Hands-On: Segmenting Individual Signs from Continuous Sequences

Low Jian He, Harry Walsh, Ozge Mercanoglu Sincan, Richard Bowden

CVSSP, University of Surrey, Guildford, United Kingdom

{jianhe.low, harry.walsh, o.mercanoglusincan, r.bowden}@surrey.ac.uk

Abstract— **This work tackles the challenge of continuous sign language segmentation, a key task with huge implications for sign language translation and data annotation. We propose a transformer-based architecture that models the temporal dynamics of signing and frames segmentation as a sequence labeling problem using the Begin-In-Out (BIO) tagging scheme. Our method leverages the HaMeR hand features, and is complemented with 3D Angles. Extensive experiments show that our model achieves state-of-the-art results on the DGS Corpus, while our features surpass prior benchmarks on BSLCorpus.**
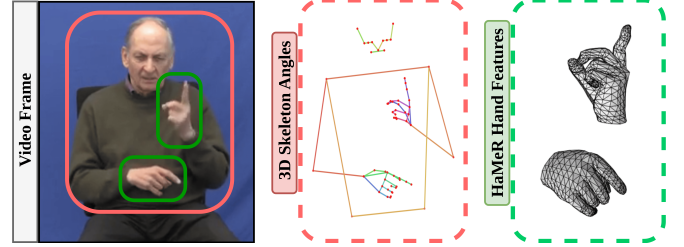
Fig. 1. **Illustration of Feature Representations:** The figure showcases a sample frame from a sign language video, accompanied by its corresponding 3D skeleton-based angular pose representation, and HaMeR-generated mesh visualizations for the left and right hands.

I. INTRODUCTION

Sign language is an expressive visual language that conveys meaning through hand gestures, facial expressions, and body movements [23]. As the primary means of communication for Deaf communities, sign languages reflect both a deep cultural and linguistic identity.

Advancements in sign language translation (SLT) systems offer a promising way to bridge communication gaps as they are designed to translate sign language into text and vice versa. However, the multi-channel nature of sign language, involving intricate hand movements, body positions, and facial expressions, poses significant challenges. This is further exacerbated by the scarcity of annotated data, particularly at the frame level, as annotating sign language videos is a labor-intensive, expertise-driven, and costly process [18]. This limits the availability of large-scale datasets and greatly hinders the generalization of SLT systems.

Sign segmentation, the task of identifying the temporal boundaries of signs in continuous videos, is thus critical to tackle these challenges. A robust segmentation model could automate much of the gloss[1] annotation process, allowing linguists to refine the predicted segments instead of manually labeling them, significantly reducing dataset annotation costs. Additionally, accurate sign segmentation could enhance the performance of SLT systems. While gloss-free models have gained traction due to limited annotated data [25], [28], gloss-based models still consistently outperform them [5], [27]. Hence, a strong segmentation model could revitalize gloss-based approaches, unlocking new opportunities in SLT.

Prior linguistic studies have shown that sign boundaries can be identified even by non-signers, through cues such as pauses and hand lowering motions [11]. Furthermore, it was also found that over 60% of signs in the SignBank dataset [24] consisted of a single hand shape [16], highlighting the importance of hand shape features in boundary detection.

---

[1]A gloss is a written representation of a sign in sign language, typically using a word or phrase from the local spoken language.

Building on these insights, we propose a novel sign segmentation model that integrates robust linguistic and visual cues. Our approach leverages the state-of-the-art Hand Mesh Recovery (HaMeR) model [17] to extract fine-grained hand shape features and incorporates 3D skeleton angles [13] to capture body-hand dynamics. We provide examples of these features on a single video frame in Fig. 1. These features are combined into a unified multi-modal framework, which leverages a transformer-based architecture to model temporal dependencies, enabling precise identification of pauses and transitions within sign language sequences.

In summary, our contributions are as follows: *(i)* We are the first to leverage HaMeR features for sign segmentation and demonstrate their effectiveness when combined with body pose features. *(ii)* We provide extensive experiments analyzing the impact of our design choices on performance. *(iii)* We achieve state-of-the-art segmentation results on the DGS Corpus. *(iv)* We demonstrate our feature's versatility by showing they surpass prior benchmarks when integrated into alternative sign segmentation frameworks on the BSLCorpus.

