Sign Language Production: A Review

Razieh Rastgoo[1,2], Kourosh Kiani[1], Sergio Escalera[3], Mohammad Sabokrou[2]
[1]Semnan University [2]Institute for Research in Fundamental Sciences (IPM)
[3]Universitat de Barcelona and Computer Vision Center
rrastgoo@semnan.ac.ir, kourosh.kiani@semnan.ac.ir, sergio@maia.ub.es, sabokro@ipm.ir

Abstract

Sign Language is the dominant yet non-primary form of communication language used in the deaf and hearing-impaired community. To make an easy and mutual communication between the hearing-impaired and the hearing communities, building a robust system capable of translating the spoken language into sign language and vice versa is fundamental. To this end, sign language recognition and production are two necessary parts for making such a two-way system. Sign language recognition and production need to cope with some critical challenges. In this survey, we review recent advances in Sign Language Production (SLP) and related areas using deep learning. This survey aims to briefly summarize recent achievements in SLP, discussing their advantages, limitations, and future directions of research.

1. Introduction

Sign Language is the dominant yet non-primary form of the communication language used in large groups of people in society. According to the World Health Organization (WHO) report in 2020, there are more than 466 million deaf people in the world [88]. There are different forms of sign languages employed by different nationalities such as USA [87], Argentina [26], Poland [71], Germany [36], Greek [27], Spain [3], China [2], Korea [61], Iran [28], and so on. To make an easy and mutual communication between the hearing-impaired and the hearing communities, building a robust system capable of translating the spoken languages into sign languages and vice versa is fundamental. To this end, sign language recognition and production are two necessary parts for making such a two-way system. While the first part, sign language recognition, has rapidly advanced in recent years [64, 65, 66, 67, 68, 69, 50, 62, 59, 8, 48], the latest one, Sign Language Production (SLP), is still a very challenging problem involving an interpretation between visual and linguistic information [79]. Proposed systems in sign language recognition generally map signs into the spoken language in the form of text transcription [69]. However, SLP systems perform the reverse procedure.

Sign language recognition and production are coping with some critical challenges [69, 79]. One of them is the visual variability of signs, which is affected by hand-shape, palm orientation, movement, location, facial expressions, and other non-hand signals. These differences in sign appearance produce a large intra-class variability and low inter-class variability. This makes it hard to provide a robust and universal system capable of recognizing different sign types. Another challenge is developing a photo-realistic SLP system to generate the corresponding sign digit, word, or sentence from a text or voice in spoken language in a real-world situation. The challenge corresponding to the grammatical rules and linguistic structures of the sign language is another critical challenge in this area. Translating between spoken and sign language is a complex problem. This is not a simple mapping problem from text/voice to signs word-by-word. This challenge comes from the differences between the tokenization and ordering of words in the spoken and sign languages.

Another challenge is related to the application area. Most of the applications in sign language focus on sign language recognition such as robotics [20], human–computer interaction [5], education [19], computer games [70], recognition of children with autism [11], automatic sign-language interpretation [90], decision support for medical diagnosis of motor skills disorders [10], home-based rehabilitation [17] [57], and virtual reality [82]. This is due to the misunderstanding of the hearing people thinking that deaf people are much more comfortable with reading spoken language; therefore, it is not necessary to translate the reading spoken language into sign language. This is not true since there is no guarantee that a deaf person is familiar with the reading and writing forms of a speaking language. In some languages, these two forms are completely different from each other. While there are some detailed and well-presented reviews in sign language recognition [30, 69], SLP suffers from such a detailed review. Here, we present a survey, including recent works in the SLP, with the aim of discussing

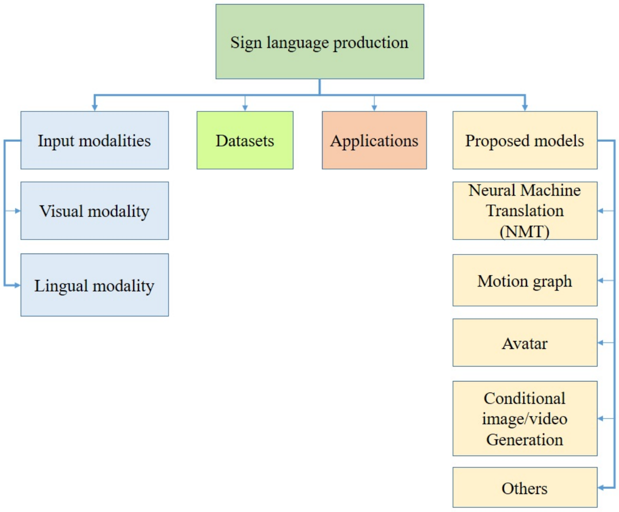


Figure 1. The proposed taxonomy of the reviewed works in SLP.

advances and weaknesses of this area. We focus on deep learning-based models to analyze state-of-the-art on SLP. The remainder of this paper is organized as follows. Section 2 presents a taxonomy that summarizes the main concepts related to SLP. Finally, section 3 discusses the developments, advantages, and limitations in SLP and comments on possible lines for future research.

2. SLP Taxonomy

In this section, we present a taxonomy that summarizes the main concepts related to deep learning in SLP. We categorize recent works in SLP providing separate discussion in each category. In the rest of this section, we explain different input modalities, datasets, applications, and proposed models. Figure 1 shows the proposed taxonomy described in this section.

2.1. Input modalities

Generally, vision and language are two input modalities in SLP. While the visual modality includes the captured image/video data, the linguistic modality for the spoken language contains the text input from the natural language. Computer vision and natural language processing techniques are necessary to process these input modalities.

