Towards Automatic Speech to Sign Language Generation

Parul Kapoor[1], Rudrabha Mukhopadhyay[2], Sindhu B Hegde[2], Vinay Namboodiri[1,3], C V Jawahar[2]

[1]Department of Computer Science and Engineering, IIT Kanpur
[2]Center for Visual Information Technology, IIIT Hyderabad
[3]Department of Computer Science, University of Bath

kparul@iitk.ac.in, {radrabha.m, sindhu.hegde}@research.iiit.ac.in, vpn22@bath.ac.uk,
jawahar@iiit.ac.in

Abstract

We aim to solve the highly challenging task of generating continuous sign language videos solely from speech segments for the first time. Recent efforts in this space have focused on generating such videos from human-annotated text transcripts without considering other modalities. However, replacing speech with sign language proves to be a practical solution while communicating with people suffering from hearing loss. Therefore, we eliminate the need of using text as input and design techniques that work for more natural, continuous, freely uttered speech covering an extensive vocabulary. Since the current datasets are inadequate for generating sign language directly from speech, we collect and release the first Indian sign language dataset comprising speech-level annotations, text transcripts, and the corresponding sign-language videos. Next, we propose a multi-tasking transformer network trained to generate signer's poses from speech segments. With speech-to-text as an auxiliary task and an additional cross-modal discriminator, our model learns to generate continuous sign pose sequences in an end-to-end manner. Extensive experiments and comparisons with other baselines demonstrate the effectiveness of our approach. We also conduct additional ablation studies to analyze the effect of different modules of our network. A demo video containing several results is attached to the supplementary material.

Index Terms: Sign Language Generation, Speech to Sign Language, Speech Recognition, Human-Computer Interaction, Computational Paralinguistics

1. Introduction

According to a World Health Organization report [1], over 466 million people worldwide, or roughly 5% of the world's population, suffer from hearing loss. Sign language is often the primary means of communication used by people with hearing disabilities. Sign language consists of manual communication features including hand gestures, hand-shape, location, movement, orientation. It includes non-manual gestures like eye gaze, eyebrows, and mouth movement.

Sign-language is popularly represented in human-readable form using glosses and text transcripts. However, annotating at the gloss level is a tedious task and limits datasets to a smaller size. While efforts [2, 3] have been focused on using text as the input modality, this is inherently limiting. Generally, sign-language replaces speech as a communication medium for people suffering from hearing loss. Thus, apart from the content, the speaker's emotions and other attributes are also communicated via sign language. This additional stream of information is completely lost when converting text to corresponding sign language videos. Text transcripts are also annotated at the sen-
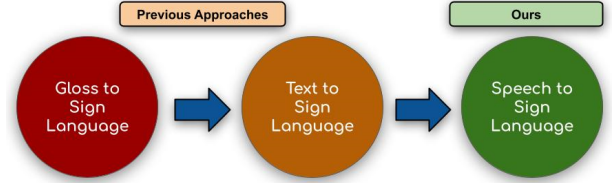

Figure 1: *Previous approaches have only attempted to generate sign-language from text transcripts, we focus on directly converting speech segments into sign-language. Our work opens up several assistive technology applications and can help effectively communicate with people suffering from hearing loss.*

tence level, limiting the amount of higher-level context information provided for better continuous sign language generation. Finally, most applications like interviews and news reporting require direct translation from continuous speech to sign language that cannot be segmented into sentences.

In this work, we focus on overcoming the current setup's shortcomings and directly translate natural speech to the corresponding sign language poses for the first time. This allows us to generate improved sign-language gestures by inherently transferring content and style from the speech. To generate sign language videos from speech, we collect and release the first large-scale dataset tailor-made for our task. The dataset contains videos of a signer signing in Indian Sign Language standards [4] accompanied by corresponding speech segments. Each speech segment also contains the corresponding text transcripts. We then train a multi-task transformer network on the collected data coupled with a cross-modal discriminator to generate intelligible sign-language videos from the given speech segment. The proposed dataset and code will be released for future research at https://kapoorparul.github.io/S2SL/

2. Text to Sign Language Generation

Sign-language generation from text has always been an essential part of assistive technologies. Initial works [5–7] for this problem depended on classical grammar-based approaches to stitch together signs for isolated words to achieve continuous sign language production. Like in many other cases, deep-learning has had a massive impact on modern techniques. Researchers considered sign-language as another "language" and modeled converting the text in language A to "sign-language" as a Neural Machine Translation (NMT) problem. This approach was used in [8, 9] with varying degrees of success. Zelinka et al. [3] proposed generating Czech sign language from words rather than glosses but they restrict the length of signs to 7 frames for each word and work with a very limited vocabu-

lary of 598 Czech lemmas. Recent methods [2, 10] use transformer networks to learn long-term dependencies between the input and output modalities. Most of these works provide text tokens as input and generate a sequence of human poses as the signed output.