II. RELATED WORK

Sign segmentation is the task of identifying the temporal boundaries of signs—where a sign begins and ends—within continuous sign language videos. Its difficulty lies in the ambiguity of defining precise boundaries and the labor-intensive nature of annotating frame-level data, resulting in limited datasets. Additionally, unlike spoken language, sign languages are further complicated by the simultaneous use of hand configuration, movement, location, and facial expressions [26]. Despite these difficulties, significant progress has been made in sign segmentation and related tasks such as sign sequence alignment and sign localization.

Early works in sign segmentation addressed the lack of boundary annotations by aligning sign sequences to weak

labels, such as TV subtitles [2], [6] or speech recognition sequences [20]. These approaches relied on traditional algorithms, such as distance metrics [2] and mining techniques [6]. The utilization of Hidden Markov Models (HMM) in [20] marked a significant step forward, as it formally defined the task as sign segmentation, and laid the groundwork for future research. Following this, more advanced machine learning techniques were then explored. For instance, a random forest classifier [9] was used to segment sign boundaries in a small-scale Japanese sign language dataset [1].

Subsequent advancements then leveraged deep learning techniques. An Inflated 3D ConvNet (I3D) [4] with a Multi-Stage Temporal Convolutional Network (MS-TCN) [10] was introduced in [18] and able to achieve superior performance on the BSLCorpus [22] and the unreleased BSL-1k dataset. Further advancements were then made in [19], which proposed a changepoint-modulated pseudo-labeling algorithm to address source-free domain adaptation for sign segmentation. In parallel, Bull et al. [3] enhanced I3D embeddings by integrating them with BERT [8] embeddings within a transformer-based architecture to perform subtitle segmentation and alignment in sign language videos.

More recently, a Begin-In-Out (BIO) tagging scheme was introduced for sign segmentation [16], moving beyond the traditional binary in/out frame classification. The approach highlighted the importance of prosodic cues such as facial expressions and extended pauses. Thus, it analyzed the effects of multi-modal features by combining facial landmarks, body keypoints, optical flow, and 3D hand data. This approach also established a benchmark on the MeinDGS corpus [15], which has remained unchallenged. Therefore, we closely align our implementation with the methodology in [16] and evaluate our results in comparison.

III. METHODOLOGY

A. Problem Formulation

Prior work in sign segmentation has typically employed IO tagging or binary classification [18], where each frame is labeled either 'in' (I) or 'out' (O) of a sign, with 'in' indicating that the frame is part of a sign, and 'out' denoting that the frame is outside the sign. However, we instead adopt the more recently suggested formulation of sign segmentation based on a BIO tagging scheme [16], as it yields more precise sign boundaries and has demonstrated superior performances over IO tagging across various frame rates.

Thus, given a continuous sign language video with a sequence of frames $x = (x_1, x_2, ..., x_T)$, where T denotes the total number of frames, the goal is to produce an output sequence of $y = (y_1, y_2, ..., y_T)$. Each element $y_t \in \{0, 1, 2\}$ corresponds to a tag indicating the phase of the sign, with 0 indicating 'outside' (O), 1 for 'inside' (I) and 2 for the 'beginning' (B) of a sign segment.

B. Features

We hypothesize that hand shape and body position are critical features for effective sign segmentation. Thus, two carefully selected feature extraction methods were employed to capture this information:

1) HaMeR: The hand shape features are derived from the open-sourced HaMeR model [17], which reconstructs 3D hand meshes. Specifically, we utilize the transformer head section of HaMeR, which was originally used to regress hand and camera parameters. In our work, we repurpose the transformer head to extract hand pose and global orientation parameters, as they were highly effective hand shape descriptors in our tests. The hand pose features are represented as a tensor of $\mathbf{H} \in \mathbb{R}^{2 \times 15 \times 3 \times 3}$, while the global orientation is parameterized as $\mathbf{G} \in \mathbb{R}^{2 \times 3 \times 3}$. For simplicity and computational efficiency, we flatten these parameters, resulting in a 1D feature vector of $\mathbf{H}' \in \mathbb{R}^{270}$ for the hand pose, and $\mathbf{G}' \in \mathbb{R}^{18}$ for the global orientation. The flattened features are then concatenated to form a feature vector of $\mathbf{F} \in \mathbb{R}^{288}$, serving as our input hand shape representation.