Visual modality: RGB and skeleton are two common types of input data used in SLP models. While RGB images/videos contain high-resolution content, skeleton inputs decrease the input dimension necessary to feed to the model and assist in making a low-complex and fast model. Only one letter or digit is included in an RGB image input. The spatial features corresponding to the input image can be extracted using computer vision-based techniques, especially deep learning-based models. In recent years, Convolutional Neural Networks (CNN) achieved outstanding performance for spatial feature extraction from an input image [53]. Furthermore, generative models, such as Generative Adversarial Networks (GAN), can use the CNN as an encoder or decoder block to generate a sign image/video. Due to the temporal dimension of RGB video inputs, the processing of this input modality is more complicated than the RGB image input. Most of the proposed models in SLP use the RGB video as input [13, 72, 73, 79]. An RGB sign video can correspond to one sign word or some concatenated sign words, in the form of a sign sentence. GAN and LSTM are the most used deep learning-based models in SLP for static and dynamic visual modalities. While successful results have been achieved using these models, more effort is necessary to generate more lifelike sign images/videos in order to improve the communication interface with the Deaf community.

Lingual modality: Text input is the most common form of linguistic modality. To process the input text, different models are used [76, 80]. While text processing is low-complex compared to image/video processing, text translation tasks are complex. Among the deep learning-based models, the Neural Machine Translation (NMT) model is the most used model for input text processing. Other Seq2Seq models [80], such as Recurrent Neural Network (RNN)-based models, proved their effectiveness in many tasks. While successful results were achieved using these models, more effort is necessary to overcome the existing challenges in the translation task. One of the challenges in translation is related to domain adaptation due to different words styles, translations, and meaning in different languages. Thus, a critical requirement of developing machine translation systems is to target a specific domain. Transfer learning, training the translation system in a general domain followed by fine-tuning on in-domain data for a few epochs is a common approach in coping with this challenge. Another challenge is regarding the amount of training data. Since a main property of deep learning-based models is the mutual relation between the amount of data and model performance, large amount of data is necessary to provide a good generalization. Another challenge is the poor performance of machine translation systems on uncommon and unseen words. To cope with these words, byte-pair encoding, such as stemming or compound-splitting, can be used for rare words translation. As another challenge, the machine translation systems are not properly able to translate long sentences. However, the attention model [86] partially deals with this challenge for short sentences. Furthermore, the challenge regarding the word alignment is more critical in the reverse translation, that is translating back from the target language to the source language.

2.2. Datasets

While there are some large-scale and annotated datasets available for sign language recognition, there are only few publicly available large-scale datasets for SLP. Two public datasets, RWTH-Phoenix-2014T [14] and How2Sign [22] are the most used datasets in sign language translation. The former includes German sign language sentences that can be used for text-to-sign language translation. This dataset is an extended version of the continuous sign language recognition dataset, PHOENIX-2014 [29]. RWTH-PHOENIX-Weather 2014T includes a total of 8257 sequences performed by 9 signers. There are 1066 sign glosses and 2887 spoken language vocabularies in this dataset. Furthermore, the gloss annotations corresponding to the spoken language sentences have been included in the dataset. The later dataset, How2Sing, is a recently proposed multi-modal dataset used for speech-to-sign language translation. This dataset contains a total of 38611 sequences and 4k vocabularies performed by 10 signers. Like the former dataset, the annotation for sign glosses have been included in this dataset.

Though RWTH-PHOENIX-Weather 2014T and How2Sign provided SLP evaluation benchmarks, they are not enough for generalization of SLP models. Furthermore, these datasets just include German and American sentences. In line with the aim of providing an easy to use application for mutual communication between the Deaf and hearing communities, new large-scale datasets with enough variety and diversity in different sign languages is required. The point is that the signs are generally dexterous and the signing procedure involves different channels, including arms, hands, body, gaze, and facial expressions simultaneously. To capture such gestures requires a trade-off between capture cost, measurement (space and time) accuracy, and the production spontaneity. Furthermore, different equipment is used for data recording such as wired Cybergloves, Polhemus magnetic sensors, headset equipped with an infrared camera, emitting diodes and reflectors. Synchronization between different channels captured by the aforementioned devices is key in data collection and annotation. Another challenge is related to the capturing complexity of the hand movement using some capturing devices, such as Cybergloves. Hard calibration and deviation during data recording are some difficulties of these acquisition devices. The synchronization of external devices, hand modeling accuracy, data loss, noise in the capturing process, facial expression processing, gaze direction, and data annotation are additional challenges. Given these challenges, providing a large and diverse dataset for SLP, including spoken language and sign language annotations, is difficult. Figure 2 shows existing datasets for SLP.

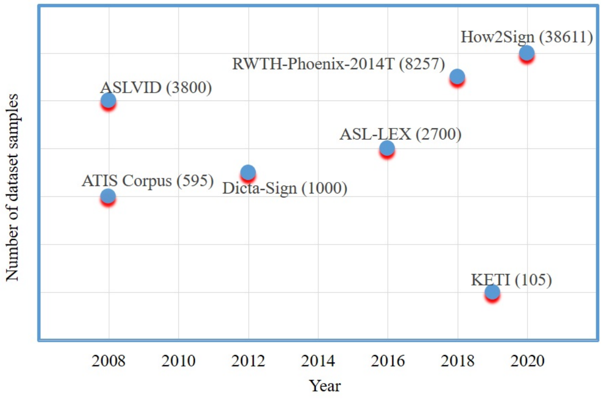


Figure 2. SLP datasets in time. The number of samples for each dataset is shown in brackets

2.3. Applications

With the advent of the potent methodologies and techniques in recent years, machine translation applications have become more efficient and trustworthy. One of the early efforts on machine translation is dated back to the sixties, where a model was proposed to translate from Russian to English. This model defined the machine translation task as a phase of encryption and decryption. Nowadays, the standard machine translation models fall into three main categories: rule-based grammatical models, statistical models, and example-based models. Deep learning-based models, such as Seq2Seq and NMT models, fall into the third category, and showed promising results in SLP.

