

SIGN LANGUAGE SEGMENTATION WITH TEMPORAL CONVOLUTIONAL NETWORKS

Katrin Renz^{1,2} Nicolaj C. Stache² Samuel Albanie¹ Gül Varol^{1,3}

¹ Visual Geometry Group, University of Oxford, UK

² University of Heilbronn, Germany

³ LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France

<https://www.robots.ox.ac.uk/~vgg/research/signsegmentation/>

ABSTRACT

The objective of this work is to determine the location of temporal boundaries between signs in continuous sign language videos. Our approach employs 3D convolutional neural network representations with iterative temporal segment refinement to resolve ambiguities between sign boundary cues. We demonstrate the effectiveness of our approach on the BSLCORPUS, PHOENIX14 and BSL-1K datasets, showing considerable improvement over the state of the art and the ability to generalise to new signers, languages and domains.

Index Terms— Sign Language, Temporal Segmentation

1. INTRODUCTION

Sign languages are languages that have evolved among deaf communities that employ movements of the face, body and hands to convey meaning [32]. Despite significant recent progress in neural machine translation for spoken languages [34] and fine-grained visual action recognition [29], automatic recognition and translation of sign languages remains far from human performance [19]. A key challenge in closing this gap is the prohibitive annotation cost of constructing high-quality labelled sign language corpora, which are consequently orders of magnitude smaller than their counterparts in other domains such as speech recognition [3]. The high annotation cost is driven by: (1) the limited supply of annotators who possess the skill to annotate the data, and (2) the laborious time-per-annotation (taking 100 hours to densely label 1 hour of signing content [11]).

Motivated by these challenges, the focus of this work is to propose an automatic sign segmentation model that can identify the locations of temporal boundaries between signs in continuous sign language (see Fig. 1). This task, which has received limited attention in the literature, has the potential to significantly reduce annotation cost and expedite novel corpora collection efforts. Key challenges for such an automatic segmentation tool include the fast speed of continuous signing and the presence of motion blur, especially around hands.

We make the following contributions: (1) We demonstrate the effectiveness of coupling robust 3D spatio-temporal convolutional neural network (CNN) representations with an iter-

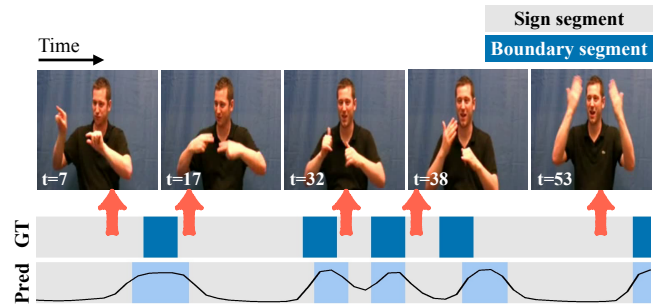


Fig. 1: Task: We illustrate the task of temporal sign segmentation for an example continuous sign language sequence from BSLCORPUS. Ground truth (GT) and predictions of our model (Pred) are shown, together with a sample frame and its frame number from each sign segment (whose location is denoted with a red arrow). Despite the fast transitions between signs, our model is able to accurately detect boundaries.

ative 1D temporal CNN refinement module to produce accurate sign boundary predictions; (2) we provide comprehensive experiments to study different components of our method. (3) We contribute a test set of human-annotated British Sign Language (BSL) temporal sign segmentation labels for a portion of BSL-1K to provide a benchmark for future work. (4) We show that our approach strongly outperforms prior work on the BSLCORPUS and BSL-1K datasets and investigate its cross-lingual generalisation on PHOENIX14.

2. RELATED WORK

The linguistic definition of sign boundaries has been non-trivial in prior work [4]. The research of [2, 14] showed that the same signs were annotated differently across teams. Therefore, several works have attempted standardising the definition of sign boundaries [9, 10, 15]. The annotation guidelines of BSLCORPUS [9], which we follow, define the position of a boundary at the time when the hands start moving away from the previous sign, an event typically indicated by a change in direction, orientation or handshape.

Although the task of automatically identifying temporal boundaries between signs has received attention in the literature, it has typically been tackled with methods that require



BSLCORPUS BSL-1K PHOENIX14

Fig. 2: Datasets: We provide samples from each dataset we use in this work: BSLCORPUS [27, 28], BSL-1K [1], PHOENIX14 [20].

access to a semantic labelling of the signed content (e.g., in the form of glosses or free-form sentence translations) [18, 21, 26]. As we show in Sec. 4, our approach can operate effectively with access to only category-agnostic annotation on the domain of interest. Other works tackled segmenting sign language content into sentence-like units [6], identifying whether a person is signing or not [25], or segmenting signs given subtitles [8]. Unsupervised sign sub-unit segmentation has also been explored [33].

Recently, [12] proposed to tackle the category-agnostic sign boundary segmentation problem using a random forest in combination with geometric features computed from 3D skeletal information obtained via motion capture. They demonstrate their approach on a small-scale Japanese Sign Language (JSL) dataset [5]—we compare our approach with theirs in Sec. 4.3.

3. SIGN SEGMENTATION

Problem formulation. Given a sequence of video frames of continuous signing, $\mathbf{x} = (x_1, \dots, x_N)$, the objective of sign language segmentation is to predict a corresponding vector of labels $\mathbf{y} = (y_1, \dots, y_N) \in \{0, 1\}^N$, where label values of 0 and 1 denote the interior of a sign segment (or “token”) and boundaries between sign segments, respectively. In the sign language corpus construction literature, several different approaches have been used to define the precise extent of a sign token [15]. In this work, we follow the set of conventions for parsing signs prescribed by [9].

Training. Motivated by its effectiveness for human action recognition and recently for sign language recognition [1, 16, 22] and sign spotting [24], our approach adopts the spatio-temporal convolutional I3D architecture [7] and couples it with the Multi-Stage Temporal Convolutional Network (MS-TCN) module proposed by [13]. Each stage of the latter—known as a Single-Stage TCN (SS-TCN)—comprises a stack of dilated residual layers of 1D temporal convolutions, followed by a linear classifier which is used to predict frame-level labels, (y_1, \dots, y_N) . Each SS-TCN is run sequentially, such that the predictions of one stage are used as input to the following stage, refining the segmentations. Our model is trained to perform frame-level binary classification with a cross-entropy loss, together with a smoothing truncated mean squared loss (following the formulation described in [13]) to reduce over-segmentation errors. For constructing the ground truth, we assign a video frame as boundary (i.e., $y = 1$) if it is at the start or end time of the sign segment. In addition, any frames between the end