# MyVoice: Continuous End-to-End Sign Language to Text Translation

**Nitin Pillai**
nitin.pillai@berkeley.edu

**Manohar Madhira**
rmadhira@berkeley.edu

**Isabel Garcia Pietri**
isabelgarpietri@berkeley.edu

**Riyaz Kasmani**
rkasmani@berkeley.edu

* Authors are listed in random order

## Abstract

Previous research on Sign Language Translation was mainly involved with techniques such as Action Recognition, translating word level sign language data and recent work on continuous sign language translation (CSLT) focused on domain specific datasets. Not much work has gone into addressing CSLT on broad domain datasets that closely model real world communication.

This paper introduces CSLT done on a Multimodal American Sign Language (ASL) dataset: How2Sign which consists of more than 80 hours of ASL videos. We base our work on previous state of the art on Sign Language Translation using Transformers for jointly learning both sign language recognition and translation tasks. Since this approach depends on Glosses which is an intermediary representation and has proven to increase model performance, we generate glosses due to the absence of them in the How2Sign dataset.

We evaluate the translation performance of our model on the How2Sign dataset and compare it with the previous state of the art. Our model outperforms the previous state of the art by more than 50% on BLEU-4 scores. As part of our work, we provided enhancements to the How2Sign dataset with gloss generation, sentence alignment, manual cleaning of the videos along with making all our code repository for modeling and experimentation public on Github.

## 1 Introduction

During the recent pandemic, the deaf and hard of hearing (DHH) often found themselves excluded. Information broadcast on television was not always made available in sign language. And at schools around the world, remote learning alternatives failed to meet the needs of children who use sign language leaving them feeling isolated, excluded, and frustrated. Even without a crisis, the hearing impaired face daily issues of isolation and miscommunication (1). Just like the rest of us, the DHH have a right to quality education, healthcare and an environment that maximizes their potential. This paper is motivated to further the research in continuous sign language technologies that can bring deaf people a step closer to the realities of day-to-day communication.

According to the World Health Organization (2), over 5% of the worlds population (430 million people) have a 'disabling' hearing loss. And that number is expected to increase to 700 million by 2050. The prevalence of hearing loss increases with age, among those older than 60 years, over 25% are affected by disabling hearing loss. A person who is not able to hear as well as someone with normal hearing – hearing thresholds of 20 dB or better in both ears – is said to have hearing loss. 'Disabling' hearing loss refers to hearing loss greater than 35 decibels (dB) in the better hearing ear. People who are hard of hearing usually communicate through spoken language and can benefit from hearing aids, cochlear implants, and other assistive devices as well as captioning. 'Deaf' people mostly have very little or no hearing and often use sign language for communication.

Sign language is a natural visual-spatial language. It typically uses manual (hand shape, palm orientation, etc.) and non-manual markers

(mouthing, eye gazes, facial expressions, etc.) (3). Contrary to popular belief, at least 200 sign languages exist worldwide (4). Each with its own unique linguistic similarities, such as grammar and semantics, and is different from the spoken language of the region it belongs to. There is no one to one mapping between sign and spoken language translations and this makes the task very complex. Additionally, there is Glossing (5) which is a written transcription (not translation) form of sign language. And while glossing can accurately capture the signing process, there is no equivalent structure in spoken languages.

The National Center for Health Statistics estimates 28 million Americans have some degree of hearing loss (6). About 2 million of these 28 million people are classified as deaf (they can't hear every day sounds or speech even with a hearing aid). American Sign Language (ASL) is the natural language of around 500,000 deaf people in the US and Canada and it is used in 20 other countries.