Even though these works show decent improvements in the generation of sign-language from text, there has been no push to include speech as the input. Some previous works [11–13] propose speech to sign language but rely on intermediate automatic speech recognition step and finally use rule based approaches to convert text to sign language videos for a very restrictive vocabulary. Despite bringing an additional stream of information while communicating, speech as a modality brings a lot of challenges to the current setting. For example, unlike text, speech cannot be segmented into discrete units like words or glosses. The lack of datasets containing speech-level annotations also complicates the task. To help solve this problem, we collect a dataset amenable to our task. We then propose a multi-task transformer-based architecture for learning high-level contextual information between speech and the signer's poses. We also use a cross-modal discriminator to classify whether the generated poses match the input speech. In Section 3, we discuss our dataset collection procedure and the proposed network in detail.

3. Speech To Sign Language Generation

In this work, we make the first-ever attempt to generate sign-language directly from the speech input by proposing a fully end-to-end model. Formally, given a speech input sequence S as $\{s_1, s_2, .., s_n\}$, our aim is to design a model that will generate a human pose key point sequence P as $\{p_1, p_1, .., p_m\}$ corresponding to the sign language. Here p_i corresponds to the pose key points for the ith frame in the sign language sequence. For each input speech segment S, the corresponding text is also transcribed as $T = \{t_1, t_2, .., t_Q\}$. Due to the unavailability of datasets containing speech level annotation, we curate the first continuous sign language dataset adhering to requirements. Our sign language dataset is based on Indian Sign Language (ISL) standards as it is curated from Indian news reports that a signer annotates in real-time.

3.1. Indian Sign Language Dataset

Over time, several datasets have been proposed for American [14], German [15], Korean [16], and Chinese [17] sign languages. Since these datasets do not contain the required speech annotations necessary for our task, we collect and release a large scale corpus containing all three modalities: (i) continuous sign-language videos, (ii) text and (iii) speech. Our dataset is based on ISL standards [4] and is far more extensive than the current ISL datasets [18, 19].

Our proposed corpus spans a vast vocabulary of $\approx$ 10K English words. We have extracted a total of 9092 high-quality video recordings of the news for the hearing impaired available freely on the internet. Specifically, these videos consist of front-facing professional signers signing the spoken sentences that a hidden speaker says, as shown in Figure 2. Our data consists of over 18 hours of videos across 5 professional signers covering various topics, including current affairs, sports, and world news. A comparison of our dataset with the existing datasets is presented in Table 1. As seen from the table, German sign language dataset Phoenix 14T [15] is significantly less challenging in comparison to our dataset due to its limited vocabulary of 2887 words.
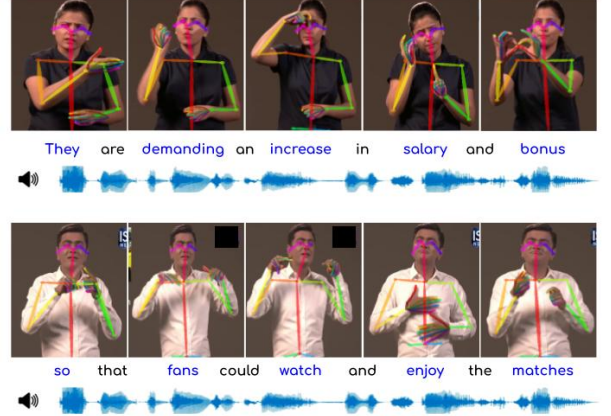

Figure 2: *We introduce a new, large-scale Indian Sign language (ISL) dataset containing continuous sign language videos from 5 signers along with the corresponding text and speech transcripts. Our dataset covers a large number of topics and has a vocabulary of around 10K English words.*