2) 3D Skeleton Angles: To complement hand shape information, we incorporate 3D skeleton angle features extracted using the model in [13]. This model employs a Convolutional Neural Network (CNN) backbone, and utilizes a Forward Kinematics (FK) layer to generate a 3D human pose representation in the form of joints and angles. We adopt this model as our feature extractor due to its ability to deliver fast and accurate 3D pose estimations, while having a compact representation of $\mathbf{S} \in \mathbb{R}^{104}$. These features encode both body and hand positions, providing critical spatial context for hand movements relative to the body in 3D space.

C. Model Architecture

To achieve robust segmentation, we utilize a transformer-based architecture as the classification module due to its well-documented ability to model sequential dependencies and contextual relationships. As detailed in Sec. III-B.1 and III-B.2, hand shape and 3D skeleton features are extracted using pre-trained models; thus, they remain frozen during training. Meanwhile, to enhance feature-specific representations and facilitate multi-modal learning, we introduce two auxiliary modules and a mixer module, illustrated in Fig. 2. The auxiliary modules are feature-specific and are each a three-layer multi-layer perceptron (MLP) that up-projects HaMeR and 3D Angle features to $\mathbf{F}' \in \mathbb{R}^{512}$ and $\mathbf{S}' \in \mathbb{R}^{512}$.

To optimize computational efficiency, the features are downsampled temporally by a factor of 2. The downsampled features are concatenated along the feature dimension to form a multi-modal representation of $\mathbf{M} \in \mathbb{R}^{1024}$. The concatenated feature is then processed by the multi-modal mixer, also a three-layer MLP, to capture cross-modal interactions. The refined features are then passed to the transformer classifier to generate final per-frame predictions.

D. Experimental Setup

1) Training: Our model is trained for multi-class classification with target labels $y_t \in \{0, 1, 2\}$ using a unified loss function that combines frame-level and gloss-level objectives. For frame-level classification, we utilize multi-class cross-entropy, where $y = 2$ marks the start of a sign, $y = 1$
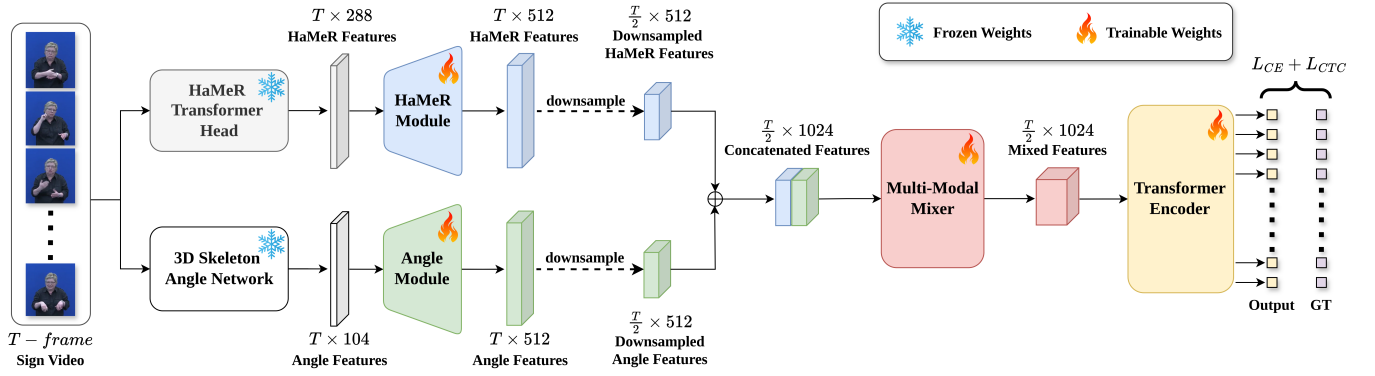
Fig. 2. **Architecture Overview:** Our approach consists of three stages: **(a) Feature Extraction:** Hand shape features are extracted using the HaMeR model's transformer head, while pose features are obtained from the 3D Skeleton Angle network. Feature-specific auxiliary modules are introduced to capture the characteristics of each feature. **(b) Feature Mixing:** Extracted features are concatenated and passed through a multi-modal module to learn inter-feature relationships. **(c) Classification:** A transformer encoder processes the multi-modal features to generate (Begin-In-Out) BIO scheme predictions.

denotes the continuation of the sign, and $y = 0$ represents non-sign frames. Concurrently, gloss-level supervision is incorporated using the Connectionist Temporal Classification (CTC) loss [12], allowing the model to leverage linguistic annotations while also providing a global constraint.

