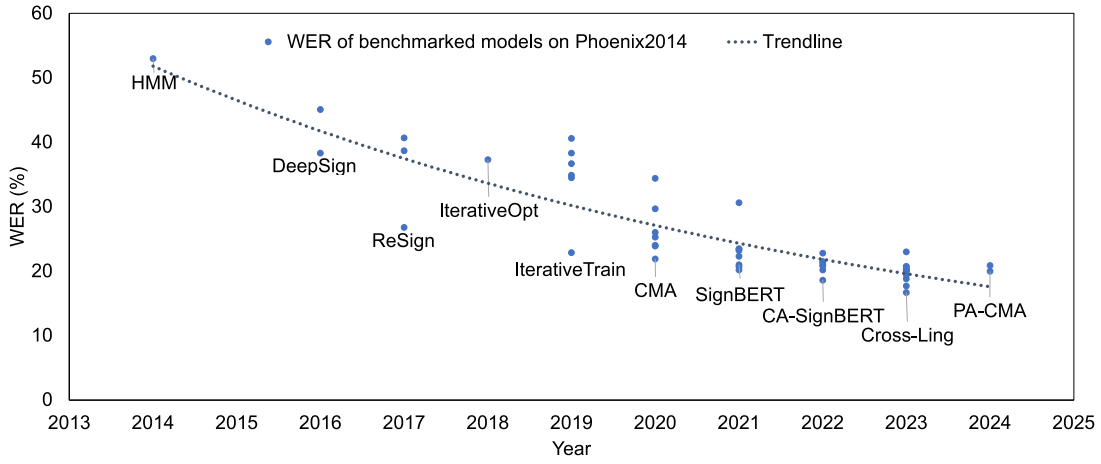

Fig. 11. Performance of best-performing CSLR models on Phoenix2014 dataset per year.

limitations and open research gaps in the area of CSLR. We categorize these gaps into data-centric and model-centric gaps. These gaps need to be investigated further by researchers.

6.1. Data-centric gaps

- Lack of CSLR datasets: The availability of comprehensive and diverse CSLR datasets is essential for advancing CSLR. Evidently, publicly available CSLR datasets are limited, and the currently available datasets cover few sign languages. Hence, hundreds of sign languages are neglected by CSLR researchers. Therefore, more efforts should be directed towards building CSLR datasets for the resource-scarce sign languages, which will enable understanding features of these sign languages and creating CSLR models tailored to their specifications.
- Limited signing domain: Most CSLR datasets cover only a limited set of signs. The Phoenix2014, which is the standard benchmarking dataset contains only 1389 unique signs used within a limited domain, namely weather forecasts. Hence, the current SOTA are expected to operate within a limited domain of discourse and are far from commercial use. Therefore, to enable the use of the CSLR model in the real world, there is a need to create linguistically rich CSLR datasets that cover a large set of signs spanning various domains and scenarios.
- Unconstrained and naturalistic signing: The majority of the publicly available datasets are recorded in controlled environments, whereas there is a need for realistic training data spanning various environment settings, including different camera angles, backgrounds, video quality, and aspect ratio. Also, we note that the Phoenix2014 dataset is the only dataset collected from real-life data where the signs are performed on the fly in a speedy manner. Moreover, it is observed that all the publicly available CSLR datasets have a fixed background, while CSLR systems should be able to operate with complex and moving backgrounds. Therefore, creating realistic data will enable building robust CSLR models that can be employed by various consumers with unrestricted settings.
- Selfie-view datasets: Although it is expected that the CSLR models would eventually be deployed to smart devices, such as mobile phones and tablets, the majority of CSLR datasets were not recorded in selfie view using smartphones. Hence, there is a need to create CSLR datasets using a variety of smartphones and tablets, where the videos are recorded in close front selfie view.

6.2. Model-centric gaps

- Training with few samples: Given that creating CSLR datasets is a challenging task that requires employing fluent sign language signers, it is often difficult and costly to create a large number of examples in the datasets. Thus, efforts should be directed at building CSLR frameworks that can operate with limited training data. For this purpose, newly emerged concepts such as few-shot learning can be adopted for CSLR. Also, semi-surprised learning techniques can be leveraged to exploit the large unlabeled body of sign language content available on public websites, such as YouTube.
- Signer-independence: The performance of CSLR models steeply declines when faced with unseen signers. This is mainly attributed to the limited number of unique signers in CSLR datasets. This issue can be addressed by increasing the number of signers in the training set, which may not be always possible. Another approach is investigating effective data augmentation methods reflecting various appearances and signing speeds. In addition, pose-based techniques do not depend on the signer, and they can be used for developing robust CSLR systems.

- Non-manual features: To correctly recognize sign language, both manual and non-manual gestures need to be integrated. However, different cue types have unequal importance. Modeling the dependencies between multi-cue features is still a non-trivial issue. Moreover, DL-based methods primarily focus on strong features to quickly converge, which can result in overlooking other relevant cues. Incorporating non-manual features has not been studied deeply in the literature. We also note that no recent work has studied eye gaze, which is an informative parameter in interpreting sign language.
- Multi-modal fusion: Integrating information from multiple modalities, such as RGB video, depth data, skeletal information, and textual data can improve recognition accuracy. According to the conducted survey, the lowest WERs using the Phoenix2014 dataset were obtained when integrating RGB with pose data, as observed in Chen et al. (2022) and Zuo and Mak (2024). Nonetheless, multi-modal CSLR systems are computationally complex, and further research is needed to develop effective methods for fusing and leveraging these diverse sources without burdening the system with additional computation.
- Real-time CSLR: Although there has been an increased interest in developing real-time isolated SLR (Abdul et al., 2021; Chan-Wah & Surendra, 2002; Pan et al., 2016; Rashid & Albelwi, 2012; Rastgoo et al., 2021; Starner, Weaver, & Pentland, 1998), CSLR researchers have neglected to investigate this aspect. Developing real-time and online CSLR systems is a challenge due to the need for low latency and robust performance. Therefore, developing lightweight fast models that can meet the requirements for real-time applications is needed. We noted that Hu, Gao, et al. (2024, 2023) and Zhu et al. (2023) are the only recent works devoted to creating lightweight CSLR models.
- Multi-Task Learning (MTL): It involves training a model on multiple related tasks simultaneously, enabling better generalization on each task by exploiting shared information. Jointly training systems for CSLR and SLT has seen great potential, as both systems can benefit from a shared representation of visual features, which can improve the accuracy of sign recognition and the quality of the translation (Camgöz et al., 2020; Papastratis, Dimitropoulos, & Daras, 2021; Xie et al., 2021). In addition, collaborative Multilingual CSLR achieved admirable results outperforming individually trained recognition models, by taking advantage of shared low-level visual patterns amongst sign languages (Hu, Pu, et al., 2023). MTL can be explored between other sign language understanding tasks such as sign localization and spotting. Also amongst other related fields including gesture and action recognition.
- Interactive CSLR: Interactive CSLR systems allow for real-time interaction between users and the system, enabling dynamic feedback and adaptation. For example, incorporating user feedback and corrections during CSLR prediction can improve the accuracy and robustness of the system over time.
- Under exploitation of language modeling: Most studies addressed CSLR as a video understanding task only, whilst CSLR is also a language modeling task. Few studies (Guo et al., 2023; Zhang, Guo, et al., 2023; Zheng et al., 2023; Zhou, Tam, & Lam, 2021) incorporated this aspect to produce accurate glosses.
- Multi-person CSLR: All previous CSLR works assume that there is only one person present in the frame. Hence, these models would fail in situations where more than one signer is present in the scene. Evidently, multi-person conversation-aware SLR is needed.

