IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 April 2024, accepted 2 May 2024, date of publication 8 May 2024, date of current version 3 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3398806

SURVEY

Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches, Datasets, and Challenges

TANGFEI TAO, YIZHE ZHAO, TIANYU LIU, AND JIELI ZHU
School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Tangfei Tao (taotangfei@mail.xjtu.edu.cn)

This work was supported in part by Shaanxi Province Key Research and Development Project under Grant 2020KWZ-003.

ABSTRACT The Deaf are a large social group in society. Their unique way of communicating through sign language is often confined within their community due to limited understanding by individuals outside of this demographic. This is where sign language recognition (SLR) comes in to help people without hearing impairments understand the meaning of sign language. In recent years, new methods of sign language recognition have been developed and achieved good results, so it is necessary to make a summary. This review mainly focuses on the introduction of sign language recognition techniques based on algorithms especially in recent years, including the recognition models based on traditional methods and deep learning approaches, sign language datasets, challenges and future directions in SLR. To make the method structure clearer, this article explains and compares the basic principles of different methods from the perspectives of feature extraction and temporal modelling. We hope that this review will provide some reference and help for future research in sign language recognition.

INDEX TERMS Sign language recognition, traditional method, deep learning, SLR datasets.

I. INTRODUCTION

Helping the deaf lead a normal life is meaningful but complicated, for the number of people in this group is growing rapidly year by year all over the world. According to statistics from WHO [1], by 2021, about 1.5 billion people worldwide suffer from some degree of hearing loss, and about 430 million people need medical rehabilitation for hearing loss, including 34 million children. By 2050 [2], the number of people who are projected to have some degree of hearing loss will climb to 2.5 billion. On the other hand, the deaf community has difficulty integrating into society because of language barriers. They can only communicate through sign language, which few hearing people can read and understand. Therefore, sign language recognition can be very important to help hearing people understand the intentions of the deaf. One way is to recognize sign language manually. However, manual recognition is time-consuming and labor-intensive. The use of algorithm-based sign language recognition can largely avoid the above problem and provide sign language recognition for the deaf anywhere, anytime. It bridges the communication gap and contributes greatly to the integration of deaf people into society. Sign Language Recognition, as an instantiated research problem in the field of action recognition and trajectory tracking, encompasses a wide range of research areas, including image segmentation, key point extraction, temporal modelling, etc. The core problem of sign language recognition is how to transform a piece of sign language information into text, as shown in Fig. 2. It requires focus on not only hand gestures but also hand movement trajectories, body posture, facial expressions, etc. [3]. Thus, sign language recognition is a highly integrated and interdisciplinary study. By far, many scholars have made significant contributions to the field of sign language recognition and their methods have achieved excellent results. Therefore, we want to make a summary of these methods. There are also some review papers [4], [5], [6], [7], [8], [9], and [10] on sign language recognition published in recent years, however,

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

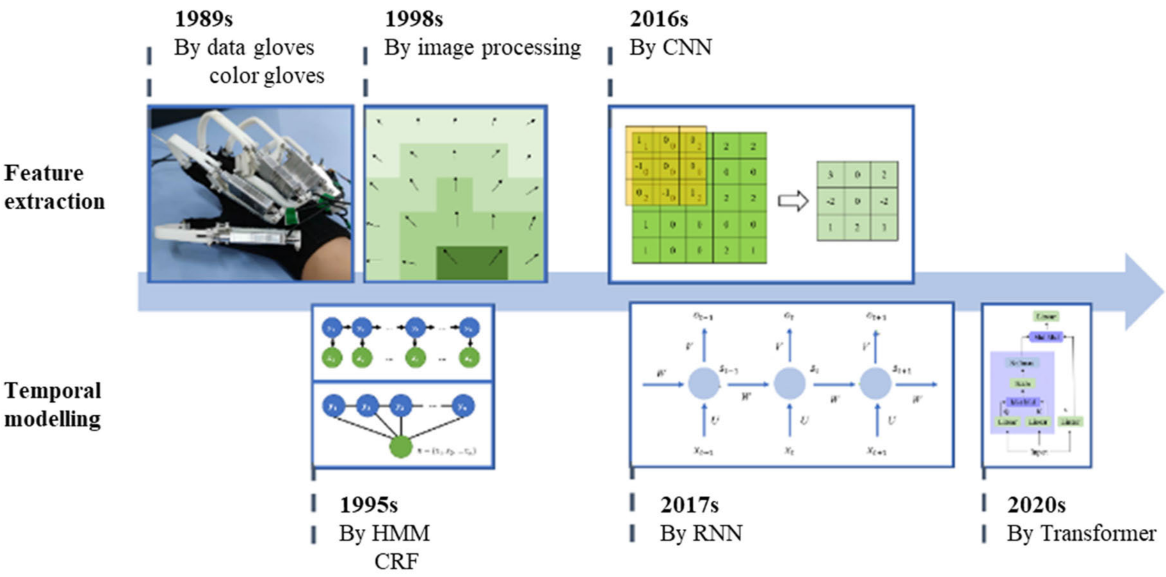

FIGURE 1. The whole development of sign language recognition.

some of them [6], [8], [9], [10] fail to give an introduction to the latest methods, like Transformer-based sign language recognition networks, etc. In addition, most of them [4], [6], [7], [8], [9], [10] do not elucidate the characteristics of the existing datasets. Most importantly, the structures of the methods are not classified clearly in [7], [8], and [10], and the pros and cons of different methods are not analyzed in [5] and [9]. In this paper, we have analyzed the current better methods and grouped them by structure from the perspectives of traditional methods and deep learning-methods. Specifically, we do not treat each method as a whole but split it into feature extraction section and temporal modelling section. We believe this would make the structure of methods much clearer.

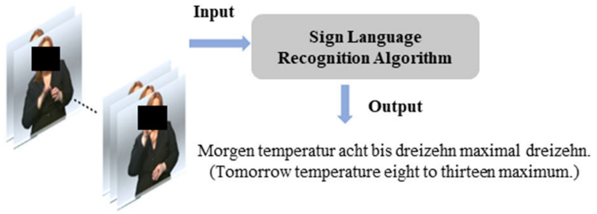
Input
Sign Language Recognition Algorithm
Output
Morgen temperatur acht bis dreizehn maximal dreizehn.
(Tomorrow temperature eight to thirteen maximum.)

FIGURE 2. The sign language recognition process (The sample in this figure is collected from the German sign language dataset RWTH-PHOENIX-Weather 2014).

The whole development of sign language recognition is shown in Fig. 1. In the early stage of SLR, some auxiliary sensing equipment, such as data gloves [11] and color gloves [12], are used to record key points or regions of the hand. These captured data are then subjected to classification and recognition processes using traditional feature extraction operators. This is the most basic method of recognition, using sensors to capture a range of features such as hand position, movement trajectory and speed. Then, sign language is recognized using some temporal model like Hidden Markov Model (HMM) [13] or other methods. However, the complexity of these devices limits their application. As a result, a number of methods [14] based on image processing have been proposed to get rid of the limitations of the device. Nevertheless, those operators used in image processing are not designed specifically for sign language, which cause another bottleneck in recognition due to the limited representational capacity. Sign language is a complex and delicate movement and therefore requires more precise features for its representation, while the development of deep learning in recent years has provided a solution to this problem. Deep learning has powerful feature extraction capabilities to fully exploit intrinsic features from limited data automatically, which is exactly what is required for sign language recognition. Scholars have designed a variety of sign language recognition networks [15], [16], [17] based on deep learning methods, and have achieved a significant improvement on recognition results in public sign language datasets. However, the current research on sign language recognition is equally deficient. The limitations of the datasets and the complexity of the algorithms make sign language recognition a long way from practical application.

sign language recognition covers a wide range of research areas, including object detection, trajectory tracking, pose estimation, action recognition and more, making it a highly comprehensive research area. From another point of view, sign language recognition, especially continuous sign language recognition, is a typical seq2seq problem. The focus of solving the sign language recognition problem is how to establish the mapping relationship between two sequences. There are still many unsolved problems with sign language recognition, such as the recognition of unseen sentences. These problems are the main focus of today's research. In this

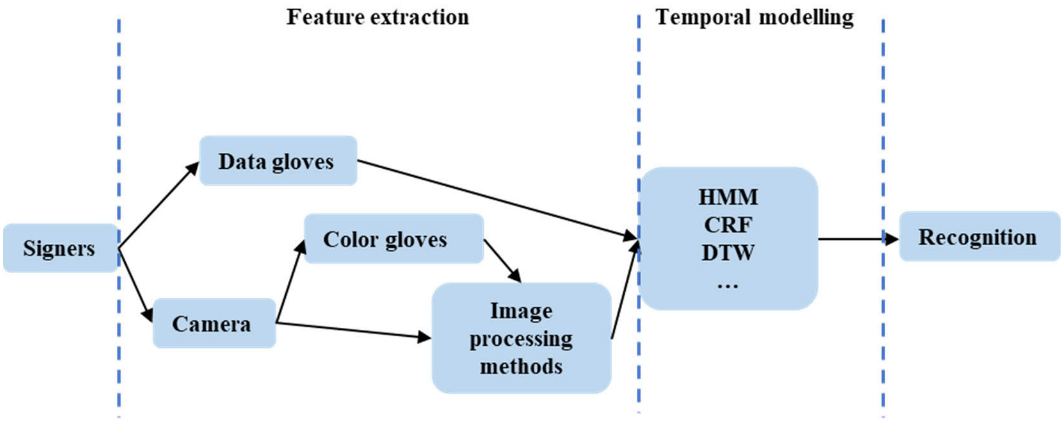

FIGURE 3. The process of the traditional sign language recognition.

paper, we try to highlight the important parts of the above issues and analysis of current challenges and directions for sign language recognition that need to be addressed. We list the contributions of this paper as follows:

1) Summarize many representative papers, including the latest approaches based on generative adversarial network (GAN), Graph Convolutional Network (GCN), Transformer;
2) Break down the different method structures into a feature extraction section and a temporal modelling section, and analyze them from the perspectives of traditional method and deep-learning method;
3) Present the common and latest sign language recognition datasets, with a description of the distinctive parts of each dataset;
4) Analyze the existing challenges and future developing trend of sign language recognition.

We hope to present a comprehensive review of sign language recognition based on traditional methods and deep learning methods. In the traditional-method part, we present feature extraction methods in three areas: data gloves, color gloves and image processing, and introduce temporal processing methods such as Hidden Markov Model (HMM), Conditional Random Field (CRF), etc. In the deep-learning-method part, we present the main framework for sign language recognition, such as Convolutional Neural Network (CNN), Transformer, etc., and give common methods for continuous sign language recognition. In addition to the methods described above, we present evaluation metrics and datasets in recent years, and analyze the current challenges with sign language recognition. The remainder of this paper is organized as follows. Section II introduces the sign language recognition based on traditional methods. In Section III, methods based on deep learning are presented and analyzed. The datasets, challenges and future directions are given in Section III-C and Section IV respectively. Finally, conclusions are provided in Section V.

II. SIGN LANGUAGE RECOGNITION BASED ON TRADITIONAL METHODS

In a sign language presentation video, the basic unit of sign language information is the gloss, which is a complete sign language word containing several gestures. The task that contains only one gloss per video is called Isolated Sign Language Recognition (ISLR). If the message in a video is a sentence consisting of several glosses, this is called Continuous Sign Language Recognition (CSLR). The sign language message is conveyed by a continuous sequence of gestures, with specific meanings conveyed through each different gesture and the interrelationship between the gestures. Therefore, in recognition tasks, it needs to be clear what the gesture is and what the relationship is between the gesture and the gesture. These are two basic components of sign language recognition: feature extraction and temporal modelling. The process of traditional sign language recognition can be summarized as Fig. 3. The specific method of the traditional sign language recognition is often done with the help of some sensors placed on data gloves. The features are obtained by capturing the changes of body parts such as hands, body posture, etc. by these sensors. There are also some ways to get the features by using the image processing approaches with the data collected by color gloves and cameras. After that, some temporal models, such as Hidden Markov Model (HMM), Conditional Random Field (CRF), etc., are used to predict the result. In this section, we present the sign language recognition based on these traditional methods.

A. FEATURE EXTRACTION

The purpose of feature extraction is to get the key information. Most of the sensor-based methods get information such as hand position, angle, movement speed, etc. and use them as hand features; while the image-processing-based methods can get deeper information such as gradient, contour, histogram, etc.

1) METHOD BASED ON GLOVES

The gloves used in traditional SLR are mainly data gloves and color gloves, which were proposed by scholars as early as the last century. The data glove, which is shown in Fig. 4(a), mainly uses attached sensors to collect hand shape, position and other information as a basis for judging hand gestures.

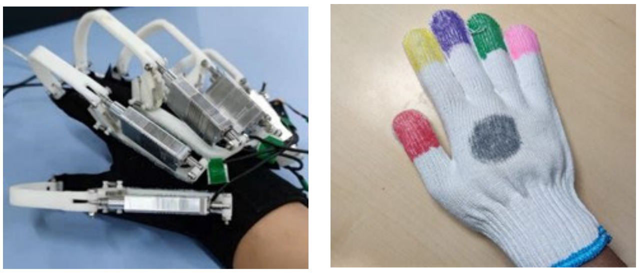
(a) The data glove (b) The color glove

FIGURE 4. The data glove and the color glove.

Waldron and Kim [18] obtained the hand shape and position data, including the coordinates of the hand in space and the pitch, roll, yaw of the hand, etc., from a data glove mounted with a Polhemus sensor and input them into a two-stage neural network to recognize isolated American Sign Language. The system proposed by them is shown in Fig. 5. This is a typical traditional sign language recognition system based on data gloves. It contains almost all relevant shallow features of the hand derived by sensors. Fu et al. [19] used a data glove that captures not only the gesture of the hand but also the motion information of the forearm, which was later trained to classify the numbers 0-10 using a BP neural network.

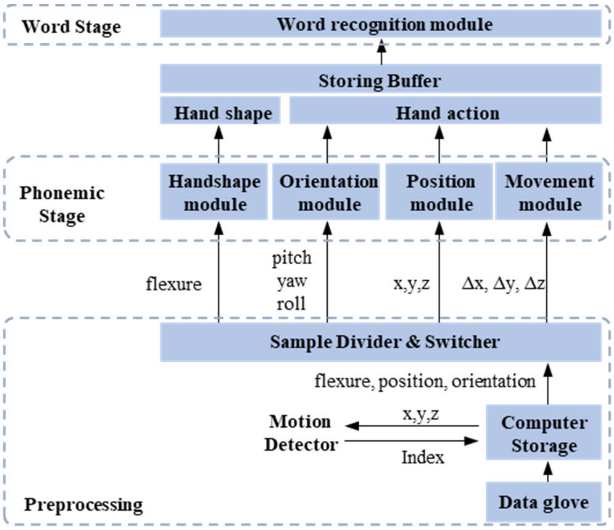

FIGURE 5. The sign language recognition system.