Table 1: *Comparison of our dataset with several other publicly available datasets containing sign language information. Ours is the first continuous Indian sign dataset to contain natural speech annotations.*

Dataset	lang	vocab	#hours	#signers	continuous	speech
KETI [16]	KSL	524	20	14	✓	✗
Boston104 [14]	ASL	104	0.7	3	✓	✗
Phoenix 14T [15]	GSL	3k	11	9	✓	✗
INCLUDE [19]	ISL	263	-	7	✗	✗
Ours	ISL	10k	18	5	✓	✓

3.2. Multi-tasking Transformer Network

Generating sign language poses solely from speech is a highly challenging task. We propose a multi-tasking transformer network coupled with cross-modal discriminator to generate poses directly from speech segments. We describe the different modules of our network in this section.

The backbone of our proposed model is a modified transformer architecture introduced in [20]. In most natural language processing tasks, both the input and the target domain consist of discrete vocabulary. On the contrary, for our task, both input and target belong to the continuous space. To assist our primary goal of generating sign language poses from speech, we introduce an auxiliary task of recognizing the input speech and train jointly. Our transformer architecture consists of three major components: (i) a joint speech encoder, (ii) a pose decoder and (iii) a text decoder decoding the complementary modalities. Finally, we also employ a cross-modal matching network as a discriminator to help the transformer learn the high-quality translations from speech to sign language. A pictorial representation of our architecture is given in Figure 3.

3.2.1. Generating Poses from Videos

We use OpenPose [21] to represent the body poses using the human pose joint keypoints for the signers. We consider 50 human upper body key points including hands and fingers. To minimize the effect of noisy and missed detections of OpenPose, we use an iterative back-propagation based method using inverse kinematics to lift key points from $2D$ to $3D$ as suggested by [3].
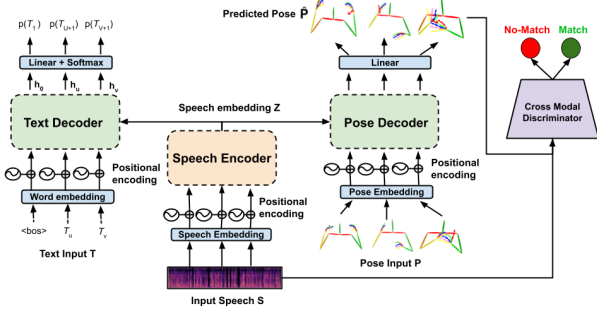
Figure 3: *Multi-task transformer architecture which generates sign language pose sequences along with the secondary task of generating the text translations for the given speech input.*

3.2.2. Speech Encoder

Our network starts with a speech encoder SE used to learn contextual embeddings from the input speech segment. We use melspectrograms representation of speech S with 80 mel frequency bins. We project the input melspectrogram of shape $T_s \times 80$ where T_s is the number of STFT time steps into a dense continuous space using a linear layer. Following this, a standard positional encoding is applied similar to [20]. We then use N multi-head self-attention layers with M heads and also a position-wise feed-forward layer. Each of the two sublayers has residual connections and layer normalization. Our speech encoder produces outputs of dimension d_{model} and the final embedding is represented by $Z = \{z_1, z_2, ..z_n\}$.

3.2.3. Pose Decoder

The task of the pose decoder PD is to attend to the learned context from the speech encoder and the previous time steps to generate pose at the current time step. Input to pose decoder is the masked pose sequence of dimension $m \times 151$. Pose p_i for every ith frame is initially projected to a dense continuous space using a linear layer. To guide the generation process without the special symbol denoting the start and end of the sequence as done while handling discrete vocabulary, we add a counter in the range of $0 - 1$ to every frame's pose similar to [2]. We then use positional encoding and N masked multi-head self-attention layers with M heads which produce embeddings of dimension d_{model}. This is followed by N multi-head cross-modal attention layers attending to speech embeddings modeling the translation between speech and pose. Finally, a position-wise feed-forward layer followed by a dense layer is used to generate the predicted pose $\hat{p}_i$ for ith frame given as,

$$\hat{p}_i = PD(\tilde{p}_{i-1}|z_{1:n}, \tilde{p}_{1:i-2}) \qquad (1)$$

Here, $z_{1:n}$ represents the contextual embedding of the speech input segment S obtained from the speech encoder, $\tilde{p}_{i-1}$ is the pose embedding for pose frame $i - 1$. The predicted pose is used to calculate the regression loss for the network given as,