7. Conclusion

CSLR has gained increased attention recently due to the rapid advances in machine learning techniques. This paper offers a comprehensive review of CSLR research efforts spanning the past 25 years. Its purpose is to identify existing gaps in the research and to facilitate the advancement of the CSLR field. This study included a discussion on various challenges associated with the task of CSLR, including sign boundary detection and co-articulation effect. Moreover, publicly available CSLR datasets were listed and reviewed in terms of sign language, modality, numbers of signs, sentences and samples. Furthermore, a critical analysis of 126 studies was presented and organized into a taxonomy comprising seven dimensions: sign language, data acquisition, input modality, leveraged sign language cues, recognition techniques, utilized datasets, and overall performance. Additionally, DL-based CSLR frameworks were further categorized based on spatial, temporal, and alignment learning methods.

The conducted survey revealed many insightful observations, such as scarcity of annotated CSLR datasets, which cover few sign languages, mainly ASL, DSL, and CSL. This can justify the lack of CSLR research on other sign languages. In addition, most of publicly available datasets were recorded in controlled environments and cover a limited number of signs within specific domains. This prohibits the creation of robust CSLR models capable of performing in real-world scenarios. Regarding CSLR techniques, a shift from sensor-based to vision-based approaches was observed, along with a trend to adopt the full-frame image for a global representation of the features. We also observe the great potential of multi-modal CSLR, specifically, complementing RGB with pose data. Incorporating prior language information in CSLR models has also revealed promising results, which was overlooked by the majority of existing CSLR models. As for recognition approaches, most recent CSLR systems adopt pre-trained CNN backbones for spatial feature extraction, followed by TempConvs and BLSTMs for temporal encoding and sequence learning. The application of Transformers was mostly limited to temporal learning, and few studies utilized ViTs as their visual backbone. Moreover, most DL-based models adopt CTC loss to learn the frame-gloss alignments. Given the over-fitting tendency of CTC, a significant part of CSLR research steered towards developing solutions to enforce sufficient model training, including iterative training strategies and auxiliary losses. Additional research gaps and future research prospects were highlighted and discussed in this survey, including fusion of multi-modal CSLR systems, vision-language modeling, and multi-person scenes in CSLR models. The conducted survey can be extended to include other related topics, such as finger-spelling, isolated SLR, and SLT systems. Furthermore, the topic of sign language generation, which is a new and exciting area in sign language understanding, can be discussed in future surveys.

CRediT authorship contribution statement

Sarah Alyami: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Hamzah Luqman:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Mohammad Hammoudeh:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors would like to acknowledge the support received from the Saudi Data and AI Authority (SDAIA), Saudi Arabia and King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant no. JRC-AI-RFP-14.

References

Abbas, S., Al-Barhamtoshy, H., & Alotaibi, F. (2021). Towards an Arabic Sign Language (ArSL) corpus for deaf drivers . *PeerJ Computer Science*, *7*, Article e741. http://dx.doi.org/10.7717/peerj-cs.741.

Abdul, W., Alsulaiman, M., Amin, S. U., Faisal, M., Muhammad, G., Albogamy, F. R., et al. (2021). Intelligent real-time arabic sign language classification using attention-based inception and bilstm. *Computers & Electrical Engineering*, *95*(April), Article 107395. http://dx.doi.org/10.1016/j.compeleceng.2021.107395.

Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., & Th, G. (2021). A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, 1–14, arXiv:arXiv:2007.12530v2.

Adeyanju, I. A., Bello, O. O., & Adegboye, M. A. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, *12*, Article 200056. http://dx.doi.org/10.1016/j.iswa.2021.200056.

Aditya, W., Shih, T. K., Thaipisutikul, T., Fitriajie, A. S., Gochoo, M., Utaminingrum, F., et al. (2022). Novel spatio-temporal continuous sign language recognition using an attentive multi-feature network. *Sensors*, *22*(17), http://dx.doi.org/10.3390/s22176452.

Al-Qurishi, M., Khalid, T., & Souissi, R. (2021). Deep learning for sign language recognition: Current techniques, benchmarks, and open issues. *IEEE Access*, *9*, 126917–126951. http://dx.doi.org/10.1109/ACCESS.2021.3110912.

AL-Rousan, M., Assaleh, K., & Tala'a, A. (2009). Video-based signer-independent arabic sign language recognition using hidden Markov models. *Applied Soft Computing*, *9*(3), 990–999. http://dx.doi.org/10.1016/j.asoc.2009.01.002.

Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., et al. (2020). BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): vol. 12356 LNCS*, (pp. 35–53). http://dx.doi.org/10.1007/978-3-030-58621-8_3, arXiv:2007.12131.

Aloysius, N., & Geetha, M. (2020). Understanding vision-based continuous sign language recognition. *Multimedia Tools and Applications*, *79*(31–32), 22177–22209. http://dx.doi.org/10.1007/s11042-020-08961-z.

Ananthanarayana, T., Srivastava, P., Chintha, A., Santha, A., Landy, B., Panaro, J., et al. (2021). Deep learning methods for sign language translation. *ACM Transactions on Accessible Computing*, *14*(4), http://dx.doi.org/10.1145/3477498.