As for continuous sign language recognition, the research can be dated back to 2002. Fang and Gao et al. [20] segmented the video into glosses and then recognized them one by one. The results show that this divide-and-conquer approach worked well in isolated sign language recognition after segmentation and was capable of dealing with some short sentences. The team used the same idea in 2007 and proposed transition movement models (TMMs), which aimed to recognize the ME (movement epenthesis, i.e., interval signs between two sign glosses) after clustering it by its similar end-start sequences between two signs, to handle transition parts between two adjacent signs in large-vocabulary continuous sign language recognition [21].

The data glove method is accurate but complex and heavy. It captures hand information quickly but ignores facial expression and body posture. Validation on small datasets limits its credibility. To minimize the limitation and impact of the equipment on the signers, some scholars have proposed the use of color gloves, which are shown inFig.4(b), to replace data gloves.

Instead of using sensors, color gloves use different color areas to represent different key areas of the hand, so they are often used jointly with the camera. Okayasu et al. [22] assigned distinct colors to glove parts, locating each part using the colored region's center of gravity. They utilized an optical camera to obtain hand data like trajectory, position, and velocity, feeding it into Hidden Markov Model (HMM) for ranking likelihood.

Researchers explored color-glove-based continuous sign language recognition. Bauer and Hienz et al. [23] used two gloves, one multicolored for the dominant hand and the other unicolored. They extracted location, handshape, and orientation features, feeding them into HMM, and marking early continuous sign language recognition exploration. This work confirmed HMM feasibility and maintained natural sign flow, crucial for practical use. The outcomes of this research substantiated the feasibility of HMM and upheld the preservation of natural pace of sign language, a pivotal aspect for practical applicability.

Color glove based features mainly include position, angle, and handshape. While it simplifies the complexity of the device, strict background conditions and lack of portability limit its practical use. This method aims to reduce device limitations but does not solve the problem thoroughly. Instead, it sacrifices recognition accuracy, making it unpopular.

2) METHOD BASED ON TRADITIONAL IMAGE PROCESSING METHODS

To overcome the limitations of gloves and sensors for signers, researchers are using image processing to extract features directly from images. Digital image processing extracts richer features that enable the recognition of different sign languages. This approach allows for isolated and continuous sign language recognition, and although it's less accurate than sensors, it can greatly reduce the constraints of the device on the signer, making the method more widely applicable.

In the part of the isolated sign language recognition. Lin and Ding et al. [24] extracted histograms of oriented gradients (HOG) features from hand and background images to train a support vector machine (SVM) for hand detection. They then combined hand position, velocity, and angle to form a 4-dimensional vector as the trajectory feature, and calculated Mahalanobis distance for gesture recognition. Auephanwiriyakul et al. [25] utilized Scale Invariant Feature Transform (SIFT) for key point descriptors to match input frames with standard gestures, using HMMs to translate sequences into words. Rahim and Miah et al. [26] proposed an optimal segmentation method for identifying hand gestures from input images by comparing various segmentation techniques including YCbCr, SkinMask, and HSV (hue, saturation, and value). They processed the images using the selected method and utilized the resultant images as inputs to the model, thereby enhancing recognition performance. Subsequently, they devised a concatenated segmentation approach leveraging the YCbCr, HSV, and the watershed algorithm. Additionally, multiple data augmentation techniques were employed to enhance the model's generalization capabilities [27], [28]. Ming [29] combined the RGB and depth information, putting forward 3D Mesh MoSIFT feature descriptor to extract the information of hand motion affected by occlusions and subtle changes in local areas. Lim et al. [30] proposed a method called the block-based histogram of optical flow (BHOF) which establishes the histogram of optical flow of right hand and left hand as features. Katoch et al. [31] created the Bag of Visual Words (BOVW), which is similar to the Natural Language Processing (NLP) Bag of Words (BOW) but uses image features instead of words, by taking advantage of Speeded Up Robust Features (SURF) method to extract the features.

For continuous sign language recognition, Yang et al. [32] used the mixed Gaussian model [33] to complete the classification of skin pixel. They then used the histogram and processed by principal component analysis (PCA), of the distance from each point of the detected hand contour to a particular reference point in the picture as features. Yu et al. [34] utilized CamShift to determine the size and position of moving hands, employing the color histogram mode for tracking moving objects. Additionally, they manually designated three regions (head, chest, and bottom) as gesture regions. The characterization of the hand's shape involved the utilization of 7Hu moments, and its orientation was ascertained by calculating the angle between the hand's long axis and the x-axis. Integration of these features, combined with the count of hands, facilitated the creation of hybrid feature vectors. The two methods mentioned above try to recognize the continuous sign language by segmenting it into isolated glosses and eliminating the effects of movement epenthesis (ME). While some scholars tried to recognize the sequence holistically without these segmentations. Koller et al. [35] used HOG-3D Features [36] to capture the edges of the hands spatially and temporally. As for trajectories, they calculated the covariance matrix of velocity vectors within a time window and use the eigenvalues of the matrix to characterize the motion. In addition, they extracted seven continuous distance measurements across landmarks around the signer's face as high-level face features using active appearance model (AAMs) [37]. Hassan et al. [38] conducted research on both their sensor-based and vision-based datasets. For the sensor-based dataset, they used window-based statistical feature extraction techniques, calculating the mean and standard deviation to serve as features. For the vision-based dataset, they first detected motion through pixel difference analysis and selected optimal thresholds to transform image differences into binary images. Subsequently, features were extracted using 2D Discrete Cosine Transform (DCT).

It can be concluded that the traditional image processing approach is effective for sign language recognition, but feature design is complex and lacks robustness. Although some image processing approaches such as HOG, SURF and SIFT may increase the number of features to a large extent, they are not specifically designed for extracting hand features and some features gathered from these methods are even invalid, which means they still have their limitation when facing the large amount of sign language data.

B. TEMPORAL MODELING

Temporal modelling is a necessary part when dealing with the sequence problem, which is exactly what sign language recognition needs. Given that video is the main input for recognition, understanding the relation among video frames becomes vital for both continuous and isolated sign language recognition. This is particularly significant in continuous SLR where labels change over time. Hidden Markov Models or Conditional Random Fields approaches are the most commonly used traditional methods. In addition, there are some other methods such as Dynamic Time Warping methods.

1) HMM AND CRF

The most commonly used methods are Hidden Markov Models (HMM) and Conditional Random Fields (CRF), which are adopted from the area of natural language processing (NLP).

The utilization of HMM in SLR can be dated back to 2000 or even earlier. Bauer and Hienz et al. [23] designed the different HMM models for each sign and validated on a lexicon of 97 signs of German sign language. Gao et al. [39] proposed self-organizing feature maps (SOFM)/HMM for modeling signer independent isolated signs. Maebatake et al. [40] put forward a method using multi-stream HMM for sign language recognition, modeling the information of hand positions and movements respectively. To enhance the correlation between variables in different streams in multi-stream HMM, Theodorakis et al. [41] applied Product-HMMs (PHMM) for partial asynchrony between streams. Park and Lee et al. [42] proposed a cascade of two HMMs to recognize the point gesture.

The implementation of Conditional Random fields in sign language recognition can be dated back to 2006 or even earlier. Yang and Sarkar et al. [43] segmented the video

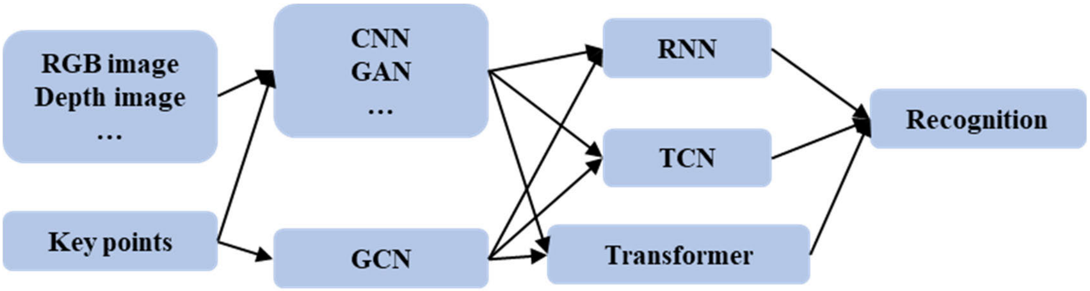

FIGURE 6. The process of the sign language recognition based on deep learning.

sequence into coarticulation frames (which has the same meaning as ME) and sign frames manually and fed key frames into CRF. Yang and Lee et al. [44] proposed a two-layer CRF structure which consisted of a T-CRF [45] and a conventional CRF. The T-CRF can discriminate signs, fingerspelling and non-sign patterns with the information of hand motion and place of articulation, and the conventional CRF can recognize subsign patterns between signs. They validated their approach on both isolated sign language and continuous sign language. Kong and Ranganath et al. [46] used a segmentation algorithm [47] based on minimum velocity and maximum directional angle change in movement channel to segment the continuous sign language sequence and classified the results into sign or ME. They then used a two-layer CRF to recognized the sign part of the segmentation. The first layer concentrates on the phoneme level consisted the information of handshape, movement, orientation and location, and the second layer is a Semi-Markov CRF for sign recognition. It can be concluded that most methods for continuous sign language use a two-stage process: segmenting gloss signs and intervals first, then recognizing them. Few scholars use CRF for sign language recognition due to its complexity in feature selection and training.

2) DYNAMIC TIME WARPING

Dynamic Time Warping (DTW) was originally used to measure sequence similarity in speech recognition. Sign language faces similar issues due to different signing speeds, leading to attempts at DTW-based recognition.

The DTW algorithm is mostly used for isolated sign language recognition, as isolated sign language recognition can provide a standard reference for verification data. It is generally not applied to continuous sign language because there is too much variation in sentence and the sign language sentences used for verification may not appear in the training data. Mathur and Sharma et al. [47] collected key frames from videos and extracted features of hand trajectory, then used DTW to recognize sign language. Wöllmer et al. [48] proposed a three-dimensional dynamic time warping (3D-DTW) algorithm to synchronize multimodal data, overcoming the computational complexity of AHMM. For continuous sign language recognition, the testing sentence may differ from the reference and it is impossible to list all sentences in the dataset. But if one sentence can be segmented into the units of glosses in advance, DTW may still be useful in the area of continuous sign language recognition.

3) OTHER METHODS

The distance comparison is a less commonly used method. The main purpose of this method is to accomplish the task of sign language recognition by comparing the differences between the feature vectors of isolated sign languages and the standard sign language feature vectors. Lin and Ding et al. [24] extracted gesture features and trajectory features of the standard sign language to build up a database. Then they recognized the gesture by extracting the same features of the given sequence and compared the Mahalanobis distance with the standard one. Ming [29] introduced 3D Mesh MoSIFT feature extraction method by firstly detecting key points in mesh domain which is transformed from 3D point clouds data, and then computed the 3D gradient and 3D motion features by calculation of image gradient along the horizontal and vertical directions, and subtraction of corresponding points in depth image. The Levenshtein distance is used to measure the similarity of predicted label and truth label finally. We could see that two methods mentioned above all cope with gesture recognition. This is because the feature space created by these methods is insufficient to represent the more complex sign language information. What's more, these methods are not robust enough if the interclass distance is not well controlled.

C. CONCLUSION

We introduced traditional sign language recognition methods with feature extraction and temporal modeling. Comparing results was difficult due to different datasets. Most of the datasets used in the above methods are small in size and lacked practicality. Traditional methods had limitations in hand-specific feature extraction and were sensitive to factors like illumination and occlusion. Manual feature design was costly and time-consuming, leading to accuracy bottlenecks. However, the traditional approach is more interpretable, facilitating the researcher to explore the importance of different features and is instructive.

III. SIGN LANGUAGE RECOGNITION BASED ON DEEP LEARNING

The typical deep-learning-based methods were proposed after 2012, where scholars began to extract hand features automatically using deep neural network. Deep-learning-based sign language recognition is able to mine the data for deeper information and automatically generate the features that best represent the feature of gesture. This is extremely important for improving the accuracy of sign language recognition. However, this approach requires large-scale datasets and computing power to support. Nowadays, the emergence of large-scale datasets and the increasing computing power of GPUs have given deep learning-based sign language recognition methods a lot of room for development. And some problems like weakly supervised learning of continuous sign language recognition can be resolved well, so more and more well-performed end-to-end models are starting to emerge.

The study of sign language recognition involves a range of elements such as action recognition and trajectory tracking.

Since sign language videos are composed of a series of frames, many typical convolutional neural networks (CNN) are often used for feature extraction like GoogLeNet [49], ResNet [50] and etc. In terms of temporal feature extraction, a number of temporal networks have been applied to temporal modelling like recurrent neural network (RNN), temporal convolutional network (TCN) [51] and etc. In addition to the extraction of spatial and temporal features separately, some 3D convolutional neural networks have been proposed for the simultaneous extraction of spatio-temporal information such as C3D [52] and I3D [53] and etc. Apart from the approaches mentioned above, some methods integrate traditional methods into deep learning and show good performance. Nowadays, many scholars make use of the Transformer [54] or its variants for sign language recognition and similarly obtain better recognition results. The process of sign language recognition based on deep learning is shown in Fig. 6.

A. EVALUATION METRICS

For isolated sign language recognition, commonly used evaluation metrics are accuracy rates. Here, we focus on common evaluation metrics for continuous sign language recognition. Commonly used metrics for continuous sign language recognition include Word Error Rate (WER), BLEU (Bilingual Evaluation Understudy) [55] and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [56], mostly inherited from the field of speech recognition translation.

1) WER (WORD ERROR RATE)

The WER is the percentage of the total number of words that need to be replaced, deleted or inserted in order to make the recognized word sequence consistent with the original word sequence. The expressions are shown as follows:

$$WER = 100 \times \frac{Substitution + Deletion + Insertion}{Total} \tag{3}$$

In other words, WER indicates the proportion of words that need to be altered to align the recognized text with the correct reference text. The definition of WER shows that the lower the metric, the better the recognition. It is important to note that the WER can be greater than 100% as the number of words to be replaced, deleted or inserted may be greater than the total number of words in the original sequence. The WER metric is commonly used for sequence recognition tasks and therefore for sign language recognition, most methods use WER as an evaluation metric.