In this paper, we adapt the Transformer-based model introduced by Camgoz et al. (7) and apply it to the How2Sign dataset (8). Camgoz et al's approach outperformed previous state-of-the-art results for continuous sign language recognition and translation tasks as applied on the Phoenix14T dataset (9). The Camgoz et al. model was also leveraged by Cabot et al. (10) on the How2Sign dataset and provides a baseline set of results that we worked to improve upon further. Our main contributions for this paper can be summarized as follows:

1. Enhance the How2Sign dataset with gloss generation, text reviews, sentence alignment, and manual cleaning of videos.

2. Adapt the state-of-the art Sign Language Transformer model to How2Sign and improve on Cabot et al. baseline results.

3. Make all our code for all the experiments we did public on Github: https://github.com/sign2text/myvoice

The rest of this paper is structured as below: In section 2, we summarize the work done by Camgoz et al., Cabot et al. and other related work by authors in this field of research. In section 3, we introduce the How2Sign dataset. In section 4, we explain the model architecture and our methodology. In section 5, we discuss our results and analyze the data quality. In Section 6, we conclude with explaining limitations we faced and present future lines of work.

## 2 Related work

Most of the machine learning projects in this domain start with either finger spelling, word, or continuous sign language as their source elements. The projects fall under two main categories: Sign Language Recognition (SLR) and Sign Language Translation (SLT). The key difference between the two being Recognition (which is the more difficult task) is focused on understanding what information is being conveyed and Translation taking that embedded information and generating spoken sentences.

Figure 1 (11) represents tasks of converting between one data representation to another. The ones on the left hand side represent computer vision tasks, are inherently language-agnostic and generalized between sign languages. The ones on the right are in the realm of natural language processing. They are sign language and spoken language specific. In total there's about 20 tasks with various amounts of research.

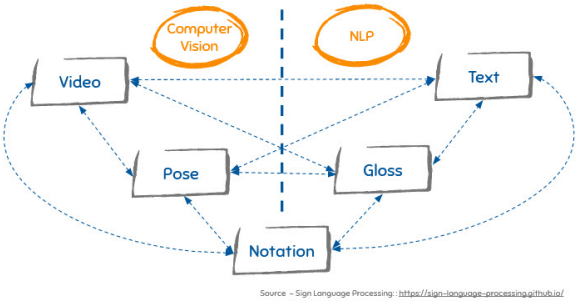


Figure 1: Sign Language Processing Paths (11))

Sign Language Recognition (SLR) - is concerned with extracting the sequence of sign glosses

from a video of someone performing a continuous sign. Most recent work has focused on leveraging Convolutional Neural Networks (CNN) on continuous data to effectively model spatio-temporal representations from sign language (12) (13). Some have focused on breaking the task into sub tasks such as alignment learning, single-gloss, and sequence construction (18) (19).

Sign Language Translation (SLT) - extracts spoken language from a video of someone performing a continuous sign. Camgoz et al. approached this task as a spatio-temporal neural machine translation (NMT) project. Using CNN's in combination with NMT, they extracted gloss-level features from video and used sequence-to-sequence attention based NMT to perform continuous German Sign Language translation. Following this Cabot et al. leveraged the same approach on How2Sign creating a baseline for American Sign Language. However, given the inherent difficulty and complexity of this task, their CSLR/CSLT performance was sub-optimal, with high Word Error Rates (WER) and low BLEU-4 scores. Work has also been done by Ko et al. (17), where they propose a similar approach but use body key-point coordinates as input, and evaluate their method on a Korean Sign Language dataset.

Datasets - some word and phrase level datasets exist publicly and have been used by various authors for their works. Key among them is the PHOENIX-Weather 2014T (9). This dataset is extracted from weather forecast reports on a German TV Station PHOENIX. It consists of a parallel corpus of German sign language videos, gloss-level annotations, and a vocabulary of 2,887 unique words and 1,066 different signs. The dataset is split into 7096 training, 519 development, and 642 test pairs. Another is ASLG-PC12 (21) which uses a rule-based approach on English data of Project Gutenberg to construct American Sign Language (ASL) glosses. This data does not contain sign language videos and contains 87,710 training pairs.

## 3 Data

In this study we use the How2Sign dataset (8). This dataset consists of more than 80 hours of sign language videos and their corresponding English transcripts. The dataset originally contains about 35k examples between training, development and test data. To reduce the cycle time, in this study we carry out some experiments using a subset of the dataset. The subset of the dataset includes only the examples of sentences with 10 words or less, which consists of approximately 11k examples. Table 1 contains the statistics of the data.