Assaleh, K., Shanableh, T., Fanaswala, M., Amin, F., & Bajaj, H. (2010). Continuous arabic sign language recognition in user dependent mode. *Journal of Intelligent Learning Systems and Applications*, *02*(01), 19–27. http://dx.doi.org/10.4236/jilsa.2010.21003.

Athira, P. K., Sruthi, C. J., & Lijiya, A. (2019). A signer independent sign language recognition with co-articulation elimination from live videos: An Indian scenario. *Journal of King Saud University - Computer and Information Sciences*, *34*(3), 771–781. http://dx.doi.org/10.1016/j.jksuci.2019.05.002.

Bauer, B., Hienz, H., & Kraiss, K. F. (2000). Video-based continuous sign language recognition using statistical methods. In *Proceedings-international conference on pattern recognition, vol. 15, no. 2* (pp. 463–466). http://dx.doi.org/10.1109/ICPR.2000.906112.

Bauer, B., & Kraiss, K. F. (2002). Towards an automatic sign language recognition system using subunits. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): vol. 2298*, (pp. 64–75). http://dx.doi.org/10.1007/3-540-47873-6_7.

Boháček, M., & Hrúz, M. (2022). Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV) workshops*.

Brock, H., Farag, I., & Nakadai, K. (2020). Recognition of non-manual content in continuous Japanese sign language. *Sensors (Switzerland)*, *20*(19), 1–21. http://dx.doi.org/10.3390/s20195621.

Buehler, P., Everingham, M., & Zisserman, A. (2009). Learning sign language by watching TV (using weakly aligned subtitles). In *2009 IEEE conference on computer vision and pattern recognition* (pp. 2961–2968). http://dx.doi.org/10.1109/CVPRW.2009.5206523.

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2017). SubUNets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision, vol. 2017-October* (pp. 3075–3084). http://dx.doi.org/10.1109/ICCV.2017.332.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 7784–7793). http://dx.doi.org/10.1109/CVPR.2018.00812.

Camgöz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 10020–10030). http://dx.doi.org/10.1109/CVPR42600.2020.01004, arXiv:2003.13830.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). Openpose: Realtime multi-person 2D pose estimation using part affinity fields. arXiv:1812.08008.

Chan-Wah, N., & Surendra, R. (2002). Real-time hand gesture recognition system and application. *Image and Vision Computing*, *20*(4), 993–1007. http://dx.doi.org/10.1016/S0262-8856(02)00113-0.

Chen, Y., Zuo, R., Wei, F., Wu, Y., Liu, S., & Mak, B. (2022). Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, *35*, 17043–17056.

Cheng, K. L., Yang, Z., Chen, Q., & Tai, Y.-W. (2020). Fully convolutional networks for continuous sign language recognition. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer vision – ECCV 2020* (pp. 697–714). Cham: Springer International Publishing.

Choudhury, A., Talukdar, A. K., Bhuyan, M. K., & Sarma, K. K. (2017). Movement epenthesis detection for continuous sign language recognition. *Journal of Intelligent Systems*, *26*(3), 471–481. http://dx.doi.org/10.1515/jisys-2016-0009.

Cortés, G., García, L., Benítez, C., & Segura, J. C. (2006). HMM-based continuous sign language recognition using a fast optical flow parameterization of visual information. In *INTERSPEECH 2006 and 9th international conference on spoken language processing, vol. 3, no. January* (pp. 1288–1291). http://dx.doi.org/10.21437/interspeech.2006-379.

Cui, R., Liu, H., & Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings - 30th IEEE conference on computer vision and pattern recognition, vol. 2017-January* (pp. 1610–1618). http://dx.doi.org/10.1109/CVPR.2017.175.

Cui, R., Liu, H., & Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, *21*(7), 1880–1891. http://dx.doi.org/10.1109/TMM.2018.2889563.

Cui, Z., Zhang, W., Li, Z., & Wang, Z. (2023). Spatial–temporal transformer for end-to-end sign language recognition. *Complex & Intelligent Systems*, 1–12.

Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., & Ney, H. (2007). Speech recognition techniques for a sign language recognition system. In *Proceedings of the annual conference of the international speech communication association, vol. 1* (pp. 705–708). http://dx.doi.org/10.21437/interspeech.2007-668.

Dreuw, P., Stein, D., & Ney, H. (2009). Enhancing a sign language translation system with vision-based features. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): vol. 5085 LNAI*, (pp. 108–113). http://dx.doi.org/10.1007/978-3-540-92865-2_11.

Dreuw, P., Steingrube, P., Deselaers, T., & Ney, H. (2009). Smoothed disparity maps for continuous American sign language recognition. In H. Araujo, A. M. Mendonça, A. J. Pinho, & M. I. Torres (Eds.), *Pattern recognition and image analysis* (pp. 24–31). Berlin, Heidelberg: Springer Berlin Heidelberg.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., et al. (2021). How2Sign: A large-scale multimodal dataset for continuous American sign language. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 2734–2743). http://dx.doi.org/10.1109/CVPR46437.2021.00276, arXiv:2008.08143.

Ekiz, D., Kaya, G. E., Bugur, S., Guler, S., Buz, B., Kosucu, B., et al. (2017). Sign sentence recognition with smart watches. In *2017 25th signal processing and communications applications conference*. Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/SIU.2017.7960255.

El-Alfy, E. S. M., & Luqman, H. (2022). A comprehensive survey and taxonomy of sign language research. *Engineering Applications of Artificial Intelligence*, *114*(July), Article 105198. http://dx.doi.org/10.1016/j.engappai.2022.105198.

Elakkiya, R. (2020). Machine learning based sign language recognition: a review and its research frontier. *Journal of Ambient Intelligence and Humanized Computing*, *12*(7), 7205–7224. http://dx.doi.org/10.1007/s12652-020-02396-y.

Elakkiya, R., & Selvamani, K. (2019). Subunit sign modeling framework for continuous sign language recognition. *Computers & Electrical Engineering*, *74*, 379–390. http://dx.doi.org/10.1016/j.compeleceng.2019.02.012.

Elakkiya, R., Vijayakumar, P., & Kumar, N. (2021). An optimized generative adversarial network based continuous sign language classification. *Expert Systems with Applications*, *182*(May), Article 115276. http://dx.doi.org/10.1016/j.eswa.2021.115276.