$$\mathcal{L}_{Reg} = \frac{1}{m}\sum_{i=1}^{m}(p_i - \hat{p}_i)^2 \qquad (2)$$

3.2.4. Text Decoder

The text decoder TD is similar to a classical transformer decoder block for discrete vocabulary. The input to this block is the masked text translations of length Q. It uses positionally encoded input tokens at time step i to decode the subse-
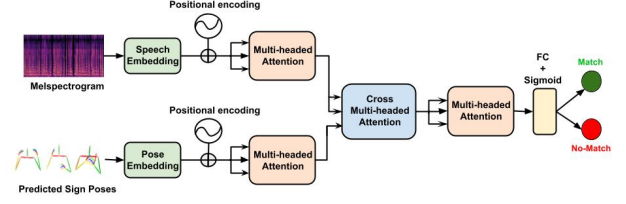

Figure 4: *We fuse the two modalities with a cross-modal discriminator, allowing the network to correlate between poses and speech time-steps. This fused representation is used to perform a binary classification to classify "match" or "no-match".*

quent tokens in the sequence until an end-of-sequence token is produced. It also consists of masked multi-head self-attention blocks followed by cross attention with the speech context embeddings and then non-linear point-wise feed-forward layers.

$$h_i = TD(\tilde{t}_{i-1}|z_{1:n}, \tilde{t}_{1:i-2}) \qquad (3)$$

where h_i is the output embedding from the text decoder, $\tilde{t}_{i-1}$ represents the word embedding for spoken language token t_{i-1} at $(i - 1)th$ time step and $z_{1:n}$ represents the contextual embedding of the speech input S. The output at every time step h_i is then used to get a softmax probability over the target vocabulary V. A cross entropy loss given in equation 4 is computed and back-propagated over the network.

$$\mathcal{L}_{Xent} = 1 - \prod_{i=1}^{Q}\sum_{j=1}^{V}p(t_i^j)p(\hat{t}_i^j|h_i) \qquad (4)$$

Here, Q is the sentence length, the V is the target vocabulary, $p(t_i^j)$ is the ground truth probability and $p(\hat{t}_i^j|h_i)$ is the predicted probability for the ith token to be t^j .

3.2.5. Cross-Modal Discriminator

We introduce a matching network which is used as a cross-modal CM discriminator. This network is used to match the speech segments and the corresponding sign pose sequences. Figure 4 shows the design of our cross-modal matching network. It consists of separate speech and pose embedding layers which first learn a high dimensional embedding of input speech and pose sequence. We have multiple self attention blocks to learn attention aware representations for each modality which are then fused using a multi-headed cross attention block. The cross attention fusion of two modalities learns to embed speech encoding into the pose space and learns the relationship between them. The cross-modal attention block is followed by a position-wise feed-forward layer and a few self-attention blocks. Finally, a fully-connected layer with sigmoid non-linearity is used to obtain a probability signifying "match" or "no-match". In our experiments, we show the effectiveness of cross-modal discriminator in improving the network's performance.

3.2.6. Joint Training of Transformer with Cross-Modal Discriminator

We train the multi-task transformer with cross-modal discriminator jointly in a conditional generative adversarial network [22] setting to produce more realistic pose sequences. Our proposed multi-task transformer acts as a generator G, and the cross-modal matching network acts as a discriminator D, and they compete in a min-max game given by the objective function,

Table 2: *We report scores from various baseline approaches along with our proposed network. We can see that our network outperforms other baselines in the text to sign language and speech to sign language settings. Since [10] was trained on Phoenix 14T [15] dataset in the original paper, we also report the scores on the same data as a reference (last row).*