## 4 Methodology

### 4.1 Re-alignment of the data

The How2Sign sign language videos for the sentences are not well aligned with the corresponding sentence's transcripts. Hence, as a first step of the data preparation process, we realigned the videos with their corresponding transcripts.

### 4.2 Crop and resize videos

The How2Sign sign language videos are recorded at 1280x720 resolution. To reduce memory requirements, all the videos are cropped to include only the area where the person is performing sign language, and converted to 224x224 resolution.

### 4.3 Fix transcripts

We resolved all contractions in the text of the video transcripts.

### 4.4 Model Architecture

The model architecture we use is an encoder-decoder architecture introduced by Camgoz et al. (7). Figure 2 shows an overview of the architecture, which is designed to generate written translations from the sign language videos, using an intermediate gloss supervision.

| | All examples | All vocab | Subset examples | Subset vocab |
|---|---|---|---|---|
| Train | 31,165 | 15,097 | 10,027 | 6,133 |
| Development | 1,741 | 3,177 | 554 | 928 |
| Test | 2,357 | 3,631 | 776 | 1,037 |

Table 1: How2Sign dataset statistics.

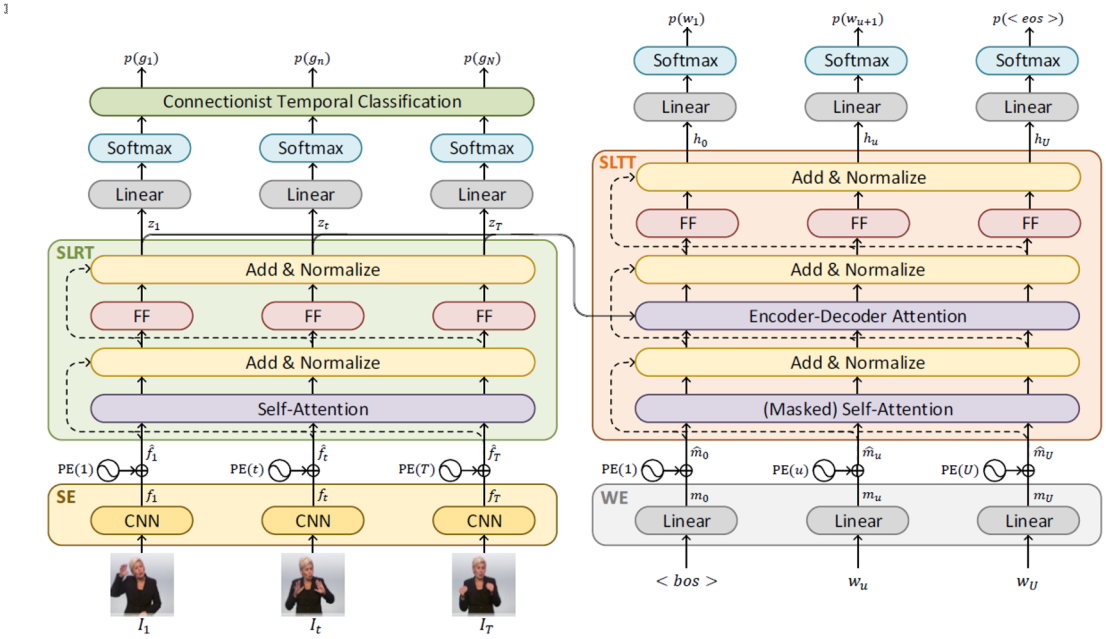


Figure 2: Model architecture.