2) BLEU (BILINGUAL EVALUATION UNDERSTUDY)

BLEU is a precision-based metric for evaluating similarity, utilized to analyze the extent to which n-grams in the candidate translation appear in the reference translation. The idea of BLEU is to calculate precision by comparing n-gram model between the output and the reference, and the expressions are as in (1), shown at the bottom of the page. The *candidates* in the formula are the sentences generated by the models, and they will be compared with the references to get the final scores. *n-gram* means the clip formed by n words adjacent to each other. The common idea of the BLEU metric is to calculate the percentage of n-gram clips in candidate, which appear both in candidates and references. Depending on the size of n, the commonly used metrics are BLEU-1, BLEU-2, BLEU-3 and BLEU-4. When n is smaller, the accuracy of the words is measured; when n is larger, the fluency of the sentences and the accuracy of the syntactic structure are measured. To make a more balanced result, the BLEU with different n are often sum up by weighting.

3) ROUGE (RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION)

ROUGE measures the extent to which the content of the reference summaries is covered in the system's output, primarily focusing on whether the machine-generated summaries have captured the information from the reference summaries. The ROUGE evaluation metric is similar to the BLEU metric, but it measures recall, that is the percentage of n-gram clips in reference, which appear both in candidates and references. The expressions are as (2), shown at the bottom of the page.

$$BLEU_n = \frac{\sum_{C\in\{Candidates\}} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in\{Candidates\}} \sum_{n-gram'\in C'} Count_{clip}(n-gram')} \tag{1}$$

$$ROUGE_n = \frac{\sum_{S\in\{ReferenceSummaries\}} \sum_{n-gram\in S} Count_{match}(n-gram)}{\sum_{S\in\{ReferenceSummaries\}} \sum_{n-gram\in S} Count(n-gram)} \tag{2}$$

Although some translation models could generate fluent sentences under the metrics of BLEU, the meanings of some generated sentences are too far apart from the references. The recall metrics ROUGE emerged as a solution to the problem of recognizing what is actually needed. There are also some variants like ROUGE-L (Longest common subsequence), ROUGE-W (Weighted longest common subsequence) and ROUGE-S (Skip-bigram co-occurrence statistics). The emergence of these metrics complemented the accuracy metrics, and to this day both BLEU and ROUGE metrics and their variants remain the main metrics in the field of sign language recognition.

Strictly, WER focuses on the overall accuracy of recognition but does not take into account the types of errors nor their potential impact on the semantic meaning of sentences. In contrast, BLEU and ROUGE focus on consistency in recognition outcomes. BLEU overlooks the semantic and syntactic correctness, emphasizing primarily the literal correctness of the translation outputs. Therefore, BLEU is more apt for assessing the global quality of translation rather than local accuracy or fluency. ROUGE, on the other hand, emphasizes the completeness of content and information encapsulated within the reference summaries. In addition, there are other evaluation metrics such as METEOR, CIDEr (Consensus-based Image Description Evaluation), etc. but they are not commonly used.

B. DEEP-LEARNING-BASED SIGN LANGUAGE RECOGNITION METHODS

The deep learning approach has the ability to express a larger feature space and therefore achieves better results in both ISLR and CSLR. In addition, deep learning methods can focus not only on manual components such as hand shape, orientation, and movement, but also on non-manual components such as mouthing, eye gaze, and eyebrow movements. This greatly enriches the features of the data, resulting in a significant improvement in recognition accuracy. The specific steps of recognition can be divided into pre-processing, feature extraction and temporal modelling, and recognition.

1) PRE-PROCESSING

A well-designed neural network has powerful redundancy removal capabilities, so that only the full image is needed as input to obtain deep features. However, the full image input is computationally intensive and makes the network take longer to converge. So how to balance this contradiction is important. There are usually two methods of pre-processing, one is to crop the hand region in the image and feed it into the neural network for training, and the other is to feed the full frame into the network.

Koller et al. [17] experimented and pointed out that using full frame as input gives better results than using cropped hand images. This is because complete frame information can provide more information than local features, such as hand trajectories, facial expressions, etc. In the later research [57], they cropped a rectangle of 92 × 132 pixels around the center of the hand on the RWTH-PHOENIX-Weather dataset [58] as the input of the network. De Coster et al. [59] cropped hand regions by OpenPose BODY-135 model [60]. To alleviate the noise that may occur near hand key points, they determine a suitable location for the hand crop in the extension of the forearm: based on the position of the elbow and wrist key points. Cihan Camgoz et al. [61] combined two SubUNets, which concentrated on hand patches and full frames respectively, to recognize continuous sign language. The WER indicator decreased by up to 2.4 and 1.0 compared to using hand patch and full frame separately, reaching the value of 42.1.

Although hand patches can be regarded as a more relevant information about hand, the size of cropping region needs to vary according to the datasets or real scenarios. However, the distances between the signers and camera are almost same within a given dataset, so it is unnecessary to change the size of the cropping box. Nevertheless, if the dataset is changed, cropping box with the original size may not be able to cover the hand regions in the new one. This may affect the recognition results.

2) FEATURE EXTRACTION AND TEMPORAL MODELLING

In the realm of feature extraction, 2D CNNs (often combined with temporal modeling methods such as RNNs) and 3D CNNs have been extensively employed in sign language recognition due to their robust feature extraction capabilities. Concurrently, the rising popularity of transformers has also demonstrated increasingly promising performance in this domain. Consequently, this section provides a comprehensive review of these two categories of methods, followed by an exploration of other approaches such as GNNs and GANs. When it comes to features, deep learning-based sign language recognition methods often consider three types of information: RGB information, depth information, and skeletal key points information. All of the information can be collected easily by somatosensory equipment like Kinect, especially for RGB information which can even be obtained by camera on mobile phone. This is very beneficial for the portability of the sign language recognition system. On the other hand, the complementarity of different information allows for more accurate recognition. It is important to make tradeoffs. Accordingly, this section concludes with an examination of the various input modalities in sign language datasets, emphasizing the importance of striking a balance between them.

a: METHODS BASED ON CNN AND RNN

One of the commonly used methods is combining CNN and RNN. Convolutional Neural Networks (CNNs) are a multilayer perceptron variant inspired by biology. Their structure includes multiple layers such as convolutional, pooling and fully connected layers. As the central component of a CNN, the convolutional kernel slides over the input image and extracts features by computing dot products with localized regions of the image. As the depth of the network increases, the CNNs are able to capture a wider range of contextual information and more complex features. Through training, CNNs are able to extract features from low to high levels layer by layer and utilize these levels of features for effective prediction or classification. Due to their excellent performance, CNNs have been widely used in computer vision fields such as image classification and have achieved remarkable results. As for sign language recognition, 2D CNNs are usually used to extract features within a single frame or for gesture recognition in image data [62], while for the action context represented by a sequence of frames, recurrent neural networks (RNNs) are often utilized for interpretation. RNNs sequentially process elements of an input sequence, computing the current hidden state based on both the current input and the previous hidden state. This mechanism enables them to capture long-term dependencies within sequences. Nevertheless, standard RNNs often struggle to effectively capture such dependencies in practical applications, primarily due to issues like gradient vanishing or explosion.To address these challenges, variants like Long Short-Term Memory Networks (LSTMs) and Gated Recurrent Units (GRUs) have been introduced. These architectures incorporate sophisticated gating mechanisms, facilitating the effective capture of information across longer sequences. Reference [17] and [57] all use GoogLeNet pretrained on ILSVRC dataset [63] as the backbone and both of them used RGB image as the input to recognize continuous sign language. Reference [17] combines CNN-LSTM and HMM, using the former part to calculate the maximum likelihood of input image and converting it to emission probabilities of HMM by Bayes' rule. Then EM algorithm is used to train CNN-LSTM iteratively. Experiments on different LSTM structures are also carried out. Although the models with more layers and number of hidden units showed better performances, it made the network hard to train. The authors later [57] compared the tandem approach (an intermediate step between GMM (Gaussian Mixed Model)-HMM and the hybrid CNN-HMM) with hybrid CNN-HMM and found that the hybrid CNN-HMM showed better performance on computational cost. Shanableh [64] proposed a two-stage approach, which first detects the number of words in a sentence for segmentation, and each word is converted into a motion image, so each motion image can contain traces of previous or successive words. Then, Inception-v3 is used to extract features from motion images and BiLSTM is used for recognition. Pu [65] considered both the information of skeleton points and RGB obtained by Kinect to recognize isolated sign language. He utilizes LeNet [66] and 3D-CNN based on AlexNet to extract the features of trajectories and handshapes, reaching a Top-1 accuracy of 0.858 on CSL dataset [67]. However, He did not give the solution when the Kinect failed to detect the hand region, and there was no explanation on how to distinguish left hand and right hand. In two studies by Hu et al. References [68] and [69], both utilized 2D CNNs to extract frame features, followed by the use of 1D CNNs and BiLSTM for short-term and long-term temporal modeling. In [68], they proposed an identification module to emphasize informative regions in each frame that are beneficial in expressing a sign, along with a correlation module to capture cross-frame trajectories. These modules were placed after each stage of the feature extractor to recognize body trajectories between adjacent frames. Meanwhile, in [69], they introduced the spatial self-emphasizing module (SSTM) and temporal self-emphasizing module (TSEM), which were integrated into each block of the feature extractor to emphasize spatial and temporal features, respectively.

Some researchers perform 3D convolution on videos directly to learn spatio-temporal information. References [70], [71], and [72] all used 3D-CNN to extract features. Sarhan et al. [71] used RGB video data as the input and obtained the optical flow stream from it. They then utilized two I3D networks to extract features from two streams above, and the predictions of each stream are averaged during the evaluation to give the final label. Pu et al. [70] processed the RGB information of the continuous sign language video by 3D-ResNet and then fed the features to the encoder-decoder network. Zhou et al. [72] proposed (3+2+1)D ResNet Model, which combines a 3D ResNet to extract spatial and temporal information simultaneously, a 2D ResNet to extract features spatially and a 1D convolutional network to extract features temporally. Its recognition results on the Hong Kong Sign Language (HKSL) dataset they proposed and CSL dataset can reach up to 94.6% and 96.0% respectively. The 1D convolutional network for temporal feature extraction is in fact a Temporal Convolutional Network (TCN) [51]. Compared to the serial processing of RNNs, TCN can process temporal information in parallel. In addition, compared with 3D-CNN, it can effectively compress the amount of data while increasing the computational speed. Gao et al. [73] utilized this structure as well. They processed the image by a 2D discrete wavelet transform to enhance the image before inputting the RGB video sequence. The order of their (2+1)D and 3D modules is different from [72] and a residual module was also applied. Their Top-1 accuracy results can be up to 98.4%. Han et al. [74] also used R (2+1) D for separate spatial and temporal modeling. In addition, they proposed a lightweight spatial-temporal-channel attention module that enables the network to focus on the significant information along spatial, temporal, and channel dimensions.Cui, R., et al. [75] used VGG-S model [76] pretrained on ILSCRV and TCN jointly followed with a BiLSTM (Bidirectional LSTM) to produce the alignment proposal. Then they utilized the GoogLeNet combined with TCN to learn the features with alignment proposal.

Yang et al. [77] proposed SF-Net which concatenates the features extracted by 2D and 3D ResNet18. They divided the network into three parts to concentrate on the features from frame level, gloss level and sentence level. In gloss level, a LSTM was used to reduce dimension and form compact gloss level features and in sentence level, a BiLSTM was used to encode the context information. This level-by-level feature fusion method is also used in [67], [78], and [79], showing its capability. Reference [67] labeled 400 frames randomly from CSL dataset to fine-tune a Faster R-CNN [80] pre-trained on VOC2007 person-layout dataset for hand detection. They aimed to recognize the continuous sign language and proposed a solution by compressive tracking [81] when the detector fails to get the hand region. A video-sentence latent space was also put forward to match the video clips and words in the sentence. However, using a sliding window to extract clip features of each piece of video can be computational-intensive. And if the latent space is too big in the case of oversized datasets, it may be hard to get the mapping function by networks. Reference [78] put forward a method to enhance the gloss feature in the task of continuous sign language recognition. By introducing GFE module (which is a decoder), they calculated the cross entropy of the outputs of GFE decoder and CTC decoder, aiming to make the outcome of GFE decoder closer to the alignment proposal produced by CTC decoder. After the back propagation, the GFE could be able to represent smooth features. This method is highly dependent on the quality of the CTC results, as it uses the output of CTC as the supervision information. Reference [79] recognized continuous sign language by combining phrase-level features of video frames extracted by ResNet152 as well as BiLSTM and outcomes of a LSTM which was used to capture the sequential label information. Then, a RNN-T [82] model was used to train these concatenation features. The concatenation operation could establish links between labels and video sequences, making the features more representative when feeding into the RNN-T.

The methods based on CNN and RNN achieve good performance even on the small datasets in the field of sign language recognition due to the strong inductive bias of CNN and RNN. They concentrate on the locality of frames and cope with sequence strictly along the chronological order. With time going by, more and more large datasets are beginning to emerge and the application of CNN and RNN may come to a bottleneck, i.e., they may not perform well on large datasets, for sign language not only contains manual features like hand shape, orientation and movement, but also non-manual features like mouthing, eye gaze and eyebrow movements, etc. [59]. It is difficult to capture the relationship between manual features and non-manual features if we only look for their relationship locally. What's more, RNN may suffer vanishing gradients and cannot handle long sequence, and is uncapable to establish links among random frames.

b: METHODS BASED ON TRANSFORMER

The feature extraction based on Transformer and its variants are very popular in recent years. Transformer was originally designed to deal with NLP problems and it shows its strong power in global relationship establishment on the large datasets. Various types of Transformers [83] have been designed by scholars to cope with different tasks in computer vision field and etc. and these structures are also very popular in the field of SLR, especially in the field of sign language translation (SLT). Unlike spoken language, sign language has its unique grammatical structure, which means they have different sequence order. SLR generates the glosses in the same order with sign language while SLT aims to transfer gloss sequence into our spoken language sequence. Current Transformer-based approaches mostly integrate SLR and SLT into one framework, so we give an overall introduction here. The fundamental principle underlying the Transformer architecture lies in the utilization of self-attention to effectively capture overarching dependencies across various positions within the input sequence. This architecture comprises two primary components: the encoder and the decoder. The encoder is composed of multiple layers of encoding, each layer encompassing two principal sub-layers: the multi-head self-attention mechanism and the positionally fully connected feed-forward network. The multi-head self-attention mechanism enables concurrent focus on multiple positions within the sequence, whereas the position feedforward network conducts independent nonlinear transformations on the representation of each position. Additionally, a third sub-layer is introduced within the decoder to attend to the encoder's output.