To translate from a source language to a target language, a corpus to perform some preprocessing steps is needed, including boundary detection, word tokenization, and chunking. While there are different corpora for most spoken languages, sign language lacks from such a large and diverse corpora. American Sign Language (ASL), as the largest sign language community in the World, is the most-used sign language in the developed applications for SLP. Since Deaf people may not be able to read or write the spoken language, they need some tools for communication with other people in society. Furthermore, many interesting and useful applications in Internet are not accessible for the Deaf community. However, we are still far from having applications accessible for Deaf people with large vocabularies/sentences from real-world scenarios. One of the main challenges for these applications is a license right for usage. Only some of these applications are freely available. Another challenge is the lack of generalization of current applications, which are developed for the requirements of very specific application scenarios.

2.4. Proposed models

In this section, we review recent works in SLP. These works are presented and discussed in five categories: Avatar approaches, NMT approaches, Motion Graph (MG) approaches, Conditional image/video Generation approaches, and other approaches. Table 1 presents a summary of the reviewed models in SLP.

2.4.1 Avatar Approaches

In order to reduce the communication barriers between hearing and hearing-impaired people, sign language interpreters are used as an effective yet costly solution. To inform deaf people quickly in cases where there is no interpreter on hand, researchers are working on novel approaches to providing the content. One of these approaches is sign avatars. Avatar is a technique to display the signed conversation in the absence of the videos corresponding to a human signer. To this end, 3D animated models are employed, which can be stored more efficiently compared to videos. The movements of the fingers, hands, facial gestures, and body can be generated using the avatar. This technique can be programmed to be used in different sign languages. With the advent of computer graphics in recent years, computers and smartphones can generate high-quality animations with smooth transitions between the signs. To capture the motion data of deaf people, some special cameras and sensors are used. Furthermore, a computing method uses to be considered to transfer the body movements into the sign avatar [45].

Two ways to derive the sign avatars include the motion capture data and parametrized glosses. In recent years, some works have been developed exploring avatars animated from the parametrized glosses. VisiCast [7], Tessa [18], eSign [92], dicta-sign [25], JASigning [34], and WebSign [39] are some of them. These works need the sign video annotated via the transcription language, such as HamNoSys[63] or SigML [43]. Although, the non-popularity of these avatars made them unfavorable in the deaf community. Under-articulated, unnatural movements, and missing non-manuals information, such as eye gaze and facial expressions, are some challenges of the avatar approaches. These challenges lead to misunderstanding the final sign language sequences. Furthermore, due to the uncanny valley, the users do not feel comfortable [58] with the robotic motion of the avatars. To tackle these problems, recent works focus on annotating non-manual information such as face, body, and facial expression [23, 24].

Using the data collected from motion capture, avatars can be more usable and acceptable for reviewers (such as the Sign3D project by MocapLab [31]). Highly realistic results are achieved by avatars, but the results are restricted to a small set of phrases. This comes from the cost of the data collection and annotation. Furthermore, avatar data is not a scalable solution and needs expert knowledge to perform a sanity check on the generated data. To cope with these problems and improve performance, deep learning-based models, as the latest machine translation developments, are used. Generative models along with some graphical techniques, such as Motion Graph, are being recently employed [79].

2.4.2 NMT approaches

Machine translators are a practical methodology for translating from one language to another. The first translator comes back to the sixties where the Russian language was translated into English [38]. The translation task requires preprocessing of the source language, including sentence boundary detection, word tokenization, and chunking. These preprocessing tasks are challenging, especially in sign language. Sign Language Translation (SLT) aims to produce/generate spoken language translations from sign language considering different word orders and grammar. The ordering and the number of glosses do not necessary match the words of the spoken language sentences.

Nowadays, there are different types of machine translators, mainly based on grammatical rules, statistics, and examples [60]. As an example-based methodology, some research works have been developed by focusing on translating from text into sign language using Artificial Neural Networks (ANNs), namely NMT [6]. NMT uses ANNs to predict the likelihood of a word sequence, typically modeling entire sentences in a single integrated model.

To enhance the translation performance of long sequences, Bahdanau et al. [6] presented an effective attention mechanism. This mechanism was later improved by Luong et al. [51]. Camgoz et al. proposed a combination of a seq2seq model with a CNN to translate sign videos to spoken language sentences [12]. Guo et al. [35] designed a hybrid model including the combination of a 3D Convolutional Neural Network (3DCNN) and Long Short Term Memory (LSTM)-based [56, 37] encoder-decoder to translate from sign videos to text outputs. Results on their own dataset show a 0.071 % improvement margin of the precision metric compared to state-of-the-art models. Dilated convolutions and Transformer are two approaches also used for sign language translation [40, 86]. Stoll et al. [79] proposed a hybrid model to automatic SLP using NMT, GANs, and motion generation. The proposed model generates sign videos from spoken language sentences with a minimal level of data annotation for training. This model first translates spoken language sentences into sign pose sequences. Then, a generative model is used to generate plausible sign language video sequences. Results on the PHOENIX14T Sign Language Translation dataset show comparable results compared to state-of-the-art alternatives.

While NMT-based methods achieved successful results in translation tasks, some major challenges need to be solved. Domain adaptation is the first challenge in this area. Since the translation between different domains is affected by different rules, domain adaptation is a crucial requirement in developing machine translation systems targeted to a specific use case. The second challenge is regarding the amount of training data. Especially in deep learning-based models, increasing the amount of data can lead to better results. Another difficulty is dealing with uncommon words. The translation models perform poorly on these words. Words alignment and adjusting the beam search parameters are other challenges for NMT-based models. The promising results of current deep learning-based models set an underpin to future research in this area.