Fang, B., Co, J., & Zhang, M. (2017). 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* (pp. 1–13).

Fang, G., Gao, W., Chen, X., Wang, C., & Ma, J. (2002). Signer-independent continuous sign language recognition based on SRN/HMM. In I. Wachsmuth, & T. Sowa (Eds.), *Gesture and sign language in human-computer interaction* (pp. 76–85). Berlin, Heidelberg: Springer Berlin Heidelberg.

Forster, J., Oberdörfer, C., Koller, O., & Ney, H. (2013). Modality combination techniques for continuous sign language recognition. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): vol. 7887 LNCS*, (pp. 89–99). Berlin, Heidelberg: Springer, http://dx.doi.org/10.1007/978-3-642-38628-2_10, URL: https://link.springer.com/chapter/10.1007/978-3-642-38628-2_10.

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., et al. (2012). RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the 8th international conference on language resources and evaluation, no. May* (pp. 3785–3789).

Gao, W., Fang, G., Zhao, D., & Chen, Y. (2004). Transition movement models for large vocabulary continuous sign language recognition. In *Sixth IEEE international conference on automatic face and gesture recognition, 2004. proceedings* (pp. 553–558). http://dx.doi.org/10.1109/AFGR.2004.1301591.

Gao, L., Li, H., Liu, Z., Liu, Z., Wan, L., & Feng, W. (2021). RNN-transducer based Chinese sign language recognition. *Neurocomputing*, *434*, 45–54. http://dx.doi.org/10.1016/j.neucom.2020.12.006.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on machine learning* (pp. 369–376). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/1143844.1143891.

Guilin, Y., Hongxun, Y., Xin, L., & Feng, J. (2006). Real time large vocabulary continuous sign language recognition based on OP/Viterbi algorithm. In *Proceedings - international conference on pattern recognition, vol. 3* (pp. 312–315). http://dx.doi.org/10.1109/ICPR.2006.954.

Guo, L., Xue, W., Guo, Q., Liu, B., Zhang, K., Yuan, T., et al. (2023). Distilling cross-temporal contexts for continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10771–10780).

Gweth, Y. L., Plahl, C., & Ney, H. (2012). Enhanced continuous sign language recognition using PCA and neural network features. In *IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 55–60). http://dx.doi.org/10.1109/CVPRW.2012.6239187.

Hao, A., Min, Y., & Chen, X. (2021). Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 11283–11292). http://dx.doi.org/10.1109/ICCV48922.2021.01111.

Hassan, M., Assaleh, K., & Shanableh, T. (2017). User-dependent sign language recognition using motion detection. In *Proceedings - 2016 international conference on computational science and computational intelligence* (pp. 852–856). http://dx.doi.org/10.1109/CSCI.2016.0165.

Hassan, M., Assaleh, K., & Shanableh, T. (2019). Multiple proposals for continuous arabic sign language recognition. *Sensing and Imaging*, *20*(1), 1–23. http://dx.doi.org/10.1007/s11220-019-0225-3.

Hassan, S., Seita, M., Berke, L., Tian, Y., Gale, E., Lee, S., et al. (2022). ASL-homework-RGBD dataset: An annotated dataset of 45 fluent and non-fluent signers performing American sign language homeworks. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, M. Schulder (Eds.), *Proceedings of the LREC2022 10th workshop on the representation and processing of sign languages: multilingual sign language resources* (pp. 67–72). Marseille, France: European Language Resources Association, URL: https://aclanthology.org/2022.signlang-1.11.

Hienz, H., Bauer, B., & Kraiss, K.-F. (1999). HMM-based continuous sign language recognition using stochastic grammars. In A. Braffort, R. Gherbi, S. Gibet, D. Teil, & J. Richardson (Eds.), *Gesture-based communication in human-computer interaction* (pp. 185–196). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hu, L., Gao, L., Liu, Z., & Feng, W. (2022). Temporal lift pooling for continuous sign language recognition. In *European conference on computer vision* (pp. 511–527). Springer.

Hu, L., Gao, L., Liu, Z., & Feng, W. (2023a). Continuous sign language recognition with correlation network. http://dx.doi.org/10.48550/ARXIV.2303.03202, URL: https://arxiv.org/abs/2303.03202.

Hu, L., Gao, L., Liu, Z., & Feng, W. (2023b). Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI conference on artificial intelligence, vol. 37, no. 1* (pp. 854–862).

Hu, L., Gao, L., Liu, Z., & Feng, W. (2024). Scalable frame resolution for efficient continuous sign language recognition. *Pattern Recognition*, *145*, Article 109903. http://dx.doi.org/10.1016/j.patcog.2023.109903, URL: https://www.sciencedirect.com/science/article/pii/S0031320323006015.

Hu, L., Gao, L., Liu, Z., Pun, C.-M., & Feng, W. (2023). AdaBrowse: Adaptive video browser for efficient continuous sign language recognition. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 709–718).

Hu, H., Pu, J., Zhou, W., Fang, H., & Li, H. (2024). Prior-aware cross modality augmentation learning for continuous sign language recognition. *IEEE Transactions on Multimedia*, *26*, 593–606. http://dx.doi.org/10.1109/TMM.2023.3268368.

Hu, H., Pu, J., Zhou, W., & Li, H. (2023). Collaborative multilingual continuous sign language recognition: A unified framework. *IEEE Transactions on Multimedia*, *25*, 7559–7570. http://dx.doi.org/10.1109/TMM.2022.3223260.

Huang, S., & Ye, Z. (2021). Boundary-adaptive encoder with attention method for Chinese sign language recognition. *IEEE Access*, *9*, 70948–70960. http://dx.doi.org/10.1109/ACCESS.2021.3078638.

Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. In *32nd AAAI conference on artificial intelligence* (pp. 2257–2264). arXiv:1801.10111.

Infantino, I., Rizzo, R., & Gaglio, S. (2007). A framework for sign language sentence recognition by commonsense context. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *37*(5), 1034–1039. http://dx.doi.org/10.1109/TSMCC.2007.900624.

Jachova, Z., Kovacheva, O., & Karovska, A. (2008). Differences between American sign language (ASL) and British Sign Language (BSL). *The Journal of Special Education and Rehabilitation*, *1*(2), 41–54.

