INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multimodal Locally Enhanced Transformer for Continuous Sign Language Recognition

Katerina Papadimitriou, Gerasimos Potamianos

Department of Electrical & Computer Engineering, University of Thessaly, Volos, Greece
aipapadimitriou@uth.gr, gpotam@ieee.org

Abstract

In this paper, we propose a novel Transformer-based approach for continuous sign language recognition (CSLR) from videos, aiming to address the shortcomings of traditional Transformers in learning local semantic context of SL. Specifically, the proposed relies on two distinct components: (a) a window-based RNN module to capture local temporal context and (b) a Transformer encoder, enhanced with local modeling via Gaussian bias and relative position information, as well as with global structure modeling through multi-head attention. To further improve model performance, we design a multimodal framework that applies the proposed to both appearance and motion signing streams, aligning their posteriors through a guiding CTC technique. Further, we achieve visual feature and gloss sequence alignment by incorporating a knowledge distillation loss. Experimental evaluation on two popular German CSLR datasets, demonstrates the superiority of our model.

Index Terms: continuous sign language recognition, RNN, Transformer, RWTH-PHOENIX Weather 2014, RWTH-PHOENIX Weather 2014T

1. Introduction

Over the past three decades, significant attention has been devoted to automatic SLR from video due to its potential to provide accessibility for the hearing impaired. Since SL constitutes a complex non-verbal communication means, with numerous manual and non-manual cues participating in signing, its recognition is an intricate task suffering from articulations complexity and correlation, as well as the signing variability among subjects. Such issues naturally arise when dealing with isolated signs [1–3], but are substantially more challenging in the instance of continuous SLR [4–6] due to the absence of gloss-level segmentation. The task of CSLR from video data constitutes the focus of our paper.

A critical aspect of CSLR research lies on the sequence modeling approach used for gloss prediction. For this purpose, early systems rely on HMM-GMMs [7, 8], while most contemporary CSLR schemes employ recurrent neural networks (RNNs), typically the LSTM networks [9] or bi-directional LSTM (BiLSTM) [10], in conjunction with connectionist temporal classification (CTC) [11]. Specifically, the work in [8] addresses the CSLR task employing a 2D convolutional neural network (CNN) and a BiLSTM encoder for spatio-temporal feature extraction, as well as the CTC loss function for alignment purpose. Further, in [12] a 3D-CNN model is employed as visual feature learner and a stacked dilated convolutional network with CTC is applied for sequence learning. In addition, several proposed systems employ the Transformer encoder with CTC for sequence learning, such as the works in [13, 14]. Moreover, various approaches treat CSLR as a video-to-text translation task, employing attention-based encoder-decoders [15–17]. Recently, some efforts in the literature have focused on the combination of temporal convolutions (TCNs), which are suitable for capturing temporal features, with RNN encoders [4,5,18] or Transformers [6, 19, 20].

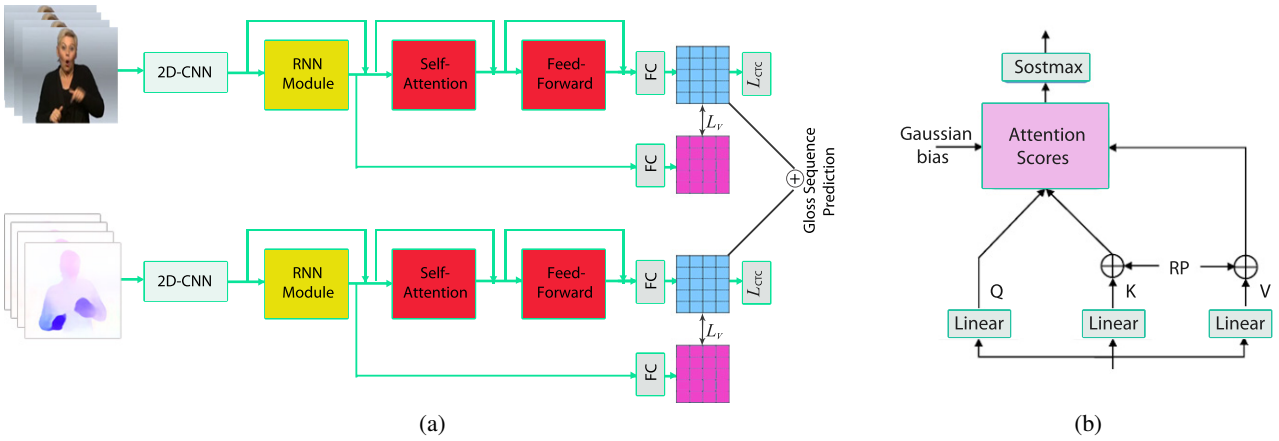
Despite their remarkable performance due to their ability to capture long-term dependencies, Transformers neglect the local structures that exhibit in SL sequences. To this end, the work in [6] introduces a local Transformer coupled with relative position encoding [21] and localness modeling through a Gaussian bias [22]. Inspired by the former and the work in [23], which combines RNN and Transformer encoder, in this work, we propose a novel model that relies on a window-based RNN (LSTM) module followed by a multi-head self-attention based Transformer encoder. Specifically, the RNN module operates on short windows of the input sequence, generating visual latent representations that are fed into the Transformer, which is enhanced with local self-attention via relative position encoding and Gaussian bias, for capturing short-term dependencies of the visual latent representations. In addition, we adopt multi-head attention mechanism in the Transformer for learning global structures.