2.4.3 Motion Graph approaches

Motion Graph (MG), as a computer graphic method for dynamically animating characters, is defined as a directed graph constructed from motion capture data. MG can generate new sequences to satisfy specific goals. In SLP, MG can be combined with NMT-based network to make a continuous-text-to-pose translation. One of the early efforts on MG is dated back to 2002, where a general framework was proposed by Kovar et al. [47] for extracting particular graph walks that satisfy a user's specifications. Distance between two frames was defined as a distance between two point clouds. To make the transitions, alignment and interpolation of the motions and positions were used between the joints. Finally, the branch and bound search algorithm was applied to the graph. In another work, Arikan and Forsyth [1] used the joint positions, velocities, and accelerations parameters to define the distance between two consecutive frames. In addition, the discontinuity between two clips was calculated using a smoothing function. After summarizing the graph, the random search was applied to the graph. Two-layer representation of motion data in another approach was proposed by Lee et al. [49]. In the first layer, data was modeled as a first-order Markov process and the transition probabilities were calculated using the distances of weighted joint angles and velocities. A cluster analysis was performed on the second layer, namely cluster forest, to generalize the motions. Stoll et al. [79] proposed an MG for continuous SLP using pose data. The sign glosses were embedded to an MG with the transition probabilities provided by an NMT decoder at each time step.

Although MG can generate plausible and controllable motion through a database of motion capture, it faces some challenges. The first challenge is regarding the access to data. To show the model potential with a truly diverse set of actions, a large set of data is necessary. Scalability and computational complexity of the graph to select the best transitions are the other challenges in MG. Furthermore, since the number of edges leaving a node increases with the size of the graph, the branching factor in the search algorithm will increase as well.

2.4.4 Conditional image/video generation

The field of automatic image/video generation has experienced a remarkable evolution in recent years. However, the task of video generation is challenging since the content between consequence frames has to be consistent, showing a plausible motion. These challenges are more difficult in SLP due to the need for human video generation. The complexity and variety of actions and appearances in these videos are high and challenging. Controlling the content of the generated videos is crucial yet difficult.

With the recent advances in deep learning, the field of automatic image/video generation has seen different approaches employing neural network-based architectures, such as CNNs [16, 84], RNNs [33, 83], Variational Auto-Encoders (VAEs) [44], conditional VAEs [89], and GAN [32]. VAEs and GANs are generally combined to benefit from the VAE's stability and the GAN's discriminative nature. Most relevant to SLP, a hybrid model, including a VAE and GAN combination, has been proposed to image generation of people [52, 77] and video generation of people performing sign language [78, 85]. Furthermore, there are some models for image/video generation that can be used in SLP. For example, Chen and Koltun [16] proposed a CNN-based model to generate photographic images given semantic label maps. Van den Oord et al. [84] proposed a deep learning-based model, namely PixelRNNs, to sequentially generate the image pixels along the two spatial dimensions. Gregor et al. [33] developed an RNN-based architecture, including an encoder and a decoder network to compress the real images presented during training and refine images after receiving codes. Karras et al. [41] designed a deep generative model, entitled StyleGAN, to adjust the image style at each convolution layer. Kataoka et al. [42] proposed a model using the combination of GAN and attention mechanism. Benefiting from the attention mechanism, this model can generate images containing high detailed content.

While deep learning-based generative models have recently achieved remarkable results, there exist major challenges in their training. Mode collapse, non-convergence and instability, suitable objective function, and optimization algorithm are some of these challenges. However, several strategies have been recently proposed to address a better design and optimization of them. Appropriate design of network architecture, proper objective functions, and optimization algorithms are some of the proposed techniques to improve the performance of deep learning-based models.

2.4.5 Other models

In addition to the previous categories, some models have been proposed to SLP using different deep learning models. For example, Saunders et al. [73] proposed a Progressive Transformers, as a deep learning-based model, to generate continuous sign sequences from spoken language sentences. In another work, Zelinka and Kanis [91] designed a sign language synthesis system focusing on skeletal model production. A feed-forward transformer and a recurrent transformer, as deep learning-based models, along with the attention mechanism have been used to enhance the model performance. Saunders et al. [75] proposed a generative-based model to generate photo-realistic continuous sign videos from text inputs. They combined a transformer with a Mixture Density Network (MDN) to manage the translation from text to skeletal pose. Tornay et al. [81] designed a SLP assessment approach using multi-channel information (hand shape, hand movement, mouthing, facial expression). In this approach, two linguistic aspects are considered: the generated lexeme and the generated forms. Using the capabilities of different methods in this category has led to successful results in SLP. However, the challenge of the model complexity still remains an open issue. Making a trade-off between accuracy vs. task complexity is a key element.

3. Discussion

In this survey, we presented a detailed review on the recent advancements in SLP. We presented a taxonomy that summarizes the main concepts related to SLP. We categorized recent works in SLP providing separate discussion in each category. The proposed taxonomy covered different input modalities, datasets, applications, and proposed models. Here, we summarize main findings:

Input modalities: Generally, vision and language modalities are two input modalities in SLP. While the visual modality includes the captured image/video data, the linguistic modality contains the text input from natural language. Computer vision and natural language processing techniques are necessary to process these input modalities. Both categories benefit from deep learning approaches to improve model performance. RGB and skeleton are two common types of visual input data used in SLP models. While RGB images/videos contain high-resolution content, skeleton inputs decrease the parameter complexity of the model and assist in making a low-complex and fast model. GAN and LSTM are the most used deep learning-based models in SLP for visual inputs. While successful results were achieved using these models, more effort is necessary to generate more lifelike and high-resolution sign images/videos acceptable by the Deaf community. Among the deep learning-based models for lingual modality, the NMT model is the most used model for input text processing. Other Seq2Seq models, such as RNN-based models, proved their effectiveness in many tasks. While accurate results were achieved using these models, more effort is necessary to overcome the existing challenges in the translation task, such as domain adaptation, uncommon words, words alignment, and word tokenization.