The model is optimized using the Adam optimizer [14] with a learning rate of 3×10^{-4}, and training is performed on a single RTX3090, employing the ReduceLROnPlateau scheduler with a patience of 5 epochs, gradient clipping of 0.1, and early stopping to prevent overfitting.

2) Dataset: **Public DGS Corpus** [15] is a German Sign Language dataset, recorded at 50 fps. It includes 330 sequences (average length 10 mins each) of free-flowing conversations between deaf participants on diverse topics, with precise frame-level gloss and sentence annotations. We utilize this dataset to train our model and evaluate its performance against [16], adhering to the MeineDGS translation protocols outlined in [21].

BSLCorpus [7], [22] is a British Sign Language dataset containing approximately 125 hours of video with gloss annotations and free translations. For sign segmentation, we use the subset with frame-level boundary annotations provided by [18], following the dataset splits and processing protocols to enable a direct comparison.

3) Evaluation Metrics: To ensure comparability with prior work, we evaluate our model using the same metrics proposed by [16]; specifically, the Frame-level F1 Score (F1), Intersection over Union (IoU), and Percentage of Segments (%). The F1 Score serves as the primary criterion for both model selection and early stopping, as it provides a balanced measure of precision and recall at the frame level. The IoU metric quantifies the overlap between predicted and ground truth segments, while the % metric complements this by assessing the ratio of predicted segments to the actual number of sign segments. Importantly, we only compute these metrics at the sign level, as phrase-level evaluation is less relevant for sign segmentation and gloss annotation.

Additionally, we adopt the mF1B and mF1S metrics in our comparisons with [18]. The mF1B metric assesses boundary detection by computing the F1 score across multiple integer-

TABLE I
ABLATION STUDY OF FEATURES AND MODULES ON DGS CORPUS [15]

Features		Modules			F1
3D Angles	HaMeR	3D Angles	HaMeR	Mixer	
✓	-	-	-	-	0.8306
-	✓	-	-	-	0.8324
✓	✓	-	-	-	0.8362
✓	✓	✓	✓	-	0.8508
✓	✓	✓	✓	✓	**0.8566**

valued thresholds within the range of [1,4]. Meanwhile, the mF1S metric assesses segmentation quality, considering a segment correct if its IoU exceeds a specified threshold.

IV. EXPERIMENTS

In this section, we present an ablation study for our model (Sec. IV-A), our comparisons to the previous state-of-the-art (Sec. IV-B), qualitative results (Sec. IV-C) and our additional feature analysis on the BSLCorpus [22] (Sec. IV-D).

A. Ablation Study

To understand the contribution of each design choice, we perform ablations on three key aspects: the input features, the feature-specific modules, and the mixer module. First, we evaluate the model using the Angle and HaMeR features individually, then as a fused multi-modal representation. As shown in Tab. I, our initial hypothesis that fusing HaMeR and Angle features would be the most effective is confirmed on the DGS Corpus [15]. While each feature performs well independently, their fusion allows the model to learn both the local hand shape information from HaMeR and the global body-hand context from the Angle features.

Additionally, the inclusion of feature-specific modules also improves performance significantly. This is likely because the modules function like adapters for the frozen feature extractors, learning representations that are tailored for sign segmentation. Finally, the mixer module's introduction also yields marginal gains, demonstrating its role in capturing interactions between modalities. It is worth noting that re-placing the MLP-based mixer with a cross-attention module

TABLE II

COMPARISONS TO STATE-OF-THE-ART ON DGS CORPUS [15]

	Sign-level		
Method	F1	IoU	%
Bi-LSTM + 3D Pose [16]	0.63	0.69	1.11
Bi-LSTM + 3D Pose + Hand Norm [16]	0.59	0.63	1.13
Ours	**0.86**	**0.76**	**0.98**

TABLE III

FEATURE COMPARISONS TO STATE-OF-THE-ART ON BSLCORPUS [7]