On the left side of the image is the encoder model: Sign Language Recognition Transformer (SLRT). The aim of SLRT is to recognize glosses from continuous sign language videos while learning meaningful spatio-temporal representations for the end goal of sign language translation. On the right side of the image is the decoder model: Sign Language Translation Transformer (SLTT). This is an autoregressive transformer decoder model, which is trained to predict one word at a time to generate the corresponding spoken language sentence. This architecture jointly learns Continuous Sign Language Recognition and Translation. This is achieved by using a Connectionist Temporal Classification (CTC) loss to bind the recognition and translation problems into a single unified architecture.

During the study we performed sensitivities with the number of heads and number of layers of the architecture.

### 4.5 Feature generation

The architecture described above does not consume raw sign videos directly, it consumes features extracted from the videos, referenced in the architecture as Spatial Embeddings (SE). To extract the features from the videos we used various approaches:

- **Inflated 3D Convolutional Neural Network (I3D)(26)**: this is a popular architecture to extract features from videos. This network was originally trained with the Kinetics dataset, which contains annotated videos of human actions. We used two slightly different implementations to extract features from this architecture: an implementation from the GLUON library (22), which generates a compressed version of the features with dimensions (1,2048), and another I3D library that generates uncompressed features with dimensions (x,1024), where x depends on the video

length.

- **InceptionV3(27)**: this is a convolutional neural network for assisting in image analysis and object detection. This network was originally trained with the Imagenet dataset. As this network is trained with images and not with videos, we extracted one feature per video frame, resulting in features with dimensions (x,2048), where x is the number of frames in a video. This way of extracting features is compute-intensive and was used only on the subset of the How2Sign dataset.

### 4.6 Glosses

Sign language datasets generally consist of sign language videos, sign language notational system (glosses) and corresponding written spoken language. Machine translation for sign language models use these three data formats to tackle the problem of sign language translation. Since the model architecture requires glosses and the How2Sign dataset does not provide glosses, we used the English transcripts as glosses.

We also explored the idea of generating synthetic glosses from the English transcripts. For this, we used a Statistical Machine Translation for Sign Language (ASL-SMT) model(23) trained with a parallel corpus of written English text to American Sign Language text (glosses). This parallel corpus has about 5k examples taken from Congress sessions transcripts.

### 4.7 Learning Rate Schedulers

One of the important factors in optimization using gradient-based methods is a hyperparameter called the learning-rate. Learning rate is the amount of change to model weights during backpropogation process. This is used by the optimizer to reach the minima of the loss function. If the learning rate is large, the algorithm learns fast, but makes large updates, or may skip the minima. If the learning rate is small, the model will slowly move towards the minima, but if it is too small it may get stuck in local minima. To find the optimum learning rate, many predefined frameworks or schedulers exist that start with higher learning rate and gradually decrease the rate as training progress. (24) demonstrated the benefits of cyclical learning rates. (20) showed that increasing learning rate during warmup improves learning for large batch sizes. We tested our sample dataset with four learning rate schedulers: **plateau** reduces learning rate in steps, **exponential** reduces learning rate exponentially, **cosine** uses a cyclical learning rate schedule and **noam** introduces an increase in learning rate during warmup. Figure 3 shows the translation losses over the training cycles for each of these schedulers. While plateau completed training quickly, the model didn't learn much. Both noam and cosine schedulers reduced translation errors with more learning, with cosine scheduler giving slightly better results.

## 5 Results

We followed a multi step journey in building our model going through several experiments in achieving our final results as outlined below:

### 5.1 Build a Model

Our model based on the Sign Language Recognition and Translation transformer is explained in the Model Architecture section above. Table 2 shows the BLEU scores we achieved on our baseline model and also the best results achieved by model optimizations done via hyperparameter tuning. As can be seen, just by doing model optimizations we were able to increase our model performance by 2x on the BLEU-4 score and get very close to the State of the Art published so far on the How2Sign dataset.