Datasets: The lack of a large annotated dataset is one of the major challenges in SLP. The collection and annotation of sign language data is an expensive task that needs the collaboration of linguistic experts and native speakers. While there are some publicly available datasets for SLP [4, 9, 14, 15, 22, 46, 54], they suffer from weakly annotated data for sign language. Furthermore, most of the available datasets in SLP contain a restricted domain of the vocabularies/sentences. To make a real-world communication between the Deaf and hearing communities, access to a large-scale continuous sign language dataset, segmented on the sentence level, is necessary. In such dataset, a paired form of the continuous sign language sentence and the corresponding spoken language sentence needs to be included. Just a few datasets meet these criteria [12, 22, 46, 91] . The point is that most of the aforementioned datasets cannot be used for end-to-end translation [12, 46, 91]. Two public datasets, RWTH-Phoenix-2014T and How2Sign, are the most used datasets in SLP. The former includes German sign language sentences that can be used for text-to-sign language translation. The later is a recently proposed multi-modal dataset used for speech-to-sign language translation. Though RWTH-PHOENIX-Weather 2014T [12] and How2Sign [21] provided the appropriate SLP evaluation benchmarks, they are not enough for the generalization of the SLP models. Furthermore, these datasets only include German and American sentences. Translating from the spoken language to a large diversity of sign languages is a major challenge for the Deaf community.

Applications: American Sign Language (ASL), as the largest sign language community in the World, is a most-used sign language in the developed applications for SLP. Since it may be hard for Deaf people to read or write the spoken language, they need some tools for communication with the other people in society. Furthermore, many interesting and useful applications in Internet are not accessible for the Deaf community. To tackle these challenges, some projects have been proposed aiming to develop such tools. While these applications successfully made a bridge between Deaf and hearing communities, we are still far from having applications involving a large vocabularies/sentences from complex real-world scenarios. One of the main challenges for these applications is a license right for usage. Another challenge is regarding the application domain. Most of these applications have been developed for very specific domains such as clinics, hospitals, and po-

Table 1. Summary of deep SLP models.

Year	Ref	Feature	Input modality	Dataset	Description
2011	[45]	Avatar	RGB video	ViSiCAST	**Pros.** Proposing a gloss-based tool focusing on the animation content evaluating using a new metric for comparing avatars with human signers. **Cons.** Need to include non-manual features of human signers.
2016	[55]	Avatar	RGB video	Own dataset	**Pros.** Automatically adding realism to the generated images, low computational complexity. **Cons.** Need to place the position of the shoulder and torso extension on the position of the avatar's elbow, rather than the IK end-effector.
2016	[31]	Avatar	RGB video	Own dataset	**Pros.** Easy to understand with high viewer acceptance of the sign avatars. **Cons.** Limited to the small set of sign phrases.
2018	[12]	NMT	RGB video	PHOENIX-Weather 2014T	**Pros.** Robust to jointly align, recognize, and translate sign videos. **Cons.** Need to align the signs in the spatial domain.
2018	[35]	NMT	RGB video	Own dataset	**Pros.** Robust to align the word order corresponding to visual content in sentences. **Cons.** Need to generalize to additional datasets.
2020	[79]	NMT, MG	Text	PHOENIX14T	**Pros.** Robust to minimal gloss and skeletal level annotations for model training. **Cons.** Model complexity is high.
2020	[73]	Others	Text	PHOENIX14	**Pros.** Robust to the dynamic length of output sign sequence. **Cons.** Model performance can be improved including non-manual information.
2020	[91]	Others	Text	Czech news	**Pros.** Robust to the missing skeletons parts. **Cons.** Model performance can be improved including information of facial expressions.
2020	[75]	Others	Text	PHOENIX14T	**Pros.** Robust to non-manual feature production. **Cons.** Need to increase the realism of the generated signs.
2020	[13]	Others	Text	PHOENIX14T	**Pros.** No need to the gloss information. **Cons.** Model complexity is high.
2020	[74]	Others	Text	PHOENIX14T	**Pros.** Robust to manual feature production. **Cons.** Need to increase the realism of the generated signs.

lice stations. Improving the amount of available data and its quality can benefit the creation of these needed applications.

Proposed models: The proposed works in SLP can be categorized into five categories: Avatar approaches, NMT approaches, MG approaches, Conditional image/video generation approaches, and other approaches. Table 1 shows a summary of state-of-the-art deep SLP models. Some samples of the generated videos and gloss annotations are shown in Figure 3, 4, and 5. Using the data collected from motion capture, avatars can be more usable and acceptable for reviewers. Avatars achieve highly realistic results but the results are restricted to a small set of phrases. This comes from the cost of the data collection and annotation. Furthermore, avatar data is not a scalable solution and needs expert knowledge to be inspected and polished. To cope with these problems and improve performance, deep learning-based models are used.

While NMT-based methods achieved significant results in translation tasks, some major challenges need to be solved. Domain adaptation is the first challenge in this area. Since the translation in different domains need different styles and requirements, it is a crucial requirement in developing machine translation systems targeted at a specific use case. The second challenge is regarding the amount of available training data. Especially in deep learning-based models, increasing the amount of data can lead to better results. Another challenge is regarding to the uncommon words. The translation models perform poorly on these words. Words alignment and adjusting the beam search parameters are other challenges in NMT-based models.