Jang, Y., Oh, Y., Cho, J. W., Kim, D.-J., Chung, J. S., & Kweon, I. S. (2022). Signing outside the studio: Benchmarking background robustness for continuous sign language recognition. arXiv preprint arXiv:2211.00448.

Jang, Y., Oh, Y., Cho, J. W., Kim, M., Kim, D.-J., Kweon, I. S., et al. (2023). Self-sufficient framework for continuous sign language recognition. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5). IEEE.

Jebali, M., Dakhli, A., & Jemni, M. (2021). Vision-based continuous sign language recognition using multimodal sensor fusion. *Evolving Systems*, *12*(4), 1031–1044. http://dx.doi.org/10.1007/s12530-020-09365-y.

Jiao, P., Min, Y., Li, Y., Wang, X., Lei, L., & Chen, X. (2023). Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 20676–20686).

Kagirov, I., Ivanko, D., Ryumin, D., Axyonov, A., & Karpov, A. (2020). TheRuSLan: Database of Russian sign language. In *LREC 2020 - 12th international conference on language resources and evaluation, conference proceedings, no. May* (pp. 6079–6085).

Kelly, D., Reilly Delannoy, J., Mc Donald, J., & Markham, C. (2009). A framework for continuous multimodal sign language recognition. In *ICMI-MLMI'09 - proceedings of the international conference on multimodal interfaces and the workshop on machine learning for multimodal interfaces* (pp. 351–358). http://dx.doi.org/10.1145/1647314.1647387.

Ko, S. K., Kim, C. J., Jung, H., & Cho, C. (2019). Neural sign language translation based on human keypoint estimation. *Applied Sciences (Switzerland)*, *9*(13), 1–18. http://dx.doi.org/10.3390/app9132683, arXiv:1811.11436.

Koishybay, K., Mukushev, M., & Sandygulova, A. (2020). Continuous sign language recognition with iterative spatiotemporal fine-tuning. In *Proceedings - international conference on pattern recognition* (pp. 10211–10218). http://dx.doi.org/10.1109/ICPR48806.2021.9412364.

Koller, O., Bowden, R., & Ney, H. (2016). Automatic alignment of hamnosys subunits for continuous sign language recognition. In *LREC 2016 proceedings* (pp. 121–128).

Koller, O., Camgoz, N. C., Ney, H., & Bowden, R. (2020). Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(9), 2306–2320. http://dx.doi.org/10.1109/TPAMI.2019.2911077.

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, *141*, 108–125. http://dx.doi.org/10.1016/j.cviu.2015.09.013.

Koller, O., Ney, H., & Bowden, R. (2015). Deep learning of mouth shapes for sign language oscar. In *2015 IEEE international conference on computer vision workshop* (pp. 477–483). IEEE, http://dx.doi.org/10.1109/ICCVW.2015.69, URL: https://www.cv-foundation.org/openaccess/content_iccv_2015_workshops/w12/papers/Koller_Deep_Learning_of_ICCV_2015_paper.pdf.

Koller, O., Zargaran, S., & Ney, H. (2017). Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-hmms. In *Proceedings - 30th IEEE conference on computer vision and pattern recognition, vol. 2017-Janua* (pp. 3416–3424).

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2016). Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *British machine vision conference 2016, vol. 2016-Septe, no. September* (pp. 136.1–136.12). http://dx.doi.org/10.5244/C.30.136.

Kong, W. W., & Ranganath, S. (2014). Towards subject independent continuous sign language recognition: A segment and merge approach. *Pattern Recognition*, *47*(3), 1294–1308. http://dx.doi.org/10.1016/j.patcog.2013.09.014.

Koulierakis, I., Siolas, G., Efthimiou, E., Fotinea, S. E., & Stafylopatis, A. G. (2021). Sign boundary and hand articulation feature recognition in sign language videos. *Machine Translation*, *35*(3), 323–343. http://dx.doi.org/10.1007/s10590-021-09271-3.

Kumar, D. A., Sastry, A. S., Kishore, P. V., & Kumar, E. K. (2018). Indian sign language recognition using graph matching on 3D motion captured signs. *Multimedia Tools and Applications*, *77*(24), 32063–32091. http://dx.doi.org/10.1007/s11042-018-6199-7.

Li, R., & Meng, L. (2022). Multi-view spatial-temporal network for continuous sign language recognition. arXiv preprint arXiv:2204.08747.

Li, K., Zhou, Z., & Lee, C. H. (2016). Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. *ACM Transactions on Accessible Computing*, *8*(2), http://dx.doi.org/10.1145/2850421.

Liang, R. H., & Ouhyoung, M. (1998). A real-time continuous gesture recognition system for sign language. In *Proceedings - 3rd IEEE international conference on automatic face and gesture recognition* (pp. 558–567). http://dx.doi.org/10.1109/AFGR.1998.671007.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., et al. (2019). Mediapipe: A framework for building perception pipelines. arXiv:1906.08172.

Luqman, H. (2023). Arabsign: A multi-modality dataset and benchmark for continuous arabic sign language recognition. In *2023 IEEE 17th international conference on automatic face and gesture recognition* (pp. 1–8). IEEE.

Luqman, H., & El-Alfy, E. S. M. (2021). Towards hybrid multimodal manual and non-manual arabic sign language recognition: Marsl database and pilot study. *Electronics (Switzerland)*, *10*(14), 1–16. http://dx.doi.org/10.3390/electronics10141739.

Luqman, H., & Mahmoud, S. A. (2019). Automatic translation of Arabic text-to-Arabic sign language. *Universal Access in the Information Society*, *18*(4), 939–951. http://dx.doi.org/10.1007/s10209-018-0622-8.

Martínez, A. M., Wilbur, R. B., Shay, R., & Kak, A. C. (2002). Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. In *Proceedings - 4th IEEE international conference on multimodal interfaces* (pp. 167–172). Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/ICMI.2002.1166987.

Meng, X., Feng, L., Yin, X., Zhou, H., Sheng, C., Wang, C., et al. (2019). Sentence-level sign language recognition using RF signals. In *2019 6th international conference on behavioral, economic and socio-cultural computing* (pp. 1–6). http://dx.doi.org/10.1109/BESC48373.2019.8963177.

Min, Y., Hao, A., Chai, X., & Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11542–11551).

Min, Y., Jiao, P., Li, Y., Wang, X., Lei, L., Chai, X., et al. (2022). Deep radial embedding for visual sequence learning. In *European conference on computer vision* (pp. 240–256). Springer.