To fully leverage the computational capabilities for long-range dependencies of transformers and the local feature extraction abilities of CNNs, Shin et al. [84] proposed a multi-branch network based on convolutional and transformer layers to parallelly extract local and long-range dependency features. De Coster et al. [59] trained a Video Transformer Network (VTN) [85] with the features extracted from RGB information by ResNet-34 pretrained on ImageNet [86] and the pose flow obtained by OpenPose BODY-135 model [60]. The result showed the accuracy of 92.92% on AUTSL isolated sign recognition dataset [87]. However, they chose to uniformly down sampling the video frames before feeding them into the VTN. This may be not a good choice as the information of the video may not be evenly distributed. Du et al. [88] utilized a tiny Swin-Transformer [89] to transform RGB images into semantic features and used another Mask-Future Transformer for temporal sign language video modeling and comprehension of sign language actions. This Mask-Future Transformer aimed to calculate the self-attention without using the future frames and they believed it is more suitable for sequence modelling and comprehension. However, Two-Transformer-based structures would make the training harder and more time-consuming. Niu et al. [90] used pretrained ResNet to extract features of each RGB frame and fed them to the Transformer. They avoided overfitting by stochastically dropping some frames during training and not computing back-propagation for a part of the input frames during spatial feature extraction. Furthermore, they introduced sub-gloss states for each gloss, and calculated extension function of gloss sequence and sub-gloss state number sequence by Monte Carlo sample method. They

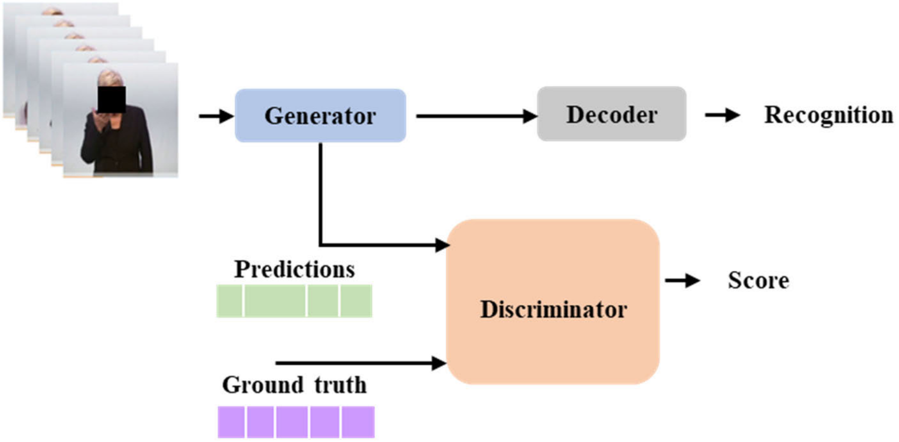

FIGURE 7. The main structure of GAN used in SLR.

then approximated the likelihood term in the sampled results and computed the lower bound of CTC, whose objective value is proved to become larger by iteration in this way. The performance showed WER of 24.9 and 25.1 on RWTH-PHOENIX-Weather 2014 and RWTH-PHOENIX-Weather 2014T respectively.

Yin et al. [91] translated the sign language by two steps. They first generated the glosses by Spatial-Temporal Multi-Cue (STMC) Network [92], which used multiple visual cues as input including face, hand, full-frame and pose and extracted the temporal features of inter-cue and intra-cue. This STMC network tried to extract almost all the possible features from the single RGB information to make the feature more representative by detecting different body parts. It is more effective than just feeding full frames into the network. These features are fed to BiLSTM and calculated the Connectionist Temporal Classification (CTC) [93] loss. Then they did the Neural Machine Translation (NMT) by Transformer network. Camgoz et al. [15] combined the recognition and translation. They utilized the spatial embedding approach [94] to extract non-linear frame level spatial representation and fed them to the self-attention layer after combining with positional encoding. They calculated the loss in gloss level by CTC as intermediate supervision in Transformer encoder, helping match video frames with glosses. In decoder, they input word embedding and attention information from encoder and generalized the final output. The results shown BLUE-4 and WER of 22.38 and 24.98 respectively on RWTH-PHOENIX-Weather 2014T dataset. They expanded channel numbers, adding the mouthing and pose information in Transformer network in the later work [95]. The results showed 19.51 and 45.90 on BLEU-4 and ROUGE metrics respectively. Li et al. [96] exploited a semantic hierarchical structure among video segments by different sliding window sizes. Instead of pooling or concatenating of multi-scale segments directly, they developed inter-scale and intra-scale attention approaches to calculate the relation locally and globally. They finally decoded the features by Transformer decoder.

Guo et al. [97] used the adaptive temporal interaction (ATI) module to incorporate the adaptive shift operation and self-attention to capture local and non-local temporal correlations concurrently. For the transformer to be able to distinguish between temporal and spatial features of sign language videos, Cui et al. [98] designed an ST dual-channel feature extraction network to extract contextual features and dynamic features, respectively. Zuo et al. [99] enhanced the transformer backbone from a consistency perspective by adding spatial attention consistency constraints and sentence embedding consistency constraints. Hinrichs et al. [100] extracted and augmented body markers using data imputation and velocity-like features, which were then used with a transformer network for continuous sign language recognition, and achieved state-of-the-art performance.

Although good performance has been achieved by Transformer-based sign language recognition and translation network, it can't be denied that the drawback of Transformer still exists. Transformer only takes the previous content of the output text into account, and it is incapable to incorporate context from both directions, which may lead to the omittance of crucial information between gestures. With the emergence and success of Bidirectional Encoder Representations from Transformers (BERT) [101], many scholars adopted it from NLP field and designed BERT-based sign language recognition network [19], [102], [103], [104], [105].

Hu et al. [102] designed a SignBERT model for isolated SLR. They utilized 2D hand pose sequence of both hands as data and train the SignBERT in an self-supervised manner by masking and reconstructing visual tokens. Instead of inputting the skeletal coordinates to the SignBERT directly, they adopted the spectral-based Graph Convolution Network (GCN) [106] to process the sequence frame-by-frame to generate the frame-level semantics representation. After adding position encoding and hand chirality embedding (to distinguish left and right hand), they trained the features with SignBERT decoded the results by MANO [107], which created a mapping from low-dimensional pose to triangulated hand mesh. Then they applied the SignBERT to downstream recognition task. Zhou et al. [103] applied (3+2+1)D ResNet () [72] to extract selected key frames features and fed it into BERT to calculate the temporal relationship among each frames. In addition, they utilized open-pose [108] to locate the region of the dominant hand in the selected key frames and processed them by (3+2+1)D ResNet and BiLSTM to extracted the spatial-temporal features. After combining the output of BERT, spatial features of key frames and the spatial-temporal features of dominant hand, they input them into a BiLSTM and optimized the CTC loss. They also designed a module, minimizing the feature distance between hand image and BERT output representing the same gloss at each time step by Jensen-Shannon divergence (JSD) loss. The WER metric reached 1.52 and 23.30 on signer independent (i.e., the signer in the training and test sets are not identical) and unseen sentence (i.e., the sentences in the testing set have never occurred but each of their words has appeared in other sentences in the training set) conditions respectively on CSL dataset, and 20.1 on RWTH-PHOENIX-Weather 2014 dataset. They then refined the structure by implementing multiple BERT models for different input modalities and designed a cross-attention mechanism to exchange inter-modality information between BERT models [104]. The WER reached 1.14 and 19.80 on signer independent test and unseen sentence test respectively on CSL dataset, and 18.3 on RWTH-PHOENIX-Weather 2014 dataset.

The Transformer-based network achieved good performance on isolated and continuous sign language recognition even on SLT tasks. The reason is that multi-head attention mechanism, which is the core of Transformer, calculated the attention among tokens globally. This trait is hard to be achieved by using CNN, so Transformer performs better than CNN theoretically especially on large datasets at the cost of larger amount of parameters. Meanwhile, because of the weak inductive bias, the Transformer needs to be trained over a longer period of time to converge. To speed up, pretraining is used [19], [102], [103], [104] to equip the network with a prior knowledge. We could also see the implementation of Transformer is in fact to deal with 1D temporal problem in the SLR field, which means CNN is involved to compress the data into vectors before feeding the data into Transformer (except for [88], which utilized two Transformer models to extract spatial and temporal features separately). So precisely, the vast majority of current Transformer-based approaches are actually a combination of the transformer and CNN approaches. From recognition with CNN and RNN to recognition with CNN and Transformer, we found that the performance of sign language recognition has been greatly improved, and the recognition criteria have also evolved from validation within the dataset to signer independent recognition and unseen sentence recognition, being much closer to real-world application scenarios.

c: OTHER METHODS

In addition to the methods mentioned above, Generative Adversarial Network (GAN) [109], [110] and Graph Neural Network (GNN) [111], [112], [113], [114], [115] are also used in the domain of SLR. The main structure of GAN used in SLR is shown as Fig. 7.

Generative Adversarial Network (GAN) consists of two parts, which are generator and discriminator. Generator aims to produce the predicted results and discriminator aims to distinguish generated results and real labels. These two parts are trained iteratively. When discriminator is unable to tell the difference between generated results and real labels, we assume that the generator has the ability to generate real data. Papastratis et al. [110] designed the generator by feeding the frames into 2D-CNN and establishing the short-term and long-term temporal dependencies within the sequence by 1D convolution and BiLSTM separately. The output sequence and the ground truth label are fed into the discriminator, in which a two-stream network is designed to process the sequence on gloss-level and sentence-level respectively. Finally, a fully connected layer is used to generate the judging scores. In decoding phase, a Transformer decoder is used to generate the translation. Elakkiya et al. [109] extracted manual and non-manual features by BPaHMM [116], and denoised and reduced dimension by variational autoencoder (VAE). The LSTM and 3D-CNN are employed as generator and discriminator. A deep reinforcement learning method is implemented to optimize the hyperparameter and regularization.

Although GAN can be used to some effect in sign language recognition, its training time is too long and it requires a large dataset to support it. What's more, GAN is still based on CNN and RNN, which means it has the same shortcomings of CNN and RNN.

Graph Convolutional Network (GCN) [117] aims to conduct convolution operation in graph. A graph can be defined as $G = (V, E)$, indicating the nodes and edges of the graph respectively. The adjacency matrix A reflects the connectivity of the nodes. The layer-wise propagation function of GCN can be written as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{4}$$

where $\tilde{A} = A + I$ represents the adjacency matrix with self-connections of the undirected graph, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ and $W^{(l)}$ is a trainable weight matrix. The function is activated by σ, which can be ReLU etc. GCN could help us find the relationship among nodes in graph, which is a more widely used form of data. Since the skeletal key points are the only data resources which can be represented as graph in SLR, the GCN-based approaches mostly take these information as input. Skeleton data focuses on joint positions rather than the entire image. This approach removes non-critical information such as background and clothing, reducing data redundancy and enhancing robustness to lighting and scene variations. References [111], [112], and [118] all utilized spatial-temporal GCN, for skeletal key points information are also in sequence manner. This spatial-temporal GCN is similar to 3D CNN, which takes both spatial and temporal adjacent nodes into consideration. Wang et al. [112] defined a layer-wise propagation rule of GCN spatially by partition and aggregation and then expanded the size of convolution kernels temporally. Jiang et al. [118] utilized the spatial decoupling graph convolution [119] method, which adopts decoupling aggregation in GCN to enhance the performance, and a STC (spatial, temporal and channel-wise) attention module to construct the basic block. Then a multi-stream approach is exploited to deal with joint, bone, joint motion and bone motion separately, where bone and motion nodes can be calculated by subtracting the joint coordinates spatially and temporally respectively. The results of different streams are concatenated in the final. They reduced the whole-body skeleton graph nodes number from 133 to 27, as the higher the number of nodes, the more noise is introduced. Miah et al. [120] firstly constructed a skeleton graph based on 27 whole-body key points. Subsequently, they conveyed information from four streams into their proposed multi-stream SL-GDN to extract features. Finally, they concatenated the four different features and performed classification. In their subsequent research, a multi-stream network approach was similarly employed. However, distinguishingly, they incorporated two graph-based neural network channels, leveraging attention mechanisms to generate temporal-spatial features and spatial-temporal features respectively. These features were then fused with the generic features extracted by the final branch's universal deep neural network module [121]. Shin et al. [122] constructed a dual-stream network, which generates graph-based features by utilizing channel attention modules and a complete skeletal joint skeleton in the first stream. The second stream focuses on capturing joint motion information, followed by feature fusion for classification. Case in the graph structure of GNN, the two hand nodes and the other body joints are equally treated, to fully explore the correlation between two hands, Guo et al. [123] designed a hand shift operation to capture detailed associations between hands. In addition, a cascaded attention module was introduced in the feature extractor, which establishes residual connections to the input of spatial, temporal, and channel attention, making the model more focused on useful information in sign language actions. Instead, Papadimitriou et al. [124] used a modulated GCN to model various correlations between different body joints beyond the physical structure of the human skeleton. Meng, L., et al [125] proposed a multi-scale attention network (MSA) to model the dependencies between remote vertices to learn the long-distance dependencies, and an attention enhanced temporal convolutional network was proposed to automatically assign different weights to different frames in order to solve the problem of inaccurate recognition of motion blur frame joints. Kan et al. [111] proposed a hierarchical spatial-temporal graph structure, which consists of models on high-level and fine-level. In high-level graph, three vertices which denote facial region, left-hand region and right-hand region are presented which indicate the relative motion between these three main components. In fine-level graph, skeletal key points of a specific region are used to analysis the detailed information.

Because sign language is performed by multiple parts of the body, the nodes of the graph structure used in sign language recognition should respond to information from those parts. However, too many nodes do not provide additional useful information to the model, but instead introduce noise into the model, which affects the accuracy of the model [118], [126]. Therefore, for skeleton-based SLR, it is important to choose the right nodes for model learning.

d: METHODS BASED ON MULTI-MODALS

We introduce the feature extraction from another perspective in this part, focusing on the types of input data. We introduced different types of data and reviewed classic multi-modal based methods. The forms of input data can be categorized into RGB information, depth information and skeletal key points information. And the data required for the body parts also varies as sign language is not just presented by hands, but also facial expression, mouthing, the movement of bodies etc., which can be divided into manual features including hand shape, orientation and movement etc. and non-manual features including mouthing, eye gaze and eyebrow movements etc.