Although MG can generate plausible and controllable motion through a database of motion capture, it faces some challenges. The first challenge is regarding limited access to data. To show the model potential with a truly diverse set of actions, a large set of data is necessary. Scalability and computational complexity of the graph to select the best transitions are other challenges in MG. Furthermore, since the number of edges leaving a node increases with the graph size, the branching factor in the search algorithm will increase as well. To automatically adjusting the graph configuration and relying on the training data, instead of user interference, Graph Convolutional Network (GCN) could be used along with some refining algorithms to adopt the graph structure monotonically.

While GANs has recently achieved remarkable results for image/video generation, there exist major challenges in the training of GANs. Mode collapse, non-convergence and instability, suitable objective function, and optimization algorithm are some of these challenges. However, several suggestions have been recently proposed to address the better design and optimization of GANs. Appropriate design of network architecture, proper objective functions, and optimization algorithms are some of the proposed techniques to improve the performance of GAN-based models. Finally, the challenge of the model complexity still remains for hybrid models.

Limitations: In this survey, we presented recent advances in SLP and related areas using deep learning. While successful results have been achieved in SLP by recent deep learning-based models, there are some limitations that need to be addressed. A main challenge is regarding the Multi-Signer (MS) generation that is necessary for providing a real-world communication in the Deaf community. To this end, we need to produce multiple signers of different appearance and configuration. Another limitation is the possibility of high resolution and photo-realistic continuous sign language videos. Most of the proposed models in SLP can only generate low resolution sign samples. Conditioning on human keypoints extracted from training data can decrease the parameter complexity of the model and assist to produce a high resolution video sign. However, avatar-based models can successfully generate high resolution video samples, though they are complex and expensive. In addition, pruning algorithms of MG need to be improved by including additional features of sign language, such as duration and speed of motion.

Future directions: While recent models in SLP presented promising results relying on deep learning capabilities, there is still much room for improvement. Considering the discriminative power of self-attention, learning to fuse multiple input modalities to benefit from multi-channel information, learning structured spatio-temporal patterns (such as Graph Neural Networks models), and employing domain-specific prior knowledge on sign language are some possible future directions in this area. Furthermore, there are some exciting assistive technologies for Deaf and hearing-impaired people. A brief introduction to these technologies can get an insight to the researchers in SLP and also make a bridge between them and the corresponding technology requirements. These technologies fall into three device categories: hearing technology, alerting devices, and communication support technology. For example, let imagine a technology that assists a Deaf person go through a musical experience translated into another sensory modality. While the recent advances in SLP are promising, more endeavor is indispensable to provide a fast processing model in an uncontrolled environment considering rapid hand motions. It is clear that technology standardization and full interoperability among devices and platforms are prerequisites to having real-life communication between the hearing and hearing-impaired communities.

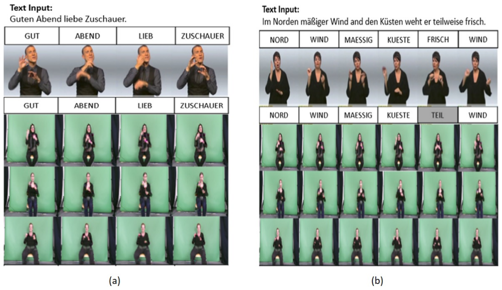
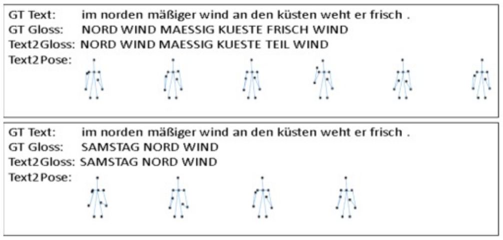


Figure 3. Translation results from [79]: (a) "Guten Abend liebe Zuschauer". (Good evening dear viewers), (b) "Im Norden maessiger Wind an den Kuesten weht er teilweise frisch". (Mild winds in the north, at the coast it blows fresh in parts). Top row: Ground truth gloss and video, Bottom row: Generated gloss and video. This model combines an NMT network and GAN for SLP.



GT Text: im norden mäßiger wind an den küsten weht er frisch .
GT Gloss: NORD WIND MAESSIG KUESTE FRISCH WIND
Text2Gloss: NORD WIND MAESSIG KUESTE TEIL WIND
Text2Pose:

GT Text: im norden mäßiger wind an den küsten weht er frisch .
GT Gloss: SAMSTAG NORD WIND
Text2Gloss: SAMSTAG NORD WIND
Text2Pose:

Figure 4. Translation results from [79]: Text from spoken language is translated to human pose sequences.

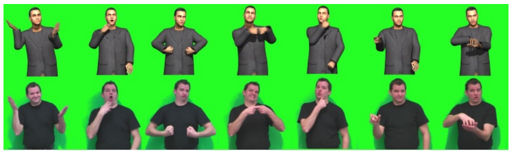


Figure 5. Translation results from [45]: A signing avatar is created using a character animation system. Top row: signing avatar, Bottom row: original video.

Acknowledgement This work has been partially supported by the HIS Company and Institute for Research in Fundamental Sciences (IPM) in Iran, Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE), and CERCA Programme/Generalitat de Catalunya, and ICREA under the ICREA Academia programme.

References

[1] Okan Arikan and D.A. Forsyth. Interactive motion generation from examples. *In Proceedings of the 29th annual conference on computer graphics and interactive techniques, SIGGRAPH '02*, pages 483–490, 2002.

[2] Start ASL. Chinese sign language. *https://www.startasl.com/chinese-sign-language/*, 2021.

[3] Start ASL. Spanish sign language. *https://www.startasl.com/spanish-sign-language-ssl/*, 2021.