Mittal, A., Kumar, P., Roy, P. P., Balasubramanian, R., & Chaudhuri, B. B. (2019). A modified LSTM model for continuous sign language recognition using leap motion. *IEEE Sensors Journal*, *19*(16), 7056–7063. http://dx.doi.org/10.1109/JSEN.2019.2909837.

MMPose Contributors (2020). OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose.

Mocialov, B., Turner, G., Lohan, K., & Hastie, H. (2017). Towards continuous sign language recognition with deep learning. In *Proceeding of the workshop on the creating meaning with robot assistants: The gap left by smart device*. URL: https://github.com/CMU-Perceptual-Computing-Lab/openpose/.

Mukushev, M., Ubingazhibov, A., Kydyrbekova, A., Imashev, A., Kimmelman, V., & Sandygulova, A. (2022). FluentSigners-50: A signer independent benchmark dataset for sign language processing. *PLoS ONE*, *17*(9 September), 1–19. http://dx.doi.org/10.1371/journal.pone.0273649.

Niu, Z., & Mak, B. (2020). Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer vision – ECCV 2020* (pp. 172–186). Cham: Springer International Publishing.

Pan, T. Y., Lo, L. Y., Yeh, C. W., Li, J. W., Liu, H. T., & Hu, M. C. (2016). Real-time sign language recognition in complex background scene based on a hierarchical clustering classification method. In *Proceedings - 2016 IEEE 2nd International conference on multimedia big data* (pp. 64–67). http://dx.doi.org/10.1109/BigMM.2016.44.

Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2021). Artificial intelligence technologies for sign language. *Sensors*, *21*(17), http://dx.doi.org/10.3390/s21175843.

Papastratis, I., Dimitropoulos, K., & Daras, P. (2021). Continuous sign language recognition through a context-aware generative adversarial network. *Sensors*, *21*(7), http://dx.doi.org/10.3390/s21072437.

Papastratis, I., Dimitropoulos, K., Konstantinidis, D., & Daras, P. (2020). Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space. *IEEE Access*, *8*, 91170–91180. http://dx.doi.org/10.1109/ACCESS.2020.2993650.

Pei, X., Guo, D., & Zhao, Y. (2019). Continuous sign language recognition based on pseudo-supervised learning. In *MAHCI 2019 - Proceedings of the 2nd workshop on multimedia for accessible human computer interfaces, co-located with MM 2019* (pp. 33–39). http://dx.doi.org/10.1145/3347319.3356837.

Pu, J., Zhou, W., Hu, H., & Li, H. (2020). Boosting continuous sign language recognition via cross modality augmentation. In *MM 2020 - Proceedings of the 28th ACM international conference on multimedia* (pp. 1497–1505). http://dx.doi.org/10.1145/3394171.3413931, arXiv:2010.05264.

Pu, J., Zhou, W., & Li, H. (2018). Dilated convolutional network with iterative optimization for continuous sign language recognition. In *IJCAI international joint conference on artificial intelligence, vol. 2018-July* (pp. 885–891). http://dx.doi.org/10.24963/ijcai.2018/123.

Pu, J., Zhou, W., & Li, H. (2019). Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol. 2019-June* (pp. 4160–4169). http://dx.doi.org/10.1109/CVPR.2019.00429.

Rao, G. A., Kishore, P. V., Kumar, D. A., & Sastry, A. S. (2017). Neural network classifier for continuous sign language recognition with selfie video. *Far East Journal of Electronics and Communications*, *17*(1), 49–71. http://dx.doi.org/10.17654/EC017010049.

Rashid, N., & Albelwi, N. R. (2012). Real-time Arabic Sign Language (ArSL) recognition real-time Arabic Sign Language (ArSL) recognition. In *International conference on communications and information technology, no. June* (pp. 497–501).

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, *164*(February 2020), Article 113794. http://dx.doi.org/10.1016/j.eswa.2020.113794.

Rekha, J., Bhattacharya, J., & Majumder, S. (2012). Improved hand tracking and isolation from face by ICondensation multi clue algorithm for continuous Indian sign language recognition. In P. S. Thilagam, A. R. Pais, K. Chandrasekaran, & N. Balakrishnan (Eds.), *Advanced computing, networking and security* (pp. 106–116). Berlin, Heidelberg: Springer Berlin Heidelberg.

Roussos, A., Theodorakis, S., Pitsikalis, V., & Maragos, P. (2010). Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*: *vol. 6553 LNCS*, (no. PART 1), (pp. 258–272). Springer, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-642-35749-7_20, URL: https://link.springer.com/chapter/10.1007/978-3-642-35749-7_20.

Sarkar, S., Loeding, B., Yang, R., Nayak, S., & Parashar, A. (2011). Segmentation-robust representations, matching, and modeling for sign language. In *Proc. IEEE conf. on computer vision and pattern recognition workshops* (pp. 13–19).

Sharma, S., Gupta, R., & Kumar, A. (2021). Continuous sign language recognition using isolated signs data and deep transfer learning. *Journal of Ambient Intelligence and Humanized Computing*, (2020), http://dx.doi.org/10.1007/s12652-021-03418-z.

Slimane, F. B., & Bouguessa, M. (2020). Context matters: Self-attention for sign language recognition. In *Proceedings - international conference on pattern recognition* (pp. 7884–7891). http://dx.doi.org/10.1109/ICPR48806.2021.9412916, arXiv:2101.04632.

Starner, T., Weaver, J., & Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(12), 1371–1375. http://dx.doi.org/10.1109/34.735811.

Suliman, W., Deriche, M., Luqman, H., & Mohandes, M. (2021). Arabic sign language recognition using deep machine learning. In *2021 4th international symposium on advanced electrical and communication technologies* (pp. 1–4). IEEE, http://dx.doi.org/10.1109/isaect53699.2021.9668405.

Suri, K., & Gupta, R. (2019). Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory. *Computers & Electrical Engineering*, *78*, 493–503. http://dx.doi.org/10.1016/j.compeleceng.2019.08.006.

Tateno, S., Liu, H., & Ou, J. (2020). Development of sign language motion recognition system for hearing-impaired people using electromyography signal. *Sensors*, *20*(20), 5807.

Tolba, M. F., Samir, A., & Aboul-Ela, M. (2013). Arabic sign language continuous sentences recognition using PCNN and graph matching. *Neural Computing and Applications*, *23*(3–4), 999–1010. http://dx.doi.org/10.1007/s00521-012-1024-0.