The second axis concerns the visual module, as well as the type of visual modalities integrated in the CSLR framework. Most works in the literature rely only on RGB appearance representations of articulation regions or of the full video frame based on 2D or 3D CNNs [12, 16, 24]. Others combine skeletal features with appearance representations [6, 18], while our previous work in [25] integrates appearance representations and optical flow features into human pose, modeled via ST-GNNs. In this work, based on the assumption that different modalities could potentially complement each other, we design a multimodal framework, where the proposed CSLR sequence learning model operates on two different streams, i.e., RGB appearance frames and optical flows, and their scores are fused after being aligned through the CTC guiding technique of [26]. Note that for image feature learning a 2D-CNN model is employed.

Finally, a crucial component in CSLR concerns the alignment module used for aligning the features extracted from the sequence learning model with the gloss sequence. Most works in the literature employ a CTC-based alignment module [8, 12, 18]. Nevertheless, solely utilizing the CTC loss may cause training problems due to the fact that the extracted features may not be sufficient to yield precise recognition out-

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty Members & Researchers and the procurement of high-cost research equipment grant" (Project "SL-ReDu", Project Number HFRI-FM17-2456).

10.21437/Interspeech.2023-2198



(a) (b)

Figure 1: *(a) An overview of the proposed multimodal CSLR model that generates a sequence of glosses via an RNN module and a Transformer encoder and (b) self-attention enhanced with local modeling via Gaussian bias and relative position encoding (RP).*

comes. To tackle this, the works in [5,12,27] employ stage optimization strategies to improve the extracted features, while the work in [6] integrates auxiliary learning to enhance the CSLR backbones consistency. Motivated by the above and inspired by the work in [5], in this work we design an alignment module, which combines the CTC loss with a knowledge distillation loss [28], which aligns the probability distribution produced by the sequence learning model and the probability distribution obtained by the RNN module.

In summary, our contributions lie on: (i) the development of a novel sequence learning model that combines a window-based RNN module and a Transformer encoder, enhanced with localness modeling via relative position encoding and a Gaussian bias; (ii) the design of a multi-modal framework, which applies the proposed sequence learning model to an RGB and an optical flow stream, and ensembles them by performing CTC guiding alignment; and (iii) the conjunction of the CTC loss with a visual alignment loss. To the best of our knowledge, the integration of a window-based RNN module with a local Transformer has never been investigated in the literature.

We evaluate the introduced approach on two popular large-scale German CSLR benchmarks, the "RWTH-PHOENIX Weather 2014" [29] multi-signer corpus and the "RWTH-PHOENIX Weather 2014T" dataset [13], and we provide in-depth ablations that highlight our innovations. We achieve competitive performance on both datasets compared to the current state-of-the-art. Further, the proposed outperforms our baseline, i.e., TCN and Transformer based sequence learning module.

2. Our Approach

As shown in Figure 1, our approach constitutes a deep-learning based model that learns from two different modalities, namely RGB frames and optical flows. The baseline CSLR model used for both streams composes of: (i) a visual module, which adopts a 2D-CNN based spatial feature learner and a window-based RNN module for local context visual features extraction; (ii) a sequence model relying on a Transformer encoder enhanced with relative position encoding and a Gaussian bias; and (iii) an alignment module integrating both CTC and knowledge distillation loss functions. To ensemble the RGB and optical flow streams, a guiding CTC technique is employed.