[4] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. The american sign language lexicon video dataset. *CVPR*, pages 1–8, 2018.

[5] Daniel Bachmann, Frank Weichert, and Gerhard Rinkenauer. Review of three-dimensional human-computer interaction with focus on the leap motion controller. *Sensors*, 2018.

[6] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[7] J. A. Bangham, S. J. Cox, R. Elliott, J. R. W. Glauert, I. Marshall, S. Rankov, and M. Wells. Virtual signing: Capture, animation, storage and transmission – an overview of the visicast project. *Speech and Language Processing for Disabled and Elderly People*, 2000.

[8] Mark Borg and Kenneth P. Camilleri. Phonologically-meaningful sub-units for deep learning-based sign language recognition. *ECCV*, 2020.

[9] Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. The atis sign language corpus. *6th International Conference on Language Resources and Evaluation*, 2008.

[10] A. H. Butt, E. Rovini, C. Dolciotti, G. De Petris, P. Bongioanni, M. C. Carboncini, and F. Cavallo. Objective and automatic classification of parkinson disease with leap motion controller. *BioMed Eng OnLine*, 17, 2018.

[11] Su Cai, Gaoxia Zhu, Ying-Tien Wu, Enrui Liu, and Xiaoyi Hu. A case study of gesture-based games in enhancing the fine motor skills and recognition. *Interactive Learning Environments*, 26, 2018.

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

[13] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. *ECCVW*, 2020.

[14] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Rwth-phoenix-weather 2014 t: Parallel corpus of sign language video, gloss and translation. *CVPR, Salt Lake City, UT*, 2018.

[15] Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. Asl-lex: A lexical database for asl. *Behavior Research Methods*, 49:784–801, 2017.

[16] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. *ICCV*, pages 1511–1520, 2017.

[17] Miri Weiss Cohen, Israel Voldman, Daniele Regazzoni, and Andrea Vitali. Hand rehabilitation via gesture recognition using leap motion controller. *11th International Conference on Human System Interaction (HSI), Gdansk, Poland*, 2018.

[18] Stephen Cox, Mike Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. *In Proceedings of the 5th international ACM conference on assistive technologies*, pages 205–212, 2002.

[19] Khalid A. Darabkh, Farah H. Alturk, and Saadeh Z. Sweidan. Vrcdea-tcs: 3d virtual reality cooperative drawing educational application with textual chatting system. *Comput Appl Eng Educ*, 26:1677–1698, 2018.

[20] Felix Dawes, Jaques Penders, and Giuseppe Carbone. Remote control of a robotic hand using a leap sensor. *The international conference of IFToMM ITALY*, pages 332–341, 2018.

[21] Amanda Duarte. Cross-modal neural sign language translation. *The 27th ACM International Conference*, 2019.

[22] Amanda Duarte, Shruti Palaskar, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier GiroiNieto1. How2sign: A large-scale multimodal dataset for continuous american sign language. *Sign Language Recognition, Translation, and Production workshop*, 2020.

[23] Sarah Ebling and Matt Huenerfauth. Bridging the gap between sign language machine translation and sign language animation using sequence classification. *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 2–9, 2015.

[24] Sarah EblingJohn and GlauertJohn Glauert. Exploiting the full potential of jasigning to build an avatar signing train announcements. *In 3rd International symposium on sign language translation and avatar technology*, pages 1–9, 2013.

[25] Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. The dicta-sign wiki: Enabling web communication for the deaf. *International Conference on Computers for Handicapped Persons (ICCHP)*, pages 205–212, 2012.

[26] Ethnologue. Argentine sign language. *https://www.ethnologue.com/language/aed*, 2021.

[27] Ethnologue. Greek sign language. *https://www.ethnologue.com/language/gss*, 2021.

[28] Ethnologue. Persian sign language. *https://www.ethnologue.com/language/psc*, 2021.

[29] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC12), Istanbul, Turkey*, page 3785–3789, 2012.

[30] Sakher Ghanem, Christopher Conly, and Vassilis Athitsos. A survey on sign language recognition using smartphones. *Proceedings of the 10th international conference on pervasive technologies related to assistive environments, Island of Rhodes Greece*, 2017.

[31] Sylvie Gibet, François Lefebvre-Albaret, Ludovic Hamon, Rémi Brun, and Ahmed Turki. Interactive editing in french

sign language dedicated to virtual signers: Requirements and challenges. *Universal Access in the Information Society*, 15:525–539, 2016.

[32] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, page 2672–2680, 2014.

[33] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *Proceedings of Machine Learning Research*, 2015.

[34] Virtual Humans Group. Virtual humans research for sign language animation. *School of Computing Sciences, UEA Norwich, UK*, 2017.

[35] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. Hierarchical lstm for sign language translation. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[36] Thomas Hanke. German sign language. *https://www.awhamburg.de/en/research/long-term-scientific-projects/dictionary-german-sign-language.html*, 2021.

[37] Sepp Hochreiter and JüRgen Schmidhuber. Long short-term memory. *Neural Computation*, 9, 1997.

[38] John Hutchins. History of machine translation. *http://psychotransling.ucoz.com/-ld/0/11-Hutchins-survey.pdf*, 2005.

[39] Mohammed Jemni, Oussama El Ghoul, Mehrez Boulares, Nour Ben Yahia, Kabil Jaballah, Achraf Othman, and Monoem Youneb. Websign. *http://www.latice.rnu.tn/websign/*, 2020.

[40] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv:1610.10099*, 2016.

[41] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, 2019.

[42] Yuusuke Kataoka, Takashi Matsubara, and Kuniaki Uehara. Image generation using adversarial networks and attention mechanism. *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016.