### 5.2 Adding generated Glosses

Research in Sign Language Translation has shown that Glosses which are an intermediary representation (while going from Sign Language to translated sentences) helps with model performance. Since

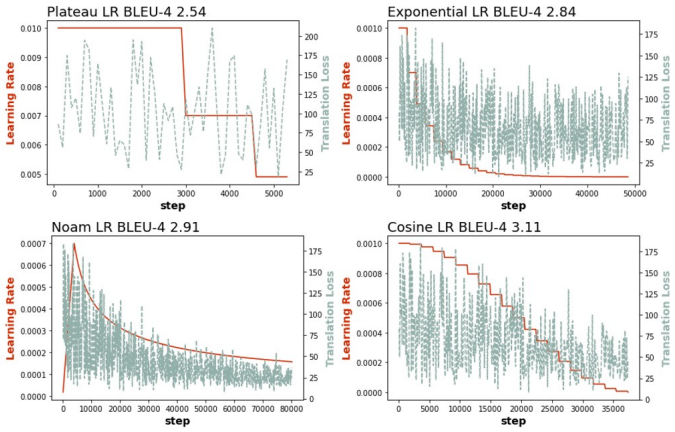


Figure 3: LR Schedules

| Test Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| SOTA (Patricia Cabot et al) | 17.40 | 7.69 | 3.97 | 2.21 |
| Our Baseline | 19.72 | 5.76 | 2.30 | 1.02 |
| Results - Optimizations | 20.96 | 7.56 | 3.96 | 2.10 |

Table 2: Results from Model Optimization

the How2Sign ASL dataset does not contain glosses, we found a way to generate glosses using a statistical machine translator which was trained on a corpus from English Congress sessions to translate English text to ASL glosses.

Shown below is an English transcript and the Glosses generated using the text-to-gloss model we used :

**English Transcript**: now together you are going to go opposite
**Glosses**: NOW DESC-TOGER X-YOU BE GOING TO GO OPPOSITE

As can be seen in Table 3, model performance on BLEU-4 increases by 35% which proves the value of having Glosses as an intermediary representation for sign language translation models.

### 5.3 Adding more data by experimenting on the full dataset

The model was initially trained on a subset of the entire dataset and we only included sentences with 10 words or lesser. The results obtained before reflected that. We now wanted to experiment with model performance by training on the entire dataset. The training took 7x more time but the results only improved marginally across all BLEU scores with a 7% increase on the BLEU 4 score as can be seen from Table 4.

### 5.4 Experiment with varying Learning Rate Schedulers

At this stage of our experimentation, since we already had the results from training on the entire dataset, we wanted to try by changing the learning rate scheduler to see if the model could learn better instead of getting stuck in some local maxima. We analyzed four learning rate schedulers : Linear, Exponential, Noam and Cosine. Cosine had the best BLEU-4 scores overall. The results of this experiment are shown in Table 5.

### 5.5 Model performance on each individual Signer

Since the BLEU scores were not increasing significantly across all our experiments, we wanted

| Test Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| Best case without Glosses | 20.96 | 7.56 | 3.96 | 2.10 |
| Best case with Glosses | 22.78 | 8.36 | 4.48 | 2.83 |

Table 3: Results from addition of generated Glosses

| Test Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| Best case on subset of data | 22.78 | 8.36 | 4.48 | 2.83 |
| Best case on entire dataset | 18.58 | 8.49 | 4.89 | 3.03 |

Table 4: Results from training on entire dataset

to pay attention to the underlying data quality of the How2Sign dataset. To verify this, we started with first validating the signer quality of each individual signer. We did this by creating separate train/dev/test datasets for each individual signer and then training our model on each signer's datasets. This was mainly done to validate our hypothesis of signer quality affecting final model scores. But as can be seen from Table 6, the BLEU scores didn't vary much across the different signers and in fact, the model performed best when we trained on the entire dataset containing all the signers as the variability between the signers helped the model learn better.

### 5.6 Study on Data Quality by native ASL Signers

During the study, we noticed that for a number of videos, the sign language in the videos did not correspond with their transcripts (even after the latest realignment provided by the authors of the dataset). And, given that it was not possible to improve the BLEU-4 score above 3-4, which is considered a very low score, we decided to carry out an analysis of the general quality of the dataset.