Tripathi, K., & Nandi, N. B. G. (2015). Continuous Indian sign language gesture recognition and sentence formation. In *Procedia computer science*: *vol. 54*, (pp. 523–531). Elsevier, http://dx.doi.org/10.1016/j.procs.2015.06.060.

Tubaiz, N., Shanableh, T., & Assaleh, K. (2015). Glove-based continuous arabic sign language recognition in user-dependent mode. *IEEE Transactions on Human-Machine Systems*, *45*(4), 526–533. http://dx.doi.org/10.1109/THMS.2015.2406692.

Tuffaha, M., Shanableh, T., & Assaleh, K. (2015). Novel feature extraction and classification technique for sensor-based continuous arabic sign language recognition. In S. Arik, T. Huang, W. K. Lai, & Q. Liu (Eds.), *Neural information processing* (pp. 290–299). Cham: Springer International Publishing.

Vassilia, P. N., & Konstantinos, M. G. (2006). Multimodal continuous recognition system for greek sign language using various grammars. In G. Antoniou, G. Potamias, C. Spyropoulos, & D. Plexousakis (Eds.), *Advances in artificial intelligence* (pp. 584–587). Berlin, Heidelberg: Springer Berlin Heidelberg.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vogler, C., & Metaxas, D. (2001). A framework for recognizing the simultaneous aspects of American sign language. *Computer Vision and Image Understanding*, *81*(3), 358–384. http://dx.doi.org/10.1006/cviu.2000.0895.

Von Agris, U., Blömer, C., & Kraiss, K. F. (2008). Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, MLLR, and MAP. In *Proceedings - international conference on pattern recognition*. Institute of Electrical and Electronics Engineers Inc., http://dx.doi.org/10.1109/icpr.2008.4761363.

Von Agris, U., Knorr, M., & Kraiss, K. F. (2008). The significance of facial features for automatic sign language recognition. In *2008 8th IEEE international conference on automatic face and gesture recognition*. http://dx.doi.org/10.1109/AFGR.2008.4813472.

Von Agris, U., & Kraiss, K.-F. (2007). Towards a video corpus for signer-independent continuous sign language recognition. (pp. 1–6).

von Agris, U., Zieren, J., Canzler, U., Bauer, B., & Kraiss, K.-F. (2008). Recent developments in visual sign language recognition. *Universal Access in the Information Society*, *6*(4), 323–362. http://dx.doi.org/10.1007/s10209-007-0104-x.

Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, *28*(3), 785–813. http://dx.doi.org/10.1007/s11831-019-09384-2.

Wang, C., Gao, W., & Xuan, Z. (2001). A real-time large vocabulary continuous recognition system for Chinese sign language. In H.-Y. Shum, M. Liao, & S.-F. Chang (Eds.), *Advances in multimedia information processing* (pp. 150–157). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, Z., & Zhang, J. (2021). Continuous sign language recognition based on multi-part skeleton data. In *2021 international joint conference on neural networks, vol. 33, no. 12* (pp. 1899–1907). IEEE, http://dx.doi.org/10.3724/SP.J.1089.2021.18816.

Wei, F., & Chen, Y. (2023). Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 23612–23621).

Wei, C., Zhao, J., Zhou, W., & Li, H. (2021). Semantic boundary detection with reinforcement learning for continuous sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(3), 1138–1149. http://dx.doi.org/10.1109/TCSVT.2020.2999384.

Wei, C., Zhou, W., Pu, J., & Li, H. (2019). Deep grammatical multi-classifier for continuous sign language recognition. In *Proceedings - 2019 IEEE 5th International conference on multimedia big data* (pp. 435–442). IEEE, http://dx.doi.org/10.1109/BigMM.2019.00027.

WHO (2021). *World report on hearing*: *Technical report*, World Health Organization, URL: https://www.who.int/publications/i/item/world-report-on-hearing.

Xiao, Q., Qin, M., & Yin, Y. (2020). Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, *125*, 41–55. http://dx.doi.org/10.1016/j.neunet.2020.01.030.

Xie, P., Cui, Z., Du, Y., Zhao, M., Cui, J., Wang, B., et al. (2023). Multi-scale local-temporal similarity fusion for continuous sign language recognition. *Pattern Recognition*, *136*, http://dx.doi.org/10.1016/j.patcog.2022.109233, arXiv:2107.12762.

Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision – ECCV 2018* (pp. 318–335). Cham: Springer International Publishing.

Xie, P., Zhao, M., & Hu, X. (2021). PiSLTRc: Position-informed sign language transformer with content-aware convolution. *IEEE Transactions on Multimedia*, 1–13. http://dx.doi.org/10.1109/TMM.2021.3109665, arXiv:2107.12600.

Yang, H.-d., & Lee, S.-w. (2011). Combination of manual and non-manual features for sign language recognition based on conditional random field and a ctive appearance model and hand configurations , while fingerspellings are a combi indicates the ends of phrases , etc [12]. sign langu. In *Proceedings of the 2011 international conference on machine learning and cybernetics* (pp. 10–13).

Yang, R., Sarkar, S., & Loeding, B. (2007). Enhanced level building algorithm for the movement epenthesis problem in sign language recognition. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8). http://dx.doi.org/10.1109/CVPR.2007.383347.

Yang, R., Sarkar, S., & Loeding, B. (2010). Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 462–477. http://dx.doi.org/10.1109/TPAMI.2009.26.

Yang, Z., Shi, Z., Shen, X., & Tai, Y.-W. (2019). SF-net: Structured feature network for continuous sign language recognition. arXiv:1908.01341.

Yang, W., Tao, J., & Ye, Z. (2016). Continuous sign language recognition using level building based on fast hidden Markov model. *Pattern Recognition Letters*, *78*, 28–35. http://dx.doi.org/10.1016/j.patrec.2016.03.030.

Ye, L., Lan, S., Zhang, K., & Zhang, G. (2020). EM-sign: A non-contact recognition method based on 24 GHz Doppler radar for continuous signs and dialogues. *Electronics*, *9*(10), 1577.

Ye, Y., Tian, Y., Huenerfauth, M., & Liu, J. (2018). Recognizing American sign language gestures from within continuous videos. In *2018 IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 2145–214509). http://dx.doi.org/10.1109/CVPRW.2018.00280.