Dataset	Approach	Task	DTW↓	PCK↑
ISL dataset	Saunders et al. [10]	Text to SL	16.26	40.14
ISL dataset	LSTM + attention	Speech to SL	17.37	39.30
ISL dataset	Ours w/o multi-tasking	Speech to SL	16.71	41.60
ISL dataset	Ours w/o discriminator	Speech to SL	14.94	49.40
ISL dataset	**Ours**	**Speech to SL**	**14.05**	**53.30**
Phoenix 14T [15]	Saunders et al. [10]	Text to SL	11.44	38.12

$$\min_G \max_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}[\log D(X|Y)]$$
$$+ \mathbb{E}[\log(1 - D(G(\hat{X}|Y)] \quad (5)$$

where, $\hat{X}$ represents the generated pose sequence, X represents the target pose sequence, and Y represents the ground truth input speech. Thus, the overall loss for the network is given as a weighted sum of losses in Equations 2, 4 and 5,

$$\mathcal{L}_{Total} = \lambda_{Reg}\mathcal{L}_{Reg} + \lambda_{Xent}\mathcal{L}_{Xent} + \lambda_{GAN}\mathcal{L}_{GAN} \quad (6)$$

where λ_{Reg} is the regression loss weight, λ_{Xent} is the recognition loss weight, and λ_{GAN} is the weight for adversarial loss.

3.3. Implementation details

We have implemented all our models using the PyTorch framework [23]. In all the transformer layers, we use an embedding size of $d_{model} = 512$, $N = 2$ layers and number of heads $M = 8$. We use Xavier initialization and Adam optimizer with an initial learning rate of 10^{-3} for training the transformer and the cross-modal discriminator. We also use data augmentation like predicting multiple-frame poses at each time step as done by [2]. We predict 10 frames at every time step to penalize the network heavily for producing mean poses. We have set $\lambda_{Reg} = 1, \lambda_{GAN} = 10^{-4}, \lambda_{Xent} = 10^{-3}$ in our experiments.

4. Results and Evaluation

We evaluate the quality of the generated sign language pose sequences from different models using Dynamic Time Warping (DTW) and Probability of Correct Keypoints (PCK) scores. DTW finds an optimal alignment between two time series by non-linearly warping them. Lower DTW corresponds to better pose generations. On the other hand, PCK is used in several pose detection and generation works [24], [25]. It evaluates the probability of pose key points to be close to the ground truth key points up to a threshold of $\alpha = 0.2$. Higher PCK corresponds to better pose generations.

We start by comparing with the state-of-the-art text to sign language model [10] trained on the proposed ISL dataset as shown graphically in Figure 5. Since ours is the first work to deal with speech to sign language generation, we do not have directly comparable networks for the given task. Thus, we also establish several baselines for the first time, as shown in Table 2. Our first proposed baseline is a modified speech-to-text [26] network consisting of an LSTM seq2seq network with an attention mechanism. We modified the network for the regression task of generating pose key points instead of
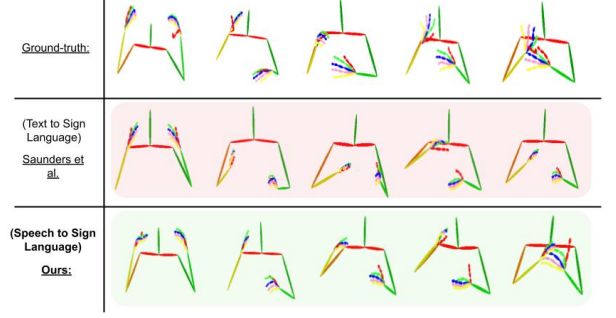

Figure 5: *We show a graphical comparison between the generated outputs from the state-of-the-art text to sign language and our speech to sign language models. We can observe that the poses generated by our model are closer to the ground-truth pose sequence indicating its superiority.*

Table 3: *Quantitative evaluation for speech to sign language generation using different discriminators.*

Approach	DTW↓	PCK↑
Ours w/o discriminator	14.94	49.40
Ours with conv. disc	15.07	45.41
Ours with pose-only disc	15.34	43.76
Ours	**14.05**	**53.33**

character probabilities. This network serves as a comparison between the RNN based approach and our transformer-based method. We observe that using an LSTM-based network leads to mean pose being generated at each time step, severely impacting the outcome. We also train a single task network by removing the text decoder block from our architecture. As evident from Table 2, incorporating an auxiliary task of generating text sequences from speech helps in the main task of sign language generation. Finally, we also evaluate the importance of the cross-modal discriminator by training a model without the discriminator. As seen from Table 2, our network with cross-modal discriminator achieves significantly better metric scores (DTW:14.05, PCK:53.33) compared to the best available text to sign language model [10] and other speech to sign language baselines. Since [10] was originally trained on the Phoenix 14T [15] dataset, we also report the reference scores for the reader in the same table.