2.1. Appearance and Optical Flow Features

As already mentioned, we employ the full-frame RGB stream, as well as the optical flow one as an additional visual representation, since it effectively captures the motion information of the numerous SL articulators. For this purpose, we use the SpyNet model [30] generating motion informative images. For both appearance and optical flow features, we consider visual representations based on the VGG11 network [31], encouraged by its prominent visual representation learning capability in the CSLR task [4, 6]. In particular, an image frame sequence of length T, $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_T}) \in \mathbb{R}^{T \times C \times H \times W}$ with C= 3 for both RGB and optical flow frames, H denoting the height, and W representing the width of frames, is appropriately rescaled (256×256) and cropped to 224×224, before being fed to the 2D-CNN model. The 2D-CNN model follows the VGG11 architecture pre-trained on the ImageNet corpus [32] and is coupled with a global average pooling layer, yielding 512-dimensional features ($\mathbf{z} = (\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_T}) \in \mathbb{R}^{\mathbf{T} \times \mathbf{512}}$).

2.2. Window-based RNN module

Before learning global dependencies via the multi-head attention mechanism, we refine the visual representations of each frame to incorporate the sequential and local information of its neighborhood. To capture the local short-term dependencies of the frame sequence, the visual representations extracted from the 2D-CNN feature learner are fed to an RNN module that operates on local windows of frames producing latent representations for each of them. In particular, inspired by [23], instead of applying RNNs to the whole sequence, we rearrange the initial frame feature sequence into many short ones using a local window of fixed size M for each target frame, such that each short sequence includes M consecutive frames with the last being the target frame. To achieve this, we prepad the beginning of the input sequence by M-1. As shown in Figure 2, M frames in the local window form a short sequence, which is subsequently processed by the RNN unit, producing the latent representations. More precisely, the local sequence $\mathbf{z_t} = (\mathbf{z_{t-M-1}}, \mathbf{z_{t-M-2}}, ..., \mathbf{z_t})$ with window size M, is passed through the RNN unit, generating the hidden state representation $\mathbf{s_t}$. Note that the RNN module relies on BiLSTM networks [9] and is followed by a normalization layer.

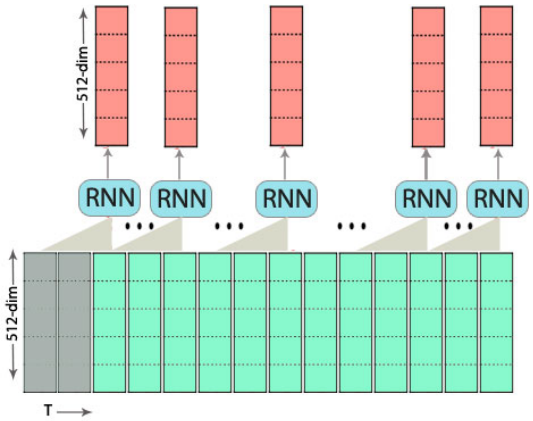


Figure 2: *Illustration of the window-based RNN module (M=3).*

2.3. Transformer encoder

Each gloss in a SL video only lasts a few frames, indicating the value of local contexts. Motivated by this, here, we employ a Transformer encoder, enhanced with local modeling, as well as with global structure modeling for sequence learning. To capture the global long-term dependencies, a multi-head attention layer followed by a feed-forward one, is applied (see also Figure 1). Both layers are coupled with normalization. Specifically, the visual feature representation sequence $\mathbf{s} \in \mathbb{R}^{T\times 512}$ extracted from the window-based RNN module is subjected to three seperate linear layers generating the queries $\mathbf{Q} \in \mathbb{R}^{T\times 512}$, the keys $\mathbf{K} \in \mathbb{R}^{T\times 512}$, and the values $\mathbf{V} \in \mathbb{R}^{T\times 512}$. Since multi-head attention is adopted, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are splitted, resulting in $\mathbf{Q^h}, \mathbf{K^h}, \mathbf{V^h} \in \mathbb{R}^{T\times 512/n_h}$, with n_h denoting the number of heads and $\mathrm{h} = 1, ..., n_h$. Subsequently, each split passes independently through a self-attention layer producing the attention scores, which are then combined together to produce the final one. The attention score for each head is computed as follows:

$$attn = \frac{\mathbf{Q^h}(\mathbf{K^h})^{'}}{\sqrt{512/n_h}} \in \mathbb{R}^{n_h\times T\times T}.$$