The forms of input data can be categorized into RGB information, depth information and skeletal key points information, of which RGB information is most commonly used, for RGB information is closer to the human eye's visual perception and easy to obtain by portable devices like mobile phones and cameras. Depth information usually serves as additional information for RGB information, which is seldom used independently. Skeletal key points are in fact a more refined information which get rid of most of the redundancy, so it often performs better on the recognition tasks. But meanwhile, the precise position of key points is not easy to obtain. References [15], [17], [57], [59], [61], [70], [71], [73], [75], [77], [88], [90], and [91] all chose RGB information as input, showing the popularity of this data form. References [65] and [103] combined the RGB information and skeletal key points information, but it is worth noting that their skeletal key points are all obtained from RGB information by algorithms. In fact, seldom scholars would acquire position of skeletal points by markers or sensors currently due to the high precision of key points detection algorithms like Openpose etc. Jiang et al. [118] combined RGB and depth information, generating pose, optical flow, HHA [127], and depth flow from original resources. It can't be denied that rich kinds of data would lead to a better performance, but it would also cause a drain on time and computing power.

The body parts information can be categorized into manual component, which refers to the features of hand shape, orientation, movement and etc., and non-manual component, which includes the information about mouthing, eye gaze, expression and etc. Non-manual component usually serves as additional information for manual component and it could provide richer information for recognition. Nevertheless, it is crucial to conduct alignment among different feature streams, otherwise it may cause confusion to the network.

Koller et al., [128] designed multi-stream CNN-HMM to process sign gloss, mouth shape and hand shape features separately and synchronize at sign ends. Gündüz et al. [129] cropped face regions and hand regions from original video frames and processed them in different streams. Liang et al. [130] developed an automatic toolkit for British sign language recognition, which took hand-arms movements and facial expressions into consideration. Zheng et al. [131] proposed a face highlight module to extract facial expression information and fused it with non-facial features. To further describe hand movements, optical flow could also be used to improve the performance network [71], [118], [132]. Slimane et al. [133] used CNN to extract features from full frame and employed self-attention for temporal modeling. In addition, a secondary stream for the cropped handshape sequences was added, and an attention mechanism was used to focus only on the required context and discard unnecessary distant information, efficiently aggregating the hand features with their appropriate spatio-temporal context. Sarhan et al. [134] used a three-stream model to process RGB, optical flow and TD attention data, respectively. For TD attention, they generate pixel-precise attention maps that focus on both hands, thereby preserving valuable hand information while eliminating distracting background information.

Nowadays, more scholars choose to input with full frames of sign language video, for they contain complete information including hand shape, movements, facial expression etc. This requires an elaborately designed network structure that can extracts useful features and discards redundant information.

3) RECOGNITION

The recognition steps for isolated and continuous SLR are different. For isolated SLR, the recognition is done by classification. This can be achieved easily by Softmax function [71], [72] or SVM [65]. Nevertheless, for continuous sign language, the recognition result is a sequence, which is essentially a seq2seq problem. Connectionist temporal classification (CTC) and Encoder-Decoder Network are two major approaches prevailed in seq2seq problem and they are then adopted from NLP field to continuous SLR by scholars. We will introduce the application of these two methods in continuous sign language recognition.

a: CONNECTIONIST TEMPORAL CLASSIFICATION

Connectionist temporal classification was first put forward by Grave et al. [93] in 2006. It aims to resolve seq2seq problem without fine-grained labels. In addition to the units, which corresponds to the words in dictionary L (the set of all words that appear in the dataset), in Softmax output layer of CTC, an extra unit representing 'blank' is added to the Softmax layer. This extra unit aims to establish all possible many-to-one mapping from output sequence to label sequence as they may not be equal in length. The dictionary turns to $L' = L \cup \{blank\}$ and the emergence possibility of a certain sentence π can be calculated as follows:

$$P(\pi|x) = \prod\nolimits_{t=1}^{T} y^t_{\pi_t} \forall \pi \in L'^T \tag{5}$$

where x is an input sequence of length T, L'^T is the set of length T sequences and y^t_k is interpreted as the probability of observing label k at time t. Then the many-to-one mapping $\mathcal{B}$ is defined by removing all blanks and repeated labels to form the possible output sequence (e.g., $\mathcal{B}(a-ab-) = \mathcal{B}(-a-abb) = aab$ Where '-' represents blank) and the conditional probability of a given labelling l can be calculating the sum of probabilities of all possible output sequence as follows:

$$P(l|x) = \sum\nolimits_{\pi \in B^{-1}(l)} P(\pi|x) \tag{6}$$

The optimization objective is to maximize the conditional probability of the given labelling l and the result can be obtained by forward-backward algorithm based on dynamic programming algorithm.

CTC is commonly used in weakly supervised learning and it was a common method for continuous sign language recognition until transformer became popular for visual tasks. References [61], [70], [77], [78], [90], [92], and [135] all used CTC as the objective function. Although the CTC-based approach can produce better results, its drawbacks cannot be neglected. The equation (6) holds when each frame in target sequence is conditionally independent, which is exactly what CTC assumes. Nevertheless, this is unreasonable, for each frame within the local scope of the sign language video is intrinsically linked. To alleviate this issue, some other methods [82], [136] could be tried in the future study. Besides, in the process of solving dynamic programming, the input video sequence must be longer than the output label sequence in length. These led to the development of Encoder-Decoder Network architectures.

b: ENCODER-DECODER NETWORK

The Encoder-Decoder Network consists of two parts to match two sequences in an intermediary latent space. Encoder aims to encode the input sequence into a fixed size vector, and decoder aims to complete the alignment between input sequence and target sequence in the latent space and output the predicting results. This structure could deal with complicated seq2seq problem, so not only Encoder-Decoder Network could resolve SLR but even SLT, especially for Transformer (which is actually an Encoder-Decoder Network essentially) because it can calculate the attention globally instead of following the sequence order. Here again we

TABLE 1. Methods based on CNN and RNN.

References	Input data	Feature types	Method
[17]	RGB	Hand + full frame	CNN + LSTM + HMM
[57]	RGB	Hand	CNN + HMM
[70]	RGB	Full frame	3D-CNN +Encoder-Decoder
[75]	RGB	Hand	CNN + TCN + BiLSTM+CTC
[64]	RGB	Full frame	CNN + BiLSTM
[77]	RGB	Full frame	2D CNN + 3D CNN + LSTM + CTC
[67]	RGB	Hand + full frame	C3D + encoder-decoder
[78]	RGB	Full frame	1D CNN +Encoder-decoder + CTC
[79]	RGB	Full frame	CNN + BiLSTM +RNN-T
[137]	RGB	Full frame	C3D+Encoder-decoder
[94]	RGB	Full frame	CNN+Encoder-decoder
[96]	RGB	Full frame	3D CNN + encoder-decoder
[92]	RGB	Face + hand + full frame + pose	CNN + TCN + BiLSTM + CTC
[131]	RGB	Face + full frame	CNN + encoder-decoder
[132]	RGB + optical flow	Hand + full frame	CNN + BiLSTM + CTC
[135]	RGB	Full frame	(2+1)D CNN + BiLSTM + CTC
[65]	RGB + key points	Hand	3D CNN + SVM
[72]	RGB	Full frame	(3+2+1)D CNN
[69]	RGB	Full frame	CNN + BiLSTM

Pros: Strong inductive bias, focus on frame locality, strict chronological processing of sequences

Cons: Finding features only locally, difficult to capture relationships between them. RNN may suffer vanishing gradients and cannot handle long sequence, and is uncapable to establish links among random frames.

TABLE 2. Methods based on transformer.

References	Input data	Feature types	Method
[90]	RGB	Full frame	CNN + Transformer + CTC
[91]	RGB	Face + hand + full frame + pose	Transformer + CTC + CNN
[15]	RGB	Full frame	CNN + Transformer
[95]	RGB + key points	Hand + mouthing + pose	CNN + Transformer
[97]	RGB	Full frame	CNN + LSTM + HMM+ Transformer
[98]	RGB	Full frame	Transformer
[99]	RGB	Full frame	Transformer
[100]	Key points	Whole body	Transformer
[59]	RGB + Key points	Hand	Transformer
[88]	RGB	Full frame	Transformer
[103]	RGB	Full frame + hand	(3+2+1)D ResNet + BiLSTM + BERT + CTC
[104]	RGB + Depth + Key points	Full frame + hand	(3+2+1)D ResNet + BiLSTM + BERT + CTC
[102]	RGB + Key points	Hand	BERT

Pros: The attention among tokens globally can be calculated due to its multi-attention mechanism.

Cons: More parameters are needed, and need to be trained over a longer time to converge.

emphasis that sign language has its unique grammatical structure as oppose to spoken language, which may lead to the different orders of words in a language sequence. SLR is to output the language sequence with same words order in inputting video sequence while SLT transform the order of words in output language sequence into our spoken language form. In fact, some SLT based on encoder-decoder network incorporates the SLR process, so a uniform presentation is made here. In the early studies, the encoder-decoder network of SLR and SLT are all based on RNN [70], [94], [137]. Camgoz et al. [94] embedded the frames in video sequence by 2D CNN and words in label sequence by linear projection and inputted them into the encoder-decoder network based on RNN.

TABLE 3. Other methods.

References	Input data	Feature types	Method
[110]	RGB	Full frame	GAN + Transformer
[109]	RGB	Hand (position, velocity) + non-manual features	HMM + VAE + GAN
[112]	Key points	Whole body	GCN + encoder-decoder
[111]	RGB	Head + left hand + right hand	GCN + Transformer
[118]	RGB + Depth + optical flow + key points	Full frame	3D CNN + GCN + TCN

GAN: Pros: can easily learn complicated distributions of data. Cons: It is still based on CNN and RNN, so has the same shortcomings of them. GCN: Pros: Effectively preventing background interference. Cons: Occlusion, viewing angle, and lighting can cause missing data at skeleton joint points, Excessive joint points introduce noise

TABLE 4. The comparison of isolated SLR methods.

References	Dataset	Accuracy
[59]	AUTSL	92.92%
[118]	AUTSL	98.53%
[72]	CSL, HKSL	96.0%, 94.6%
[102]	CSL, NMFs-CSL, MSASL, WLASL	97.6%, 78.4%, 71.24%, 54.69%
[73]	CSL	98.4%
[138]	CSL	98.54%
[88]	NMFs-CSL, WLASL	72.4%, 57.13%
[139]	MSASL, WLASL, NMFs-CSL	73.80,61.26, 83.7

Guo et al. [137] constructed the hierarchical encoder-decodernetwork, which contains two layers of LSTM and an additional LSTM was used to select key clips from input videos. Puet al. [70] combined encoder-decoder network with CTC, designing a BiLSTM encoder, a LSTM decoder and a CTC decoder. Although encoder-decoder network improves the performance of SLR and SLT, it may suffer long-term dependencies between source and target sequence due to the characteristic of RNN, so attention mechanism [140], [141] is used to provide extra information to the decoder to alleviate the long-term dependencies. On the other hand, Transformer is in fact an Encoder-Decoder Network essentially, which is most popular in SLR and SLT fields. The multi-head attention, which is the core of Transformer-based network, can establish links among all embedding features globally. References [15], [85], [88], [90], [91], [96], [102], [103], and [104] all designed their network based on Transformer, which obtained good performance.Although the encoder-decoder network is popular among SLR and SLT, it requires large computing power and it is hard to converge, so some tricks are used to assist the training, like pre-training, key pooling [137] etc. In addition, the latent space should be well designed to make the decoding successfully. What's more, most encoder-decoder networks are unable to capture bi-directional information when decoding, which may omit some import information among frames. Although BERT aims to resolve this problem, it requires great computational cost which is unaffordable for many scholars. For portable utilization, trade-off should be made.

C. CONCLUSION

We first summarise each of the previously mentioned methods in Tables 1 to 3. Each table contains the form of the input data, the feature types, and the methods. At the end of the table, we summarise the advantages and disadvantages of such methods. The comparison of isolated SLR methods is shown in Table 4. The comparison of continuous SLR methods is shown in Table 5.

From Table 4 it can be found that in relatively small datasets such as CSL, the performance can be perfectly good. But when it comes to larger dataset such as WLASL, the accuracy only reached up to 57.13%. We could also find that the input with full frame tends to obtain better performance. This is due to the rich information included in the full frame like facial expression, mouthing etc. All methods mentioned above are not signer-independent, which means the same signer may appear in both training set and testing set. This may cause confusion in practical applications if the network learns more information about signer rather than sign language. De Coster et al. [59] mention in their paper that because robust models pick up individual idiosyncrasies, validation, and test results will be overly optimistic due to data leakage if the same person appears in the training, validation,

TABLE 5. **The comparison of continuous SLR methods.**

References	Dataset	Metrics	Results
[17]	RWTH 2014	WER	26.8
	SIGNUM		4.8
[57]	RWTH 2012	WER	30
	RWTH 2014		31.6
	SIGNUM		7.4
[61]	RWTH 2014	WER	43.1
[70]	RWTH 2014	WER	37.1
[75]	RWTH 2014	WER	39.4
[77]	RWTH 2014	WER	34.9
	CSL		3.8
[67]	RWTH 2014	Accuracy	61.7%
	CSL		82.7%
[78]	RWTH 2014	WER	23.7
	RWTH 2014T		23.3
	CSL		3.0
[79]	CSL	WER	6.1 (SI)
[137]	CSL	Precision	92.9% (SI)
		BLEU-4	92.8 (SI)
		ROUGE-L	95.1 (SI)
[90]	RWTH 2014	WER	24.9
	RWTH 2014T		25.1
[91]	RWTH 2014T	BLEU-4	24.38/49.01
	ASLG-PC12	ROUGE-L	81.93/96.18
[15]	RWTH 2014T	BLEU-4	22.38
		WER	24.98
[94]	RWTH 2014T	BLEU-4	19.26
[92]	RWTH 2014	WER	21.1
	RWTH 2014T		19.6
	CSL		2.1 (SI)
[95]	RWTH 2014T	BLEU-4	19.51
		ROUGE	45.90
[96]	RWTH 2014T	BLEU-4	13.41
		ROUGE-L	34.96
[103]	RWTH 2014	WER	20.1
	CSL		1.52 (SI)
[104]	RWTH 2014	WER	18.3
	CSL		1.14 (SI)
	GSL		2.18 (SI)
	HKSL		2.63 (SI)
[110]	RWTH 2014	WER	23.4
	CSL		2.1
	GSL		2.26
[109]	RWTH 2014T	WER	20.7/-
	ASLLVD	CER	-/1.4
[112]	LMSLR	WER	2.07
	CSL		1.3
[111]	RWTH 2014	WER	19.5/-
	RWTH 2014T	BLEU-4	-/22.6
	CSL		27.6/17.8
[131]	RWTH 2014T	BLEU-4	10.94
		ROUGE	34.96
[135]	RWTH 2014	WER	20.8
	RWTH 2014T		20.8
[69]	RWTH 2014	WER	21.0
	RWTH 2014T		20.7
	CSL		0.8
	CSL-Daily		30.7
[68]	RWTH 2014	WER	19.4
	RWTH 2014T		20.5
	CSL		0.8
	CSL-Daily		30.1