We approached 3 ASL Native speakers to validate 4 randomly selected examples in the dataset (before and after realignment). We asked them to rate each video on how well it matched its transcript (% match). Results are presented in table 7. In general, realigned videos matched their transcripts better. However, only 2 out of 8 realigned videos had 90+% match. ASL Native speakers also reported that some signers were using Signed Exact English (SEE) and not American Sign Language.

## 6 Conclusion and Future Work

In order to successfully translate a continuous sign language video into a sentence, it is essential to first do some form of sign language recognition to understand the actions performed via an intermediate representation that can later be used in translating those actions into a fully understandable sentence. Previous research in Sign Language Translation heavily focused on action recognition techniques as the first step. In this paper, we base our research on Camgoz et al's work on using Transformers for jointly learning both Sign Language Recognition and Translation tasks. We described the process of adapting this transformer model for our setup of training the model with a Multimodal American Sign Language (ASL) dataset : How2Sign. Since the previous state of the art model depended on having Glosses as an intermediary representation while going from Sign Language video to sentence, we worked on generating glosses by tweaking a statistical machine translator that was trained on a corpus of congressional sessions. This helped in increasing model performance but not by much due to the quality of the generated glosses. The improvements and optimizations done here helped in surpassing the BLEU-4 scores achieved by the previous state of the art model by more than 50% establishing a new baseline for the state of the art

| Test Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| Best case on entire dataset | 18.58 | 8.49 | 4.89 | 3.03 |
| Best case on entire dataset | 20.11 | 9.30 | 5.39 | 3.34 |

Table 5: Results from varying Learning Rate Schedulers

| Test Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| Best case Cosine scheduler | 20.11 | 9.30 | 5.39 | 3.34 |
| Signer 1 | 16.92 | 7.00 | 3.98 | 2.47 |
| Signer 2 | 20.78 | 9.13 | 4.95 | 2.78 |
| Signer 3 | 17.96 | 7.27 | 3.90 | 2.23 |
| Signer 5 | 19.87 | 8.20 | 4.44 | 2.58 |
| Signer 8 | 17.96 | 7.99 | 4.31 | 2.45 |

Table 6: Results from varying signers

in Continuous Sign Language Translation on the How2Sign ASL dataset.

In the future, we would like to expand on our previous work by focusing on generating high quality glosses both via manual expert annotation by help of ASL signers and via automatic gloss generation by using models trained on corpuses belonging to a wide variety of domains. We will also like to explore other non publicly available datasets like the BBC Oxford British Sign Language dataset which has a corpus of videos across a range of domains. We would also want to explore other methods of extracting features from ASL videos by using a Video SWIN Transformer trained on sign language videos itself as we believe this will help in increasing the translation performance. After doing these suggested changes, we would want to spend more time on Model hyperparameter tuning to explore a wider range of optimal parameters to increase model performance.

Lastly, we would like to experiment on modelling individual sign articulators of the person performing sign language such as the person's facial expressions along with mouth movements, hand and body movements to see if our deep learning networks can learn any relationships between the different body parts that can help the model overall learn better.

## References


[1] Maria Fernanda Neves Silveira de Souza, Amanda Miranda Brito Ara´ujo, Luiza Fernandes Fonseca Sandes,Daniel Antunes Freitas, Wellington Danilo Soares, Raquel Schwenck de Mello Vianna, and Arlen Almeida Duarte de Sousa. (2017). Main difficulties and obstacles faced by the deaf community in health access: an integrative literature review. *Revista CEFAC, 19:395 – 405.*

[2] Deafness and Hearing Loss. (2021, April 1). World Health Organization. Retrieved July 16, 2022, from https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[3] Penny Boyes-Braem and Rachel Sutton-Spence. The Hands are the Head of the Mouth: The Mouth as Articulator in Sign Languages. *Gallaudet University Press, 2001.*

[4] Sign language. (2022, July 16). Ethnologue. Retrieved July 16, 2022, from https://www.ethnologue.com/subgroups/sign-language