[43] Richard Kennaway. Avatar-independent scripting for real-time gesture animation. *Procedural animation of sign language, arXiv:1502.02961*, 2013.

[44] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[45] Michael Kipp, Alexis Heloir, and Quan Nguyen. Sign language avatars: Animation and comprehensibility. *International Workshop on Intelligent Virtual Agents*, pages 113–126, 2011.

[46] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9, 2019.

[47] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 473–483, 2002.

[48] Agelos Kratimenos, Georgios Pavlakos, and Petros Maragos. 3d hands, face and body extraction for sign language recognition. *ECCV*, 2020.

[49] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 39–48, 1999.

[50] Xing Liang, Anastassia Angelopoulou, Epaminondas Kapetanios, Bencie Woll, Reda Al-batat, and Tyron Woolfe. A multi-modal machine learning approach and toolkit to automate recognition of early stages of dementia among british sign language users. *ECCV*, 2020.

[51] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *arXiv:1508.04025*, 2015.

[52] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *NIPS*, 2017.

[53] Nezam Majidi, Kourosh Kiani, and Razieh Rastgoo. A deep model for super-resolution enhancement from a single image. *Journal of AI and Data Mining*, 8:451–460, 2020.

[54] Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. Dicta-sign–building a multilingual sign language corpus. *In 5th LREC. Istanbul*, 2012.

[55] John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. An automated technique for realtime production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15:551–566, 2016.

[56] Long Short-Term Memory. Sepp hochreiter and jürgen schmidhuber. *Neural Computation*, 1997.

[57] Matteo Morando, Serena Ponte, Elisa Ferrara, and Silvana Dellepiane. Definition of motion and biophysical indicators for home-based rehabilitation through serious games. *Information*, 9, 2018.

[58] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics and Automation Magazine*, 19:98–100, 2012.

[59] Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. Real-time sign language detection using human pose estimation. *ECCV*, 2020.

[60] Achraf Othman and Mohamed Jemni. Statistical sign language machine translation: from english written text to american sign language gloss. *IJCSI International Journal of Computer Science*, 8:65–73, 2011.

[61] Owlcation. Korean sign language. *https://owlcation.com/humanities/Korean-Sign-Language*, 2021.

[62] Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. Exploiting 3d hand pose estimation in deep learning-based sign language recognition from rgb videos. *ECCV*, 2020.

[63] Siegmund Prillwitz. Hamnosys. version 2.0. hamburg notation system for sign languages. an introductory guide. *Hamburg Signum Press*, 1989.

[64] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Multimodal deep hand sign language recognition in still images using restricted boltzmann machine. *Entropy*, 20, 2018.

[65] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Hand sign language recognition using multi-view hand skeleton. *Expert Systems With Applications*, 150, 2020.

[66] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Video-based isolated hand sign language recognition using a deep cascaded model. *Multimedia Tools And Applications*, 79:22965–22987, 2020.

[67] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Hand pose aware multimodal isolated sign language recognition. *Multimedia Tools And Applications*, 80:127–163, 2021.

[68] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Real-time isolated hand sign language recognition using deep networks and svd. *Journal of Ambient Intelligence and Humanized Computing*, 2021.

[69] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems With Application*, 164:113794, 2021.

[70] Marco Roccetti, Gustavo Marfia, and Angelo Semeraro. Playing into the wild: A gesture-based interface for gaming in public spaces. *Journal of Visual Communication and Image Representation*, 23:426–440, 2012.

[71] Salto. Polish sign language. *https://www.salto-youth.net/tools/otlas-partner-finding/organisation/association-of-polish-sign-language-interpreters.2561/*, 2021.

[72] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Adversarial training for multi-channel sign language production. *BMVC*, 2020.

[73] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. *ECCV*, pages 687–705, 2020.

[74] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Adversarial training for multi-channel sign language production. *BMVC*, 2020.

[75] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv:2011.09846*, 2020.

[76] Abigail See and Matthew Lamm. Machine translation, sequence-to-sequence and attention. *https://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf*, 2021.

[77] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. *CVPR*, 2018.

[78] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. *BMVC*, 2018.

[79] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128:891–908, 2020.

[80] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *NIPS*, 2014.

[81] Sandrine Tornay, Necati Cihan Camgoz, Richard Bowden, and Magimai Doss. A phonology-based approach for isolated sign production assessment in sign language. *ICMI '20 Companion: Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020.

[82] Aurelijus Vaitkevičius, Mantas Taroza, Tomas Blažauskas, Robertas Damaševičius, Rytis Maskeliūnas, and Marcin Woźniak. Recognition of american sign language gestures in a virtual reality using leap motion. *Appl. Sci.*, 9, 2019.

[83] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *Proceedings of The 33rd International Conference on Machine Learning*, pages 1747–1756, 2016.

[84] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *NIPS*, 2016.

[85] Neel Vasani, Pratik Autee an, and Samip Kalyani; Ruhina Karani. Generation of indian sign language by sentence processing and generative adversarial networks. *International Conference on Intelligent Sustainable Systems (ICISS)*, 2020.

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

[87] William G. Vicars. American sign language. *http://www.lifeprint.com/*, 2021.

[88] WHO: World Health Organization. Deafness and hearing loss. *http://www.who.int/mediacentre/factsheets/fs300/en/*, 2021.

[89] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. *ECCV*, page 776–791, 2016.

[90] Hee-Deok Yang. Sign language recognition with the kinect sensor based on conditional random fields. *Sensors*, 15:135–147, 2014.

[91] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. *WACV*, pages 3395–3403, 2020.

[92] Inge Zwitserlood, Margriet Verlinden, Johan Ros, and Sanny van der Schoot. Synthetic signing for the deaf: esign. *http://www.visicast.cmp.uea.ac.uk*, pages 1–6, 2005.