RWTH 2012 is the abbreviation of RWTH-PHOENIX-Weather 2012 dataset. RWTH 2014 is the abbreviation of RWTH-PHOENIX-Weather 2014 dataset and RWTH 2014T is the abbreviation of RWTH-PHOENIX-Weather 2014T dataset. SI refers to the signer independent situation. We would like to add that the higher BLEU-4 and ROUGE are, the better result is. The lower WER is, the better result is.

and test sets. In contrast, signer-independent SLR necessitates distinct signers in the training and test sets, compelling the network to focus on sign language information. For example, the pose-based transformer proposed by Alyami et al. [142] achieves 99.74% and 68.2% accuracy in signer-dependent and signer-independent modes, respectively, on the KArSL-100 dataset. From Table 5 we could find that the majority of encoder-decoder (Transformer) based methods reduce WER to below 25 or even below 20, showing the capability of this structure. More than half scholars choose to use more than one feature types especially the combination of hand

and full frame as this combination can not only concentrate on the hand information locally, but also other information and relation among different body parts globally. Although different form of input data may provide additional information, RGB is still the most commonly selected data for it is easy to acquire by portable devices, which are also the main platforms to conduct SLR in the future. We could also see that most experiment did not test their network on signer independent and unseen sentence situation except for [70], [79], [92], [103], [137], which means the majority approaches still have limitation on practical applications. This is still a bottleneck in the field of continuous SLR.

IV. DATASETS

Sign language recognition datasets are one of the most important parts in SLR field and they support the validation of the effect of the algorithm. For sign language recognition, the early sources of data are often data collected by sensors, which include mainly the position of joint points, hand movement trajectories, etc. Due to the complexity of the acquisition device, the size of the data set is often small and unique, which are not persuasive and comparable. With the development of machine learning and deep learning, the form of data is becoming simpler and easier to obtain, all thanks to the powerful data processing capabilities of deep learning. Currently, common forms of data include RGB video, RGB-D video, skeleton key points, etc. Here we briefly introduce some of the classic datasets as well as the latest ones.

The sign language can be divided into two parts, i.e., fingerspelling and sign. The fingerspelling is a simple form of sign language, which represents the alphabet by different gestures. Strictly, fingerspelling recognition can be categorized into gesture recognition. ASL alphabet dataset [143] is an American fingerspelling dataset in which the samples are all static and presented by single hand. The dataset is split into two parts according to their difficulty of recognition. However, most fingerspelling datasets are recorded in a more homogeneous environment. In order to achieve the dataset in the wild, Shi et al. proposed two American sign language fingerspelling datasets, i.e., ChicagoFSWild [144] and ChicagoFSWild+ [145], In which ChicagoFSWild+ includes 50,402 training sequences by 216 signers, 3115 development sequences by 22 signers and 1715 test sequences by 22 signers, with no overlap in signers in the three sets.and contains 10.2% of left-handed situations and 2.6% of other situations. In addition to these large datasets, researchers in different countries also produce small datasets of local fingerspelling, however, the majority of these datasets are small in scale. Here we do not give a further introduction. Although fingerspelling is a form of sign language representation, it needs to be represented letter by letter, which is cumbersome and time-consuming. So far, fingerspelling acts as an aid to sign language expression.

The most commonly used continuous sign language datasets are RWTH-PHOENIX-Weather 2014 [35] and RWTH-PHOENIX-Weather 2014T [94], containing German Sign Language (GSL) collected from weather forecast programmes. RWTH-PHOENIX-Weather 2014 contains 45,760 video samples at 210 × 260 resolution from 9 different signers. In addition, the dataset provides videos of the signer's right hand movement trajectory. RWTH-PHOENIX-Weather 2014T is an extension of the RWTH-PHOENIX-Weather 2014 corpus and aims for SLT task. It is presented by 9 different signers with a vocabulary of 1066 different sign glosses in sign language and 2887 different words in German spoken language. The total number of video samples are 8257. SIGNUM [146] was created by Ulrich von Agris et al. from RWTH Aachen University. It is also a German Sign Language dataset, including 450 isolated glosses and 780 continuous sentences from 25 different signers. What's more, Koller et al. introduced 1miohands datasets [16], which is a dataset on common gestures in Danish Sign Language, New Zealand Sign Language and German Sign Language, containing over one million hand shapes images presented by 23 persons. It can be used as a pretraining dataset for network to acquire prior gesture information.

Apart from German Sign Language, American Sign Language (ALS) datasets take a large proportion of all datasets. American Sign Language Lexicon Video Dataset (ASLLVD) [147] was created by scholars in Boston University. It was collected from multi-angle by four cameras. Linguistic annotations include gloss labels, morphological and articulatory classifications of sign type. ASL-LEX [148] is an American Sign Language dataset covering 993 types of glosses, together with detailed lexical and phonological information about these glosses like frequency, iconicity, phonological composition, and neighborhood density. This dataset was updated to ASL-LEX 2.0 [149] later and the vocabulary was expanded to 2,723, which includes more detailed information. How2Sign [150] is a multimodal and Multi-view continuous American Sign Language dataset, including RGB, depth and 2D key points information from frontal view and side view. The total length of videos can be up to 80 hours. Word-Level American Sign Language (WLASL) [151] dataset features more than 2,000 common different words in American Sign Language by 119 signers. It consists 21,083 RGB videos without other types of information. By far it is the largest isolated American Sign Language dataset, which contains different backgrounds, illumination conditions. MS-ASL [152] is another isolated sign language dataset containing 1,000 types of signs which were recorded in challenging and unconstrained real-life conditions by over 200 signers. The types of signs in 27 Class ASL Sign Language [153] are less, but they were performed by 173 individuals and the resolution of RGB image is up to 3024 × 3024. The American Sign Language dataset is well diversified and informative, and has been one of the main choices for sign language recognition in recent years.

Chinese Sign Language dataset is also an important part in SLR field. Chinese Sign Language Recognition Dataset (CSL) is a comparatively large dataset containing isolated

TABLE 6. Fingerspelling datasets.

Dataset	Language	Data type	Dataset Scale	Recording Backgrounds	Characteristics
ASL alphabet dataset A [143]	American	RGB + depth	24 static sign by 5 persons, with up to 131,000 samples	In similar lighting and background	-
ASL alphabet dataset B [143]	American	Depth	24 static sign by 9 persons, with up to 72,676 samples	In two different environments	-
ChicagoFSWild [144]	American	RGB	7,304 fingerspelling sequences by 160 signers	In different kinds of background	Signer independent
ChicagoFSWild+ [145]	American	RGB	55,232 fingerspelling sequences by 260 signers	In different kinds of background	10.2% left-handed 2.6% other cases Signer independent

and continuous sign language. The isolated part [154] includes 500 different types of glosses while the continuous part [67] contains 100 sentences. The multi-modal information like RGB, depth and skeleton key points are provided. We also have a heard of DEVISIGN dataset but unfortunately, we haven't found any access to this dataset and none of the articles in recent years we have read utilized this dataset for validation. Currently Chinese sign language datasets are small in number and size compared to ASL and GSL.

There are also many other sign language datasets in different countries and regions. BBC-Oxford British Sign Language Dataset (BOBSL) [155] is a large-scale sign language dataset, comprising 1,940 episodes of British-Sign-Language-interpreted BBC broadcast footage. It covers a wide range of topics from horror, period and medical dramas, history to sitcoms, nature and science documentaries etc. presented by 37 signers. LSFB Dataset [156] is a comparatively large dataset for French Belgian Sign Language presented by 100 signers. It includes isolated sign language, which covers 47,551 samples for 395 glosses, and continuous sign language which is made of over 25 hours of video clips associated with a time-aligned annotation. LSA 64 [157] is a database for Argentinian Sign Language, including 64 different types of signs presented by 10 subjects. It was collected both indoor and outdoor and each subject wore black clothes and fluorescent-colored gloves. Greek Sign Language (GSL) dataset [158] is a large-scale dataset for SLR and SLT. It comprises RGB and depth information collected from 7 different signers and there are 310 unique glosses and 331 unique sentences, up to 10,290 sentence instances, 40,785 gloss instances.

It also was divided into two parts for signer dependent recognition and signer independent recognition. LSE-Sign [159] is a lexical database for Spanish Sign Language, containing 2400 individual signs presented by 2 signers. Ankara University Turkish Sign Language Dataset (AUTSL) [87] is a large-scale, multi-modal datasets on isolated sign language. It includes the RGB, depth and skeleton key points information of 226 different types of glosses performed by 43 different signers and the total number of videos reach up to 38,336. KArSL [160] is an Arabic Sign Language dataset covers 502 types of sign words presented by three professional signers and there are 75,300 samples in the whole database. It was recorded in fixed background in green and signers were not restricted to wear specific clothes. It also includes three modalities i.e., RGB, depth and skeleton key points. The aforementioned sign language datasets are categorized based on their sign language types, namely fingerspelling, isolated, continuous, and a combination of continuous and isolated signs. Detailed information about these datasets is presented in Tables 6 through 9, including data types, scale, recording backgrounds, and dataset characteristics. In recent years, scholars in different countries and regions introduced their local sign language datasets and most of them are large in scale and good in quality. Different types of sign language datasets offer more choices for researchers to validate their algorithms or build cross-language recognition models. This is a very important factor for the improvement of sign language recognition. It is evident that an increasing number of researchers are focusing on creating datasets more suitable for real-world application scenarios. This entails considerations such as recording in complex backgrounds and specific usage scenarios as considered by GSL. Additionally, factors like the number of signers, signer independence, professional level of signers, and dominant hand usage are taken into account. Consequently, the collection of datasets tailored to real-world application scenarios should comprehensively consider both recording environments and demonstrators as key factors. In terms of annotation, a growing number of continuous sign

TABLE 7. Isolated sign language datasets.

Dataset	Language	Data type	Dataset Scale	Recording Backgrounds	Characteristics
ASLLVD [147]	American	RGB	9800 samples by 6 signers on over 3,300 types of signs	By four cameras from different views	has linguistic annotations
ASL-LEX [148]	American	RGB	993 types of signs	In similar lighting and background	offers detailed lexical and phonological information
ASL-LEX 2.0 [149]	American	RGB	2723 types of signs	In similar lighting and background	More phonological descriptions of signs have been added
WLASL [151]	American	RGB	21,083 samples by 119 signers on 2,000 types of signs	In different kinds of background	-
MS-ASL [152]	American	RGB	Over 25,000 samples by over 200 signers on 1,000 types of signs	Recorded in challenging and unconstrained real-life conditions	Signer independent
27 Class ASL Sign Language [153]	American	RGB	Performed by 173 individuals. 130 photos from each person	In different kinds of background	-
1miohands [16]	Danish, New Zealand , German	RGB	over one million hand shapes images from 23 persons	In different kinds of background	Three types of sign language are included
LSA 64 [157]	Argentinian	RGB	3,200 videos from 10 signers with 64 different types of glosses	Collected indoor and outdoor. Signers are in black and wear fluorescent-colored gloves	-
LSE-Sign [159]	Spanish	RGB	2400 individual signs and a further 2,700 related non-signs	With controlled lighting conditions and a chroma background from two different angles	Each sign has a wide range of grammatical, phonological and articulatory information
AUTSL [87]	Turkish	RGB + depth + key points	226 types of signs performed by 43 signers with 38,336 video samples	Presented in real-life background	2 signers are left-handed
KArSL [160]	Arabic	RGB + depth + key points	75,300 samples by 3 signers on 502 types of signs	Recorded in fixed background in green	Performed by 3 professional signers

TABLE 8. Continuous sign language datasets.

Dataset	Language	Data type	Dataset Scale	Recording Backgrounds	Characteristics
How2Sign [150]	American	RGB + depth + key points	Total length of videos is up to 80 hours by 11signers	Recorded from multi-views in green screen studio and panoptic studio	Of the 11 signers, 5 self-identified as hearing, 4 as Deaf and 2 as hard-of-hearing with hand-shape and orientation annotations.
RWTH-PHOENIX-Weather 2014 [35]	German	RGB	1081 sign vocabulary, 7k sentences from 9 signers	Recorded in real-life TV programmes	
RWTH-PHOENIX-Weather 2014T [94]	German	RGB	1066 different sign glosses and 2887 different words from 9 signers	Recorded in real-life TV programmes for SLT	-
CSL-Daily[161]	Chinese	RGB	20,654 videos from 10 signers	In a single background	The topic revolves around people's daily lives

TABLE 9. Dataset containing isolated and continuous sign language.

Dataset	Language	Data type	Dataset Scale	Recording Backgrounds	Characteristics
SIGNUM [146]	German	RGB	450 isolated glosses and 780 continuous sentences from 25 different signers	The signers wear dark clothes with long sleeves in front of blue background	Including manual and facial features
CSL [67, 154]	Chinese	RGB + depth + key points	500 different types of glosses and 100 sentences by 50 signers	In a single background	The most popular Chinese sign language dataset
BOBSL [155]	British	RGB	1,962 episodes (approximately 1,467 hours)	In different kinds of background	Signer independent
LSFB [156]	French Belgian	RGB	Iso: 47,551 samples by 85 signers on 395 types of signs Con: 85,132 samples by 100 signers on 6,883 types of signs	The background of signers is black	Has a time-aligned annotation for continuous sign language
GSL [158]	Greek	RGB + depth	310 unique glosses and 331 unique sentences. Total number of instances is 40,785 and 10,295	Multiple individual and commonly met scenarios	Divided into two parts for signer-dependent recognition and signer independent recognition

language datasets, such as RWTH-PHOENIX-Weather 2014, provide not only sentence-level annotations but also word-level annotations.