Yu, S. H., Huang, C. L., Hsu, S. C., Lin, H. W., & Wang, H. W. (2011). Vision-based continuous sign language recognition using product HMM. In *1st Asian conference on pattern recognition* (pp. 510–514). http://dx.doi.org/10.1109/ACPR.2011.6166631.

Yuan, Q., Geo, W., Yao, H., & Wang, C. (2002). Recognition of strong and weak connection models in continuous sign language. In *2002 international conference on pattern recognition, vol. 1* (pp. 75–78). http://dx.doi.org/10.1109/ICPR.2002.1044616.

Zaboli, S., Serov, S., Mestetskiy, L., & Nagendraswamy, H. S. (2021). Gesture recognition in sign language videos by tracking the position and medial representation of the hand shapes. In S. K. Singh, P. Roy, B. Raman, & P. Nagabhushan (Eds.), *Computer vision and image processing* (pp. 407–420). Singapore: Springer Singapore.

Zadghorban, M., & Nahvi, M. (2018). An algorithm on sign words extraction and recognition of continuous Persian sign language based on motion and shape features of hands. *Pattern Analysis and Applications*, *21*(2), 323–335. http://dx.doi.org/10.1007/s10044-016-0579-2.

Zhang, H., Guo, Z., Yang, Y., Liu, X., & Hu, D. (2023). C2ST: Cross-modal contextualized sequence transduction for continuous sign language recognition. In *2023 IEEE/CVF international conference on computer vision* (pp. 20996–21005). URL: https://api.semanticscholar.org/CorpusID:267021406.

Zhang, B., Müller, M., & Sennrich, R. (2023). SLTUNET: A simple unified model for sign language translation. In *The eleventh international conference on learning representations*. URL: https://openreview.net/forum?id=EBS4C77p_5S.

Zhang, Z., Pu, J., Zhuang, L., Zhou, W., & Li, H. (2019). Continuous sign language recognition via reinforcement learning. In *2019 IEEE international conference on image processing* (pp. 285–289). IEEE.

Zhang, C., Tian, Y., & Huenerfauth, M. (2016). Multi-modality American sign language recognition. In *2016 IEEE international conference on image processing* (pp. 2881–2885). http://dx.doi.org/10.1109/ICIP.2016.7532886.

Zhang, S., & Zhang, Q. (2021). Sign language recognition based on global-local attention. *Journal of Visual Communication and Image Representation*, *80*(July), Article 103280. http://dx.doi.org/10.1016/j.jvcir.2021.103280.

Zhang, L., Zhang, Y., & Zheng, X. (2020). Wisign: Ubiquitous American sign language recognition using CommercialWi-Fi devices. *ACM Transactions on Intelligent Systems and Technology*, *11*(3), http://dx.doi.org/10.1145/3377553.

Zhang, J., Zhou, W., & Li, H. (2014). A threshold-based HMM-dtw approach for continuous sign language recognition. In *ACM international conference proceeding series* (pp. 237–240). http://dx.doi.org/10.1145/2632856.2632931.

Zhang, J., Zhou, W., & Li, H. (2015). A new system for Chinese sign language recognition. In *2015 IEEE China summit and international conference on signal and information processing* (pp. 534–538). http://dx.doi.org/10.1109/ChinaSIP.2015.7230460.

Zheng, J., Wang, Y., Tan, C., Li, S., Wang, G., Xia, J., et al. (2023). Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23141–23150).

Zhou, Z., Lui, K. S., Tam, V. W., & Lam, E. Y. (2020). Applying (3+2+1)d residual neural network with frame selection for Hong Kong sign language recognition. In *Proceedings - international conference on pattern recognition* (pp. 4296–4302). http://dx.doi.org/10.1109/ICPR48806.2021.9412075.

Zhou, M., Ng, M., Cai, Z., & Cheung, K. C. (2020). Self-attention-based fully-inception networks for continuous sign language recognition. *Frontiers in Artificial Intelligence and Applications*, *325*, 2832–2839. http://dx.doi.org/10.3233/FAIA200425.

Zhou, Z., Tam, V. W., & Lam, E. Y. (2021). SignBERT: A BERT-based deep learning framework for continuous sign language recognition. *IEEE Access*, *9*, 161669–161682. http://dx.doi.org/10.1109/ACCESS.2021.3132668.

Zhou, Z., Tam, V. W., & Lam, E. Y. (2022). A cross-attention BERT-based framework for continuous sign language recognition. *IEEE Signal Processing Letters*, *29*, 1818–1822. http://dx.doi.org/10.1109/LSP.2022.3199665.

Zhou, H., Zhou, W., & Li, H. (2019). Dynamic pseudo label decoding for continuous sign language recognition. In *Proceedings - IEEE international conference on multimedia and expo, vol. 2019-July* (pp. 1282–1287). http://dx.doi.org/10.1109/ICME.2019.00223.

Zhou, H., Zhou, W., Qi, W., Pu, J., & Li, H. (2021). Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF conference on computer vision and pattern recognition* (pp. 1316–1325). Los Alamitos, CA, USA: IEEE Computer Society, http://dx.doi.org/10.1109/CVPR46437.2021.00137, URL: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00137.

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2021). Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, *9210*(c), 1–13. http://dx.doi.org/10.1109/TMM.2021.3059098.

Zhu, Q., Li, J., Yuan, F., & Gan, Q. (2022). Multi-scale temporal network for continuous sign language recognition. (pp. 1–10). URL: https://arxiv.org/abs/2204.03864.

Zhu, Q., Li, J., Yuan, F., & Gan, Q. (2023). Continuous sign language recognition via temporal super-resolution network. *Arabian Journal for Science and Engineering*, http://dx.doi.org/10.1007/s13369-023-07718-8.

Zuo, R., & Mak, B. (2022a). C2SLR: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5131–5140).

Zuo, R., & Mak, B. (2022b). Local context-aware self-attention for continuous sign language recognition. In *Proceedings of the annual conference of the international speech communication association, vol. 2022-September, no. September* (pp. 4810–4814). http://dx.doi.org/10.21437/Interspeech.2022-164.

Zuo, R., & Mak, B. (2024). Improving continuous sign language recognition with consistency constraints and signer removal. *ACM Transactions on Multimedia Computing, Communications and Applications*, *37*(4), http://dx.doi.org/10.1145/3640815, arXiv:2212.13023.