Since local context dependencies are critical to the CSLR performance, we enhance the self-attention layer with localness modeling. As shown in Figure 1, to obtain this we add relative representations [21] to the queries $\mathbf{K}$ and values $\mathbf{V}$ enhancing neighboring relations. To further enhance local context modeling, we employ the Gaussian distribution with a fixed window size as additive bias to mask the self-attention scores. Thus, the output of the multi-head self-attention layer is formulated as:

$$F = concat\{softmax(attn^h + bias^h)V^h\}_{h=1}^{n_h}.$$

2.4. Alignment Module

As already mentioned, our CSLR model involves an alignment module, which combines the CTC loss with a knowledge distillation loss [28] aligning the probability distribution generated by the sequence learning model and the probability distribution obtained by the RNN module's visual features. In particular, the latent representations derived from the Transfromer encoder are fed to a linear projection layer followed by a softmax activation, yielding probabilities distribution $p_w(\mathbf{G}|\mathbf{F})$ for all possible signing videos $\mathbf{F}$ to gloss sequences $\mathbf{G}$ alignments, modeled as follows: $p_w(\mathbf{G}|\mathbf{F}) = \sum_{\pi\epsilon\mathbf{B}}(\pi|\mathbf{F})$, where π is a sequence path with $\pi_t \,\epsilon\, \{\_, G_1, G_2, ..., G_L\}$. Note that, L is the gloss vocabulary size, which is further complemented with the blank character, and B denotes all the possible label paths. Thus, the CTC loss is defined as: $\mathcal{L}_{CTC} = -\log p_w(\mathbf{G}|\mathbf{F})$.

Correspondingly, for the visual alignment loss function we pass the visual features extracted from the RNN module through a linear projection layer followed by a softmax activation, yielding probabilities distribution $p_v(\mathbf{G}|\mathbf{S})$. Subsequently, we incorporate the KL-divergence loss function introduced in [5], which minimize the distance between the probability distribution produced by the sequence learning model and the probability distribution generated from the output of the visual module, formulated as:

$$\mathcal{L}_V = \mathrm{KL}(softmax(\frac{\mathbf{S}}{\tau}), softmax(\frac{\mathbf{S}}{\tau})),$$

with τ being the temperature. In particular, training is conducted using a linear combination of the two loss functions, i.e: $\mathcal{L}_M = \mathcal{L}_{CTC} + \mathcal{L}_V$.

2.5. Ensemble Module

Since CTC based alignment models suffer from non-aligned spike timings in the probability distribution, direct posterior fusion (late fusion) of the two modalities, i.e, the RGB and the optical flow streams, turn out to be ineffective. To tackle this, we follow the two-step training approach of [26]. Specifically, the CSLR model trained using only the RGB modality is treated as the guiding model. During the guided model training, namely that of the optical flow stream, the probability distributions predicted by the guiding model are converted to a mask $M(F)$ with ones at the output symbol with the highest posterior and zeros at other symbols, as well as the blank symbol. Then, the probability distribution $P(F)$ generated by the guided model are subjected to element-wise multiplication with the mask M. Thus, a masked probability distribution $\hat{P}(F) = M(F) \odot P(F)$ is generated, which is used to compute the guided loss function defined as: $\mathcal{L}_G = -\sum \hat{P}(F)$. Finally, the overal training loss is formulated as: $\mathcal{L} = \mathcal{L}_{CTC} + \mathcal{L}_G$

3. Experimental Evaluation

3.1. Dataset and Experimental Framework

RWTH-PHOENIX Weather 2014 [29]: This is a continuous German SL dataset, providing 6,841 different gloss sentences, extracted from the German TV station PHOENIX news and weather forecast. The corpus signed vocabulary consists of 1,232 unique glosses (around 80,000 gloss instances) performed by 9 different signers. Video data is provided at a frame-rate of 25 Hz and 210 × 260-pixel resolution. The corpus comprises two settings: multi-signer and signer-independent. In the scope of this work, we employ the multi-signer setting, where 5,672 video samples are allocated to training, 540 to validation, and 629 to testing.

RWTH-PHOENIX Weather 2014T [13]: This dataset constitutes an expansion of the RWTH-PHOENIX Weather 2014 [29] corpus, involving both gloss and spoken language translation sentence pairs for German SL video samples of weather forecasts. The corpus volume involves in total 8,257 German SL sequences with frame resolution of 210 × 260 at a rate of 25Hz expressed by 9 signers leading to a vocabulary of 1,066 unique glosses and 2,887 spoken language words, respectively. We use the existing multi-signer set, comprising 7,096 videos for training, 519 for validation, and 642 for testing.