This dual annotation approach facilitates researchers to explore the data from both sentence and word segmentation perspectives. In contrast to most isolated sign language datasets that only annotate isolated word labels, datasets like ASL-LEX and ASLLVD offer additional phonological information and linguistic annotations. This provides researchers with more diverse entry points. For instance, Kezar [162] proposed improving sign language recognition through phonological enhancements. Therefore, it is desirable that the annotation of the dataset be as comprehensive as possible to provide researchers with a variety of research perspectives. In addition, from the perspective of dataset construction, in order to reduce bias in the dataset, it should include representative samples from different ages, genders, body types and physical characteristics, and clothing, while ensuring necessary privacy protection for signers.

V. CHALLENGES AND FUTURE DIRECTIONS

Currently, research in sign language recognition has reached a kind of bottleneck. On one hand, isolated sign language recognition is near perfect in small to medium sized datasets, but still underperforms on large scale datasets. On the other hand, continuous sign language recognition can achieve good results in simple scenarios, but there is still much room for improvement in complex scenarios. The specific questions are as follows, which are also the future trends.

A. THE LENGTH OF VIDEOS IS SHORT

To the best of our knowledge, the length of the videos is short in most datasets especially for continuous sign language datasets. Nevertheless, in reality, sign language is mostly long sentences and there are no signs of division between sentences. It is not known whether the current method can be used in real-life long-sentence sign language recognition. On the other hand, current methods lack the ability to handle long videos, RNN-based methods suffer from long-term dependencies, and Transformer-based methods increase the number of operations exponentially. Therefore, the improvement of the model's ability to process long sentences and to accurately segment between sentences, as well as its adaptability to short sentences, still requires further research.

B. UNSEEN SENTENCE AND SIGNER INDEPENDENT CONDITIONS ARE NOT CONSIDERED

The recognition of different signers and unseen sentence are important in sign language recognition. Signer independent means that signers used for validation are different from those used for training while unseen sentence means that sentences in the testing set have never occurred but each of their words has appeared in other sentences in the training set. This is relevant to whether sign language recognition can actually be used for practical applications. References [70], [79], [92], [103], and [137] all tested their models on signer independent and unseen sentence conditions but the majority of studies didn't. Due to the variations among signers in aspects such as speed and body dimensions, models may capture individual traits, leading to challenges in performance when unseen signers are presented [163]. Methods that are not demonstrator-independent often struggle to perform well with signers outside the training set, resulting in a notable performance decline when new users employ the system. Regrettably, collecting sufficient training data from each new user to retrain the sign language recognition (SLR) models is impractical. In contrast, signer-independent SLR offers greater practicality, enabling new users to directly utilize the system without the need for gathering data for training. For the recognition of unseen sentences, one aspect involves researching how to enable the model to learn lexical-level features from massive datasets of sentence sequences. On another note, pursuing more effective algorithms for vocabulary segmentation remains a significant and meaningful direction for research.

C. NO ABILITY TO RECOGNIZE SIGN LANGUAGE OUTSIDE THE DATASET

Sign language vocabulary is so large that it is difficult to capture it all by building a dataset, so getting the models to learn the new sign language themselves will be one of the main solutions. Currently, some scholars [164], [165], [166] have carried out research on isolated sign language recognition based on few-shot learning or zero-shot learning. There is still a large room for improvement. As for continuous sign language, we haven't found any research on few-shot learning or zero-shot learning. Consequently, conducting research on few-shot and zero-shot learning for isolated and continuous sign language remains a highly challenging task.

D. THE RECOGNITION ON MULTI-PERSON CONDITION IS NOT CONSIDERED

All the samples in current datasets concentrate on single signer. However, it needs to be sufficiently robust in multi-person scenarios if sign language recognition is to be used in real-world applications, for in real-life scenarios there is often interference from other people. Therefore, designing and training robust models to focus solely on the signer presents a worthwhile direction for research.

E. THE COMPLEXITY OF THE MODEL IS IN HIGH LEVEL

Although sign language recognition has achieved good results on servers, most methods have complex models and are difficult to implement on portable devices. However, the goal of sign language recognition is to provide an aid to everyday communication between the deaf and people without hearing impairments, which is impossible through computers. Therefore, a future key research direction is the lightweighting of models to facilitate deployment on portable devices.

F. ONLINE RECOGNITION IS NOT PERFORMED

Both BERT and BiLSTM models are designed to capture contextual information from both past and future elements in a sequence. Most work is now done with recorded data sets as the validation target, which means the contextual information can be easily accessed by BiLSTM or BERT [75], [79], [92], [102], [103], [104], [132], [135]. Nevertheless, in practice, sign language presentations are unpredictable and we can only predict with the information in the previous section. Therefore, some methods based on BiLSTM and BERT will no longer work. How to use one-direction sequences for prediction and achieve better recognition results will become an important research topic.

VI. CONCLUSION

Sign language recognition is a vital area of research because it is so relevant to our lives. The use of computer-aided sign language recognition is of great social importance as it can build bridges between the deaf and people without hearing impairments and help integrate deaf people into society. In this paper, we gave an introduction on sign language recognition, including methods based on traditional approaches and deep learning, metrics, datasets, challenges and future directions. We especially introduce some state-of-the-art methods like Transformer-based sign language recognition networks, and the characteristics of different datasets. The advantages and disadvantages of different methods are analyzed and compared, and the structures of different methods are split and introduced. Currently, as algorithms continue to evolve, sign language recognition is showing increasingly good results on large datasets based on deep learning. Sign language recognition based on traditional methods has gradually become less applicable for reasons such as susceptibility to factors such as light and occlusion, and poor effect on large-scale data processing. On the other hand, the availability of many large datasets has allowed for more effective validation of recognition methods. Yet we should also see that due to the complexity of the model, signer independence and some other reasons, sign language recognition still has a long way to go before practical applications.In response to the primary challenges present in SLR, we conducted an in-depth analysis and provided potential research directions for researchers to consider. We hope that this review will help scholars in the field of sign language recognition research, as well as those who will be conducting research on sign language recognition by providing them with some ideas for their research.

REFERENCES

[1] WHO. (2024). *Deafness and Hearing Loss*. Accessed: May 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] WHO. (2021). *WHO: 1 in 4 People Projected to Have Hearing Problems by 2050*. [Online]. Available: https://www.who.int/news/item/02-03-2021-who-1-in-4-people-projected-to-have-hearing-problems-by-2050

[3] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, Jan. 2019.

[4] I. A. Adeyanju, O. O. Bello, and M. A. Adegboye, "Machine learning methods for sign language recognition: A critical review and analysis," *Intell. Syst. Appl.*, vol. 12, Nov. 2021, Art. no. 200056, doi: 10.1016/j.iswa.2021.200056.

[5] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," *IEEE Access*, vol. 9, pp. 126917–126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

[6] E.-S.-M. El-Alfy and H. Luqman, "A comprehensive survey and taxonomy of sign language research," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105198, doi: 10.1016/j.engappai.2022.105198.

[7] A. Núñez-Marcos, O. Perez-de-Viñaspre, and G. Labaka, "A survey on sign language machine translation," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118993, doi: 10.1016/j.eswa.2022.118993.

[8] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113794, doi: 10.1016/j.eswa.2020.113794.

[9] S. Subburaj and S. Murugavalli, "Survey on sign language recognition in context of vision-based and deep learning," *Meas., Sensors*, vol. 23, Oct. 2022, Art. no. 100385, doi: 10.1016/j.measen.2022.100385.

[10] A. Sultan, W. Makram, M. Kayed, and A. A. Ali, "Sign language identification and recognition: A comparative study," *Open Comput. Sci.*, vol. 12, no. 1, pp. 191–210, May 2022, doi: 10.1515/comp-2022-0240.

[11] D. L. Quam, G. B. Williams, J. R. Agnew, and P. C. Browne, "An experimental determination of human hand accuracy with a DataGlove," *Proc. Hum. Factors Soc. Annu. Meeting*, vol. 33, no. 5, pp. 315–319, Oct. 1989.

[12] Y. Iwai, K. Watanabe, Y. Yagi, and M. Yachida, "Gesture recognition by using colored gloves," in *Proc. IEEE Int. Conf. Syst., Man Cybern., Inf. Intell. Syst.*, Jun. 1996, pp. 76–81.

[13] T. Starner and A. Pentland, "Real-time American sign language recognition from video using hidden Markov models," in *Proc. Int. Symp. Comput. Vision—ISCV*, IEEE, 1995, pp. 265–270.

[14] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using design and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Aug. 1998.

[15] N. Cihan Camgöz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10020–10030.

[16] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3793–3802.

[17] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3416–3424.

[18] M. B. Waldron and S. Kim, "Isolated ASL sign recognition system for deaf persons," *IEEE Trans. Rehabil. Eng.*, vol. 3, no. 3, pp. 261–271, May 1993.

[19] Q. Fu, J. Fu, J. Guo, S. Guo, and X. Li, "Gesture recognition based on BP neural network and data glove," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Oct. 2020, pp. 1918–1922.

[20] G. Fang and W. Gao, "A SRN/HMM system for signer-independent continuous sign language recognition," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nov. 2002, pp. 312–317.

[21] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *IEEE Trans. Syst., Man, Cybern., A, Syst. Hum.*, vol. 37, no. 1, pp. 1–9, Jan. 2007.

[22] Y. Okayasu, T. Ozawa, M. Dahlan, H. Nishimura, and H. Tanaka, "Performance enhancement by combining visual clues to identify sign language motions," in *Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process. (PACRIM)*, Aug. 2017, pp. 1–4.

[23] B. Bauer and H. Hienz, "Relevant features for video-based continuous sign language recognition," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 440–445.

[24] J. Lin and Y. Ding, "A temporal hand gesture recognition system based on hog and motion trajectory," *Optik*, vol. 124, no. 24, pp. 6795–6798, Dec. 2013.

[25] S. Auephanwiriyakul, S. Phitakwinai, W. Suttapak, P. Chanda, and N. Theera-Umpon, "Thai sign language translation using scale invariant feature transform and hidden Markov models," *Pattern Recognit. Lett.*, vol. 34, no. 11, pp. 1291–1298, Aug. 2013.

[26] M. A. Rahim, A. S. M. Miah, A. Sayeed, and J. Shin, "Hand gesture recognition based on optimal segmentation in human–computer interaction," in *Proc. 3rd IEEE Int. Conf. Knowl. Innov. Invention (ICKII)*, Aug. 2020, pp. 163–166.

[27] A. S. M. Miah, J. Shin, M. A. M. Hasan, and M. A. Rahim, "BenSignNet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network," *Appl. Sci.*, vol. 12, no. 8, p. 3933, Apr. 2022.

[28] A. S. M. Miah, J. Shin, M. A. M. Hasan, M. A. Rahim, and Y. Okuyama, "Rotation, translation and scale invariant sign word recognition using deep learning," *Comput. Syst. Sci. Eng.*, vol. 44, no. 3, pp. 2521–2536, 2023.

[29] Y. Ming, "Hand fine-motion recognition based on 3D mesh MoSIFT feature descriptor," *Neurocomputing*, vol. 151, pp. 574–582, Mar. 2015.

[30] K. M. Lim, A. W. C. Tan, and S. C. Tan, "Block-based histogram of optical flow for isolated sign language recognition," *J. Vis. Commun. Image Represent.*, vol. 40, pp. 538–545, Oct. 2016.

[31] S. Katoch, V. Singh, and U. S. Tiwary, "Indian sign language recognition system using SURF with SVM and CNN," *Array*, vol. 14, Jul. 2022, Art. no. 100141.

[32] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 462–477, Mar. 2010.

[33] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. Comput. Vis.*, vol. 46, no. 1, pp. 81–96, 2002.

[34] S.-H. Yu, C.-L. Huang, S.-C. Hsu, H.-W. Lin, and H.-W. Wang, "Vision-based continuous sign language recognition using product HMM," in *Proc. 1st Asian Conf. Pattern Recognit.*, Nov. 2011, pp. 510–514.

[35] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Comput. Vis. Image Understand.*, vol. 141, pp. 108–125, Dec. 2015.

[36] M. Dilsizian, P. Yanovich, S. Wang, C. Neidle, and D. Metaxas, "A new framework for sign language recognition based on 3D handshape identification and linguistic modeling," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 1924–1929.

[37] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, Nov. 2004.

[38] M. Hassan, K. Assaleh, and T. Shanableh, "Multiple proposals for continuous Arabic sign language recognition," *Sens. Imag.*, vol. 20, no. 1, p. 4, Dec. 2019.

[39] W. Gao, G. Fang, D. Zhao, and Y. Chen, "A Chinese sign language recognition system based on SOFM/SRN/HMM," *Pattern Recognit.*, vol. 37, no. 12, pp. 2389–2402, Dec. 2004.

[40] M. Maebatake, I. Suzuki, M. Nishida, Y. Horiuchi, and S. Kuroiwa, "Sign language recognition based on position and movement using multi-stream HMM," in *Proc. 2nd Int. Symp. Universal Commun.*, Dec. 2008, pp. 478–481.

[41] S. Theodorakis, A. Katsamanis, and P. Maragos, "Product-HMMs for automatic sign language recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 1601–1604.

[42] C.-B. Park and S.-W. Lee, "Real-time 3D pointing gesture recognition for mobile robots with cascade HMM and particle filter," *Image Vis. Comput.*, vol. 29, no. 1, pp. 51–63, Jan. 2011.

[43] R. Yang and S. Sarkar, "Detecting coarticulation in sign language using conditional random fields," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 108–112.

[44] H.-D. Yang and S.-W. Lee, "Simultaneous spotting of signs and fingerspellings based on hierarchical conditional random fields and boostmap embeddings," *Pattern Recognit.*, vol. 43, no. 8, pp. 2858–2870, Aug. 2010.

[45] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1264–1277, Jul. 2009.

[46] W. W. Kong and S. Ranganath, "Towards subject independent continuous sign language recognition: A segment and merge approach," *Pattern Recognit.*, vol. 47, no. 3, pp. 1294–1308, Mar. 2014.

[47] S. Mathur and P. Sharma, "Sign language gesture recognition using Zernike moments and DTW," in *Proc. 5th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2018, pp. 586–591.

[48] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, nos. 1–3, pp. 366–380, Dec. 2009.

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[51] S. Bai, J. Zico Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.

[52] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[53] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.

[56] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization ACL*, Barcelona, Spain, 2004, pp. 74–81.

[57] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1311–1325, Dec. 2018.