To highlight the importance of the cross-modal matching network, we perform an ablation experiment with other discriminators. We use a 1D convolutional discriminator to classify generated pose sequence as real or fake, similar to the one proposed by Saunder et al. in [10]. As seen in Table 3, other types of discriminators are detrimental for the network and work poorer than the "without" discriminator setup.

5. Conclusion

In this work, we introduce the task of translating the spoken language to sign language pose sequences to make way for two-sided communication between the hearing impaired and the rest of the world. We propose a new ISL dataset with speech modality and avoid the expensive gloss annotations. We are the first to achieve state-of-the-art results for the task of speech to sign language generation using a multi-task transformer network coupled with a cross-modal discriminator. Our work paves the way for future research in speech to sign language generation that may one day result in real-time sign language translation.

6. References

[1] B. O. Olusanya, K. J. Neumann, and J. E. Saunders, "The global burden of disabling hearing impairment: a call to action," *Bull World Health Organ*, vol. 92, no. 5, pp. 367–373, May 2014.

[2] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive transformers for end-to-end sign language production," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 687–705.

[3] J. Zelinka and J. Kanis, "Neural sign language synthesis: Words are our glosses," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 3384–3392.

[4] S. Sinha, *Indian Sign Language: A Linguistic Analysis of Its Grammar*. Gallaudet University Press, 2018. [Online]. Available: https://books.google.co.in/books?id=cy6_tAEACAAJ

[5] S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt, and S. Abbott, "Tessa, a system to aid communication with deaf people," in *Proceedings of the Fifth International ACM Conference on Assistive Technologies*, ser. Assets '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 205–212.

[6] J. Glauert, R. Elliott, S. Cox, J. Tryggvason, and M. Sheard, "Vanessa - a system for communication between deaf and hearing people," *Technology and Disability*, vol. 18, pp. 207–216, 2006.

[7] B. Wyvill, C. McPheeters, and G. Wyvill, "Animating soft objects," *The Visual Computer*, vol. 2, no. 4, pp. 235–242, 1986.

[8] S. Stoll, N. C. Camgöz, S. Hadfield, and R. Bowden, "Sign language production using neural machine translation and generative adversarial networks," in *29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association, 2018. [Online]. Available: http://epubs.surrey.ac.uk/848809/

[9] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2sign: Towards sign language production using neural machine translation and generative adversarial networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 891–908, Apr 2020. [Online]. Available: https://doi.org/10.1007/s11263-019-01281-2

[10] B. Saunders, N. C. Camgöz, and R. Bowden, "Adversarial training for multi-channel sign language production," *The 31st British Machine Vision Virtual Conference*, 2020. [Online]. Available: http://epubs.surrey.ac.uk/858417/

[11] R. San-Segundo, R. Barra, R. Córdoba, L. F. D'Haro, F. Fernández, J. Ferreiros, J. M. Lucas, J. Macías-Guarasa, J. M. Montero, and J. M. Pardo, "Speech to sign language translation system for spanish," *Speech Commun.*, vol. 50, no. 11–12, p. 1009–1020, Nov. 2008. [Online]. Available: https://doi.org/10.1016/j.specom.2008.02.001

[12] S. Paulson and B. Thilagavathi, "An adaptable speech to sign language translation system," *International journal of engineering research and technology*, vol. 3, 2014.

[13] K. Khalil, K. El-Darymli, O. Khalifa, and H. Enemosah, "Speech to sign language interpreter system (sslis)," in *IEEE International Conference of Computer and Communication Enginnering (ICCCE'06)*, 01 2006.

[14] M. Zahedi, D. Rybach, T. Deselaers, J. Bungeroth, and H. Ney, "Continuous sign language recognition - approaches from speech recognition and available data resources," in *In Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, 2006, pp. 21–24.

[15] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7784–7793.

[16] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Applied Sciences*, vol. 9, no. 13, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/13/2683

[17] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018, pp. 2257–2264.

[18] A. Nandy, J. S. Prasad, S. Mondal, P. Chakraborty, and G. C. Nandi, "Recognition of isolated indian sign language gesture in real time," in *Information Processing and Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 102–107.

[19] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "Include: A large scale dataset for indian sign language recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1366–1375. [Online]. Available: https://doi.org/10.1145/3394171.3413528

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[21] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, jan 2021.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, cite arxiv:1411.1784.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[24] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[25] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3492–3501.

[26] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.