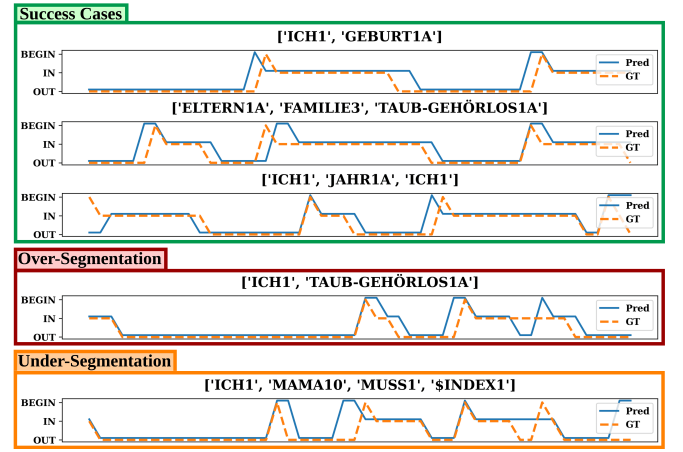| Method	mF1B	mF1S
MS-TCN + I3D [18]	68.68	47.71
MS-TCN + 3D Angles	73.09	47.37
MS-TCN + HaMeR	76.22	49.56
MS-TCN + HaMeR + 3D Angles	75.40	50.18

Fig. 3. **Qualitative Analysis:** We provide the qualitative results of our model on the DGS Corpus [15], showing the success cases (green), alongside the failure cases of over-segmentation (red) and under-segmentation (orange). Over-segmentation are instances where the model predicts a higher number of sign segments than the ground truth, while under-segmentation occurs when fewer segments are predicted than are actually present.

led to reduced performance. This performance degradation may stem from the disparity between body pose and hand shape features, causing cross-attention to introduce unnecessary complexity and hindering learning.

B. Comparisons to State-of-the-Art

We compare our model to the current state-of-the-art approach [16] on the DGS Corpus [15], which employs MediaPipe Holistic's 3D pose as input to a Bi-LSTM. As shown in Tab. II, our model performs significantly better across all evaluation metrics. Our higher F1 score shows that we have superior frame-level sign segmentation, while the improved IoU highlights our stronger segment localizations. Additionally, our model's score on the % metric also approaches the ideal value of 1.00, indicating its ability to avoid excessive over and under segmentation. Furthermore, unlike [16], who reported performance degradation when incorporating 3D hand normalizations as hand shape features, our ablation results (Tab. I) reveal that HaMeR outperforms pose-based features, showcasing its effectiveness as a robust hand shape representation for sign language tasks.

C. Qualitative Results

In Fig. 3, we present qualitative results, illustrating both success and failure cases. In success cases, signs are segmented accurately, albeit with minor frame-level deviations. For instance, while the ground truth annotates a single frame as the B tag, the model occasionally extends this annotation to 2–3 frames. Slight deviations may also occur at the boundaries where segments start and end. In failure cases, we note that over-segmentation typically occurs in long signs that involve hand shape changes or pauses during signing, as these are cues the model misinterprets as transitions between signs. Meanwhile, under-segmentation occasionally arises when consecutive signs share similar hand shapes, particularly when the latter sign is considerably shorter.

D. Comparisons of Features

We further demonstrate the effectiveness of HaMeR and 3D Angle features for alternative sign segmentation tasks by evaluating them within the state-of-the-art framework introduced in [18] on the BSLCorpus [7]. The original approach leverages an MS-TCN model trained on I3D features; however, its task formulation and evaluation metrics differ significantly from ours. Thus, to ensure a direct comparison, we retrain their model using our proposed features. As shown in Tab. III, HaMeR features consistently outperform I3D features, both individually and when combined with 3D Angles. Notably, HaMeR achieves superior performance despite having a feature dimensionality that is four times smaller than the I3D, demonstrating its computational efficiency and suitability for sign language tasks.

Furthermore, HaMeR and 3D Angle features exhibit greater robustness to variations in video quality and differences across human subjects. This robustness arises from their reliance on hand shapes and body poses, which are less influenced by variations in RGB pixel values compared to I3D features, making them more invariant to noise introduced by lower-quality video data and inter-subject variability.

V. CONCLUSION

In this work, we improved the performance of sign segmentation by proposing a transformer-based architecture trained on HaMeR and 3D skeleton angles. Extensive evaluations demonstrated that our approach outperformed state-of-the-art on the DGS Corpus. Additionally, we showed that HaMeR features generalize well to alternative sign segmentation tasks by surpassing prior benchmarks on the BSLCorpus. Future work will focus on integrating segmentation into end-to-end pipelines for sign language translation.

ACKNOWLEDGMENTS

views and the funders are not responsible for any use that may be made of the information it contains.