[5] Gloss. (n.d.). American Sign Language University (ASLU). Retrieved July 16, 2022, from https://www.lifeprint.com/asl101/topics/gloss.htm

[6] National Center for Health Statistics. (2021, March 25). Quick Statistics About Hearing. National Institute on Deafness and Other Communication Disorders (NIDCD). Retrieved July 16, 2022, from https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing

[7] Camgoz, N., Koller, O., Hadfield, S., Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

| Match | Realigned | Original |
|---|---|---|
| 90+% | 2 | 0 |
| 60-90% | 1 | 0 |
| 30-60% | 4 | 3 |
| 1-30% | 0 | 2 |
| 0% | 1 | 3 |

Table 7: Match of videos and transcripts.


[8] Duarte, A., Palaskar,S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., & Giro-i-Nieto, X. (2020). How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language.*https://how2sign.github.io/*

[9] Forster J., Schmidt C., Hoyoux T., Koller O., Zelle U., Piater J.H. , and Ney H. (2012). RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC)*

[10] Alvarez P.C., Nieto X.G., and Benet L.T. (2022). Sign Language Translation based on Transformers for the How2Sign Dataset. *cabot2022SignLT*

[11] Moryossef A. and Goldberg Y. (2021). Sign Language Processing. *Retrieved July 16, 2022, from https://sign-language-processing.github.io/*

[12] Cui R., Liu H., and Zhang C. (2017). Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[13] Molchanov P., Yang X., Gupta S., Kim K., Tyree S., and Kautz J. (2016). Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*

[14] Koller O., Zargaran S., and Ney H. (2017). Re-sign: Re-aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMS. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[15] Jihai Zhang, Wengang Zhou, and Houqiang Li. 2014. A threshold-based hmm-dtw approach for continuous sign language recognition. ACM International Conference Proceeding Series, 07

[16] Zhang J., Wengang Z., and Houqiang L. (2014). A Threshold-based HMM-DTW Approach for Continuous Sign Language Recognition. *ACM International Conference Proceeding Series.*

[17] Ko S-K, Kim CJ, Jung H, and Cho C. (2019). Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences. 2019*

[18] Jihai Zhang, Wengang Zhou, and Houqiang Li. 2014. A threshold-based hmm-dtw approach for continuous sign language recognition. ACM International Conference Proceeding Series, 07

[19] Zhang J., Wengang Z., and Houqiang L. (2014). A Threshold-based HMM-DTW Approach for Continuous Sign Language Recognition. *ACM International Conference Proceeding Series.*

[20] Huo, Z., Gu, B., Huang, H. (2021). Large Batch Optimization for Deep Learning Using New Complete Layer-Wise Adaptive Rate Scaling. *Proceedings of the AAAI Conference on Artificial Intelligence 2021*

[21] Othman, A., and Tmar, Z. (2012). English-ASL Gloss Parallel Corpus 2012: ASLG-PC12, The Second Release. *Fourth International Conference On Information and Communication Technology and Accessibility ICTA'13, Hammamet, Tunisia, October 24-26, 2013*

[22] Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., Shi, X., Wang, C., Xie, J., Zha, S., Zhang, A., Zhang, H., Zhang, Z., Zhang, Z., Zheng, S., & Zhu, Y. (2020). GluonCV and GluonNLP: Deep Learning in Computer Vision and Natural Language Processing. *Journal of Machine Learning Research, Volume 21, Number 23, Pages 1-7.*

[23] Othman, A., & Jemni, M. (2019). Designing High Accuracy Statistical Machine Translation for Sign Language Using Parallel Corpus—Case study English and American Sign Language. *Journal of Information Technology Research, Volume 12, Issue 2.*

[24] Leslie N. Smith (2017). Cyclical Learning Rates for Training Neural Networks. *Presented at WACV 2017*

[25] Camgoz, N., Hadfield, S., Koller, O., Ney H., and Bowden, R. (2018). Neural Sign Language Translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*

[26] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the Tenth Workshop on Statistical Machine Translation.*

[27] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *CoRR, Volume 1512.00567.*