Table 1: *Ablation study for the RNN module (RNN), the relative position encoding (RP), and the Gaussian bias (GB) of single and both modalities. The evaluation is conducted using the "RWTH-PHOENIX Weather 2014" dataset.*

Modalities	RNN	RP	GB	WER (%)
RGB				27.55
	✓			23.25
		✓	✓	24.05
	✓	✓	✓	21.25
Optical Flow				29.18
	✓			26.20
		✓	✓	25.87
	✓	✓	✓	25.07
Both	✓	✓	✓	**20.89**

3.2. Implementation Details

The RNN module comprises a set of two-layer BiLSTMs with hidden dimensionality equal to 512. Both layers are coupled with a normalization layer. For the sequence learning model, we employ a 3-layer Transformer encoder with hidden states of 512 and 8 heads. As in [6], for the Gaussian bias we employ a fixed window size equal to 6.3 for both datasets. Training is conducted using the Adam optimizer [33] with initial learning rate being equal to 0.0001 decayed by a factor of 0.0001, a dropout rate of 0.1, and a batch size of 2. During inference, we use beam search decoding with beam width 5. Further, temperature τ in the $\mathcal{L}_V$ loss is fixed to 8. The model is implemented in PyTorch [34] and experiments are carried out in a NVIDIA GeForce RTX 3090 GPU.

3.3. Results

Our CSLR model performance is evaluated quantitatively in terms of word error rate (WER) (%), taking into account the number of substitutions, deletions and insertions in the predicted hypotheses. As shown in Table 1, the introduced model is firstly evaluated on the "RWTH-PHOENIX Weather 2014" dataset against some variations it. Specifically, ablation study for the various components of the proposed are provided when single and both modalities are considered. As deduced from the Table 1, our model demonstrates superior performance when all modaliies are considered, yielding 20.89% WER. This reveals the benefit of using multiple feature streams that are complementary to each other. In addition, the RNN module, as well as the relative position encoding seem to be the most robust components, while Gaussian bias incorporation benefits system performance. Further, the RGB appearance modality achieves lower WER(%) than the optical flow one. Moreover, in Table 2 we examine the contribution of the various loss functions incorporated into our model. As it can be readily seen, the higher contribution to the system performance is obtained via the CTC guiding loss function, while visual alignment loss further impoves WER by 0.86% absolute.

In Tables 3 and 4 evaluation comparison of the proposed against current state-of-the-art for the "RWTH-PHOENIX Weather 2014" multi-signer corpus and the "RWTH-PHOENIX Weather 2014T" dataset is provided. As it can be observed our model achieves competitive performance on the two CSLR datasets, namely 20.89% WER on the "RWTH-PHOENIX Weather 2014" multi-signer corpus and 20.73% WER on the "RWTH-PHOENIX Weather 2014T" dataset. Specifically, in both cases, our model outperforms most results in the literature, coming very close to the state-of-the-art (19.98% WER) of [20] in the "RWTH-PHOENIX Weather 2014"dataset, which employes a spatio-temporal based Transformer and the state-of-the-art (20.40% WER) of [6] in the "RWTH-PHOENIX Weather 2014T"dataset, which relies on local Transformer encoder and two auxiliary constraints for enhancing sequence learning. It should be noted that our model involves 36M parameters and the inference speed as measured on the NVIDIA GeForce RTX 3090 GPU is 0.29 seconds per video. Finally, we evaluate the performance of our model against one variation of it, where the RNN module is substituted by a TCN layer, increasing WER by 0.78% absolute on the "RWTH-PHOENIX Weather 2014" multi-signer corpus.

Table 2: *Ablation study for loss functions. Evaluation comparison on the "RWTH-PHOENIX Weather 2014" dataset.*

Proposed Model	WER (%)
w/o $\mathcal{L}_V$	21.75
w/o $\mathcal{L}_G$	24.16
w $\mathcal{L}_V$ & $\mathcal{L}_G$	**20.89**

Table 3: *Evaluation comparison on the "RWTH-PHOENIX Weather 2014" dataset in terms of WER (%).*

Model	WER (%)
SubUnet [35]	40.70
SLT [36]	24.59
CNN-LSTM-HMM [37]	24.10
VAC [4]	22.30
SMKD [5]	21.00
STMC [18]	20.70
C2SLR [6]	20.40
STTN [20]	**19.98**
Proposed	**20.89**

Table 4: *Evaluation comparison on the RWTH-Phoenix-Weather-2014T dataset in terms of WER (%).*