[58] J. Forster, C. Schmidt, T. Hoyoux, and O. Koller, "RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus," in *Proc. 8th Int. Conf. Lang. Resour. Eval.*, 2012, pp. 3785–3789.

[59] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from RGB video using pose flow and self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3436–3445.

[60] G. H. Martinez, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh, "Single-network whole-body pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6981–6990.

[61] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3075–3084.

[62] A. S. M. Miah, M. A. M. Hasan, J. Shin, Y. Okuyama, and Y. Tomioka, "Multistage spatial attention-based neural network for hand gesture recognition," *Computers*, vol. 12, no. 1, p. 13, Jan. 2023.

[63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[64] T. Shanableh, "Two-stage deep learning solution for continuous Arabic sign language recognition using word count prediction and motion images," *IEEE Access*, vol. 11, pp. 126823–126833, 2023.

[65] J. Pu, "Video-based sign language recognition with deep learning," Ph.D. dissertation, Dept. Info. Comm. Eng., Univ. Sci. Technol. China, Hefei, China, 2020.

[66] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Jun. 1998.

[67] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 2257–2264.

[68] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous sign language recognition with correlation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2529–2539.

[69] L. Hu, L. Gao, Z. Liu, and W. Feng, "Self-emphasizing network for continuous sign language recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2023, no. 1, pp. 854–862.

[70] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4160–4169.

[71] N. Sarhan and S. Frintrop, "Transfer learning for videos: From action recognition to sign language recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1811–1815.

[72] Z. Zhou, K.-S. Lui, V. W. L. Tam, and E. Y. Lam, "Applying (3+2+1)D residual neural network with frame selection for Hong Kong sign language recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4296–4302.

[73] Y. Gao, R. Hu, T. Ma, S. Guo, Y. Yang, and X. Zhou, "Dynamic sign language recognition based on improved R(2+1)D algorithm," in *Proc. 7th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2022, pp. 7–15.

[74] X. Han, F. Lu, J. Yin, G. Tian, and J. Liu, "Sign language recognition based on R(2+1)D with spatial–temporal–channel attention," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 4, pp. 687–698, Aug. 2022.

[75] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1610–1618.

[76] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*.

[77] Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "SF-Net: Structured feature network for continuous sign language recognition," 2019, *arXiv:1908.01341*.

[78] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai, "Fully convolutional networks for continuous sign language recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 697–714.

[79] L. Gao, H. Li, Z. Liu, Z. Liu, L. Wan, and W. Feng, "RNN-transducer based Chinese sign language recognition," *Neurocomputing*, vol. 434, pp. 45–54, Apr. 2021.

[80] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[81] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2012, pp. 864–877.

[82] A. Graves, "Sequence transduction with recurrent neural networks," 2012, *arXiv:1211.3711*.

[83] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[84] J. Shin, A. S. Musa Miah, M. A. M. Hasan, K. Hirooka, K. Suzuki, H.-S. Lee, and S.-W. Jang, "Korean sign language recognition using transformer-based deep neural network," *Appl. Sci.*, vol. 13, no. 5, p. 3029, Feb. 2023.

[85] A. Kozlov, V. Andronov, and Y. Gritsenko, "Lightweight network architecture for real-time action recognition," in *Proc. 35th Annu. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 2074–2080.

[86] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[87] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340–181355, 2020.

[88] Y. Du, P. Xie, M. Wang, X. Hu, Z. Zhao, and J. Liu, "Full transformer network with masking future for word-level sign language recognition," *Neurocomputing*, vol. 500, pp. 115–123, Aug. 2022.

[89] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[90] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 172–186.

[91] K. Yin and J. Read, "Better sign language translation with STMC-transformer," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 5975–5989.

[92] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial–temporal multi-cue network for continuous sign language recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, no. 7, pp. 13009–13016.

[93] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[94] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7784–7793.

[95] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 301–319.

[96] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li, "TSPNet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 12034–12045.

[97] Z. Guo, Y. Hou, C. Hou, and W. Yin, "Locality-aware transformer for video-based sign language translation," *IEEE Signal Process. Lett.*, vol. 30, pp. 364–368, 2023.

[98] Z. Cui, W. Zhang, Z. Li, and Z. Wang, "Spatial–temporal transformer for end-to-end sign language recognition," *Complex Intell. Syst.*, vol. 9, no. 4, pp. 4645–4656, 2023.

[99] R. Zuo and B. Mak, "C^{2}SLR: Consistency-enhanced continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5121–5130.

[100] R. Hinrichs, A. J. Y. Sitcheu, and J. Ostermann, "Continuous sign-language recognition using transformers and augmented pose estimation," in *Proc. ICPRAM*, 2023, pp. 672–678.

[101] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[102] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, "SignBERT: Pre-training of hand-model-aware representation for sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11067–11076.

[103] Z. Zhou, V. W. L. Tam, and E. Y. Lam, "SignBERT: A BERT-based deep learning framework for continuous sign language recognition," *IEEE Access*, vol. 9, pp. 161669–161682, 2021.

[104] Z. Zhou, V. W. L. Tam, and E. Y. Lam, "A cross-attention BERT-based framework for continuous sign language recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 1818–1822, 2022.

[105] W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, "BEST: BERT pre-training for sign language recognition with coupling tokenization," in *Proc. AAAI Conf. Artif. Intell.*, 2023, no. 3, pp. 3597–3605.

[106] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial–temporal relationships for 3D pose estimation via graph convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2272–2281.

[107] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," 2022, *arXiv:2201.02610*.

[108] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[109] R. Elakkiya, P. Vijayakumar, and N. Kumar, "An optimized generative adversarial network based continuous sign language classification," *Expert Syst. Appl.*, vol. 182, Nov. 2021, Art. no. 115276.

[110] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, vol. 21, no. 7, p. 2437, Apr. 2021.

[111] J. Kan, K. Hu, M. Hagenbuchner, A. C. Tsoi, M. Bennamoun, and Z. Wang, "Sign language translation with hierarchical spatio-temporal graph neural network," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2131–2140.

[112] Z. Wang and J. Zhang, "Continuous sign language recognition based on multi-part skeleton data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[113] A. S. Musa Miah, J. Shin, M. Al Mehedi Hasan, Y. Fujimoto, and A. Nobuyoshi, "Skeleton-based hand gesture recognition using geometric features and spatio-temporal deep learning approach," in *Proc. 11th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Sep. 2023, pp. 1–6.

[114] A. S. M. Miah, M. A. M. Hasan, S. Nishimura, and J. Shin, "Sign language recognition using graph and general deep neural network based on large scale dataset," *IEEE Access*, vol. 12, pp. 34553–34569, 2024.

[115] A. S. M. Miah, M. A. M. Hasan, Y. Tomioka, and J. Shin, "Hand gesture recognition for multi-culture sign language using graph and general deep learning network," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 144–155, 2024.

[116] R. Elakkiya and K. Selvamani, "Extricating manual and non-manual features for subunit level medical sign modelling in automatic sign language classification and recognition," *J. Med. Syst.*, vol. 41, no. 11, p. 1, Nov. 2017.

[117] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[118] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3408–3418.

[119] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling GCN with dropgraph module for skeleton-based action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 536–553.

[120] A. S. M. Miah, M. A. M. Hasan, S.-W. Jang, H.-S. Lee, and J. Shin, "Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition," *Electronics*, vol. 12, no. 13, p. 2841, Jun. 2023.

[121] A. S. M. Miah, Md. A. M. Hasan, and J. Shin, "Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model," *IEEE Access*, vol. 11, pp. 4703–4716, 2023.

[122] J. Shin, A. S. M. Miah, K. Suzuki, K. Hirooka, and M. A. M. Hasan, "Dynamic Korean sign language recognition using pose estimation based and attention-based neural network," *IEEE Access*, vol. 11, pp. 143501–143513, 2023.

[123] Q. Guo, S. Zhang, L. Tan, K. Fang, and Y. Du, "Interactive attention and improved GCN for continuous sign language recognition," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104931.

[124] K. Papadimitriou and G. Potamianos, "Sign language recognition via deformable 3D convolutions and modulated graph convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[125] L. Meng and R. Li, "An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network," *Sensors*, vol. 21, no. 4, p. 1120, Feb. 2021.

[126] N. Naz, H. Sajid, S. Ali, O. Hasan, and M. K. Ehsan, "Signgraph: An efficient and accurate pose-based graph convolution approach toward sign language recognition," *IEEE Access*, vol. 11, pp. 19135–19147, 2023.

[127] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 345–360.

[128] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2306–2320, Sep. 2020.

[129] C. Gündüz and H. Polat, "Turkish sign language recognition based on multistream data fusion," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 29, no. 2, pp. 1171–1186, Mar. 2021.

[130] X. Liang, A. Angelopoulou, E. Kapetanios, B. Woll, R. A. Batat, and T. Woolfe, "A multi-modal machine learning approach and toolkit to automate recognition of early stages of dementia among British sign language users," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 278–293.

[131] J. Zheng, Y. Chen, C. Wu, X. Shi, and S. M. Kamal, "Enhancing neural sign language translation by highlighting the facial expression information," *Neurocomputing*, vol. 464, pp. 462–472, Nov. 2021.

[132] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1880–1891, Jul. 2019.

[133] F. B. Slimane and M. Bouguessa, "Context matters: Self-attention for sign language recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7884–7891.

[134] N. Sarhan, C. Wilms, V. Closius, U. Brefeld, and S. Frintrop, "Hands in focus: Sign language recognition via top-down attention," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 2555–2559.

[135] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11283–11292.

[136] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder–decoder neural network model for sequence to sequence mapping," in *Proc. Interspeech*, Aug. 2017, pp. 1298–1302.

[137] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical LSTM for sign language translation," in *Proc. AAAI*, 2018, pp. 6845–6852.

[138] T. Liu, T. Tao, Y. Zhao, M. Li, and J. Zhu, "A signer-independent sign language recognition method for single-frequency datasets," *Neurocomputing*, vol. 582, May 2024, Art. no. 127479.

[139] R. Zuo, F. Wei, and B. Mak, "Natural language-assisted sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14890–14900.

[140] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[141] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[142] S. Alyami, H. Luqman, and M. Hammoudeh, "Isolated Arabic sign language recognition using a transformer-based model and landmark keypoints," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 1, pp. 1–19, Jan. 2024.

[143] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1114–1119.

[144] B. Shi, A. M. Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu, "American sign language fingerspelling recognition in the wild," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 145–152.

[145] B. Shi, A. M. D. Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition in the wild with iterative visual attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5399–5408.

[146] U. von Agris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[147] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali, "The American sign language lexicon video dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–8.

[148] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey, "ASL-LEX: A lexical database of American sign language," *Behav. Res. Methods*, vol. 49, no. 2, pp. 784–801, Apr. 2017.

[149] Z. S. Sehyr, N. Caselli, A. M. Cohen-Goldberg, and K. Emmorey, "The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in American sign language," *J. Deaf Stud. Deaf Educ.*, vol. 26, no. 2, pp. 263–277, Mar. 2021.

[150] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i-Nieto, "How2Sign: A large-scale multimodal dataset for continuous American sign language," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2734–2743.

[151] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1448–1458.

[152] H. Reza Vaezi Joze and O. Koller, "MS-ASL: A large-scale data set and benchmark for understanding American sign language," 2018, *arXiv:1812.01053*.

[153] A. Mavi and Z. Dikle, "A new 27 class sign language dataset collected from 173 individuals," 2022, *arXiv:2203.03859*.

[154] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[155] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, and A. Zisserman, "BBC-oxford British sign language dataset," 2021, *arXiv:2111.03635*.

[156] J. Fink, B. Frénay, L. Meurant, and A. Cleve, "LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[157] F. Ronchetti, F. Quiroga, and L. Lanzarini, "LSA64: An Argentinian sign language dataset," in *Proc. Cong. Argentino Ciencias Comput. (CACIC)*, 2016, pp. 794–803.

[158] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 1750–1762, 2022.

[159] E. Gutierrez-Sigut, B. Costello, C. Baus, and M. Carreiras, "LSE-sign: A lexical database for Spanish sign language," *Behav. Res. Methods*, vol. 48, no. 1, pp. 123–137, Mar. 2016.

[160] A. A. I. Sidig, H. Luqman, S. Mahmoud, and M. Mohandes, "KArSL: Arabic sign language database," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, Jan. 2021.

[161] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1316–1325.

[162] L. Kezar, J. Thomason, and Z. S. Sehyr, "Improving sign recognition with phonology," 2023, *arXiv:2302.05759*.

[163] T. Liu, T. Tao, Y. Zhao, M. Li, and J. Zhu, "A signer-independent sign language recognition method for the single-frequency dataset," *Neurocomputing*, vol. 582, May 2024, Art. no. 127479.

[164] Y. C. Bilge, R. G. Cinbis, and N. Ikizler-Cinbis, "Towards zero-shot sign language recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1217–1232, Jan. 2023.

[165] A. Kamzin, V. N. S. A. Amperayani, P. Sukhapalli, A. Banerjee, and S. K. S. Gupta, "Concept embedding through canonical forms: A case study on zero-shot ASL recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6157–6164.

[166] R. A. Nihal, S. Rahman, N. M. Broti, and S. A. Deowan, "Bangla sign alphabet recognition with zero-shot and transfer learning," *Pattern Recognit. Lett.*, vol. 150, pp. 84–93, Oct. 2021.

TANGFEI TAO received the B.S., M.S., and Ph.D. degrees from Xi'an Jiaotong University, China, in 1997, 2000, and 2006, respectively. He is currently an Associate Professor with the School of Mechanical Engineering, Xi'an Jiaotong University. His research interests include sign language recognition, deep learning, machine vision, pattern recognition, condition monitoring, and fault diagnosis.

YIZHE ZHAO received the B.E. degree from Hunan University, Changsha, China, in 2022. He is currently pursuing the M.Sc. degree with Xi'an Jiaotong University. His research interests include computer vision and sign language recognition.

TIANYU LIU received the B.E. degree from Southwest Jiaotong University, Chengdu, China, in 2021. He is currently pursuing the M.Sc. degree with Xi'an Jiaotong University, China, under the supervision of Assoc. Prof. T. Tao. His research interests include computer vision and sign language recognition.

JIELI ZHU received the master's degree in instrument science and technology from Xi'an Jiaotong University, China, in 2023. He is currently pursuing the Ph.D. degree with the Department of Computer Science, UiT—The Arctic University of Norway. His research interests include machine learning and computer vision.