ETHICAL IMPACT STATEMENT

This research focuses on developing sign language segmentation models, which hold significant potential for advancing accessibility technologies for Deaf communities. However, it is critical to address the potential ethical concerns and risks associated with this work.

The primary risk arises from the use of publicly available datasets, namely the DGS Corpus and BSLCorpus, which include extensive video recordings of multiple signing participants. While these datasets were created with informed consent and released explicitly for research purposes, privacy concerns still remain paramount. Thus, to mitigate these risks, we strictly adhere to the licensing guidelines, ensure responsible handling of data, and avoid presenting participants in any negative or offensive manner.

Another concern is the potential for bias in the model due to the demographic imbalances in the datasets. While the DGS Corpus and BSLCorpus include diverse participants in terms of age and gender, they are predominantly European, with German and British participants, respectively. This therefore leads to under-representation of other racial and cultural groups. Unfortunately, these are the only large-scale datasets publicly available with the frame-level gloss annotations necessary for sign segmentation. However, this concern is addressed by our model's design, as it has been developed to be invariant to demographic biases. Specifically, it operates on features such as hand shape and body pose skeletons, which are independent of skin tone or facial characteristics, ensuring robustness across diverse populations.

This research brings huge societal benefits, including improved tools for sign language annotation and translation, which directly contribute to the Deaf community. Furthermore, as the datasets involved data contributed by Deaf individuals, this work directly builds upon resources provided by the community itself, ensuring relevance and inclusivity.

REFERENCES

[1] H. Brock and K. Nakadai. Deep JSLC: A multimodal corpus collection for data-driven generation of Japanese Sign Language expressions. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[2] P. Buehler, A. Zisserman, and M. Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2968, 2009.

[3] H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman. Aligning subtitles in sign language videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11532–11541, 2021.

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[5] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak. Two-stream network for sign language recognition and translation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[6] H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2568–2574, 2009.

[7] K. Cormier, J. Fenlon, S. Gulamani, and S. Smith. Bsl corpus annotation conventions. 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] I. Farag and H. Brock. Learning motion disfluencies for automatic sign language segmentation. *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7360–7364, 2019.

[10] Y. A. Farha and J. Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3579, 2019.

[11] J. Fenlon, T. Denmark, R. Campbell, and B. Woll. Seeing sentence boundaries. *Sign Language & Linguistics*, 10:177–200, 09 2008.

[12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, ICML '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery.

[13] M. Ivashechkin, O. Mendez, and R. Bowden. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023.

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

[15] R. Konrad, T. Hanke, G. Langer, D. Blanck, J. Bleicken, I. Hofmann, O. Jeziorski, L. König, S. König, R. Nishio, A. Regen, U. Salden, S. Wagner, S. Worseck, O. Böse, E. Jahn, and M. Schulder. Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release, 2020.

[16] A. Moryossef, Z. Jiang, M. Müller, S. Ebling, and Y. Goldberg. Linguistically motivated sign language segmentation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[17] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.

[18] K. Renz, N. C. Stache, S. Albanie, and G. Varol. Sign language segmentation with temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139, 2021.

[19] K. Renz, N. C. Stache, N. Fox, G. Varol, and S. Albanie. Sign segmentation with changepoint-modulated pseudo-labelling, 2021.

[20] P. Santemiz, O. Aran, M. Saraclar, and L. Akarun. Automatic sign segmentation from continuous signing via multiple sequence alignment. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 2001–2008, 2009.

[21] B. Saunders, N. C. Camgoz, and R. Bowden. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] A. C. Schembri, J. B. Fenlon, R. Rentelis, and K. Cormier. British sign language corpus project: A corpus of digital video data and annotations of british sign language 2008-2014. 2014.

[23] Sense. What is sign language? Accessed: December 17, 2024.

[24] SignBank.org. Signpuddle online. Accessed: December 22, 2024.

[25] R. C. Wong, N. C. Camgöz, and R. Bowden. Sign2gpt: leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations (ICLR)*.

[26] U. Zeshan. Towards a notion of 'word' in sign languages. 2002.

[27] B. Zhang, M. Müller, and R. Sennrich. Sltunet: A simple unified model for sign language translation, 2023.

[28] B. Zhou, Z. Chen, A. Clap'es, J. Wan, Y. Liang, S. Escalera, Z. Lei, and D. Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20814–20824, 2023.