Model	WER (%)
Re-Sign [8]	26.60
SFD+SGS+SFL [14]	26.10
Bi-ST-LSTM-A [16]	24.68
SLT [36]	24.59
CrossModal [24]	24.30
CNN-LSTM-HMM [37]	24.10
TDCNN [15]	23.70
SMKD [5]	22.40
ST-GCN [25]	21.34
STMC [18]	21.00
C2SLR [6]	**20.40**
Proposed	**20.73**

4. Conclusions

In this work, we focus on the intricate task of CSLR investigating the contribution of an innovative Transformer-based approach that seeks to capture both local and global signing dynamics. In particular, we employed a window-based RNN module to capture local temporal context and a Transformer encoder, enhanced with local context modeling through Gaussian bias and relative position information, as well as with global structure learning obtained via multi-head attention. Our model learns from two modalities, RGB and optical flow streams, which are fused via CTC guiding, achieving competitive performance on two large-scale German CSLR datasets.

5. References

[1] K. Grobel and M. Assan, "Isolated sign language recognition using hidden Markov models," in *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics*, 1997, pp. 162–167.

[2] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "A deep learning approach for analyzing video and skeletal features in sign language recognition," in *Proc. of the IEEE International Conference on Imaging Systems and Techniques*, 2018, pp. 1–6.

[3] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proc. of the CVPRW*, 2021, pp. 3408–3418.

[4] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *Proc. of the International Conference on Computer Vision (ICCV)*, 2021, pp. 11 522–11 531.

[5] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. of the International Conference on Computer Vision*, 2021, pp. 11 283–11 292.

[6] R. Zuo and B. Mak, "C2slr: Consistency-enhanced continuous sign language recognition," in *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5121–5130.

[7] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid CNN-HMM for continuous sign language recognition," in *Proc. of the BMVC*, 2016, pp. 1–12.

[8] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3416–3424.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the International Conference on Machine Learning*, 2006.

[12] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition," in *Proc. of the International Joint Conference on Artificial Intelligence*, 2018, pp. 885–891.

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

[14] Z. Niu and B. Mak, "Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition," in *Proc. of the ECCV*, 2020, pp. 172–186.

[15] K. Papadimitriou and G. Potamianos, "Multimodal sign language recognition via temporal deformable convolutional sequence learning," in *Proc. of the Interspeech*, 2020, pp. 2752–2756.

[16] Q. Xiao, X. Chang, X. Zhang, and X. Liu, "Multi-information spatialtemporal LSTM fusion continuous sign language neural machine translation," *IEEE Access*, vol. 8, pp. 216 718–216 728, 2020.

[17] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. of AAAI Conference on Artificial Intelligence*, 2018.

[18] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13 009–13 016.

[19] P. Xie, M. Zhao, and X. Hu, "PiSLTRc: Position-informed sign language transformer with content-aware convolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 3908–3919, 2021.

[20] Z. Cui, W. Zhang, Z. Li, and Z. Wang, "Spatialtemporal transformer for end-to-end sign language recognition," *Complex and Intelligent Systems*, 2023.

[21] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. of the Conference of the NAACL-HLT*, 2018, pp. 464–468.

[22] B. Yang, Z. Tu, D. F. Wong, F. Meng, L. S. Chao, and T. Zhang, "Modeling localness for self-attention networks," *ArXiv*, vol. abs/1810.10182, 2018.

[23] Y. Zheng, X. Li, F. Xie, and L. Lu, "Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6734–6738.

[24] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras, "Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space," *IEEE Access*, vol. 8, pp. 91 170–91 180, 2020.

[25] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Spatio-temporal graph convolutional networks for continuous sign language recognition," in *Proc. ICASSP*, 2022, pp. 8457–8461.

[26] G. Kurata and K. Audhkhasi, "Guiding CTC posterior spike timings for improved posterior fusion and knowledge distillation," in *Proc. of the Interspeech*, 2019, pp. 1616–1620.

[27] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in *Proc. of the ACM International Conference on Multimedia*, 2020, pp. 1497–1505.

[28] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *ArXiv*, vol. abs/1503.02531, 2015.

[29] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[30] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. of the CVPR*, 2017, pp. 2720–2729.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of the International Conference on Learning Representations*, 2015, pp. 1–14.

[32] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. of the NIPS-W*, 2017.

[35] N. C. Camgöz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," *Proc. of the IEEE International Conference on Computer Vision*, pp. 3075–3084, 2017.

[36] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[37] O. Koller, N. C. Camgoz, and H. N. nd R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2306–2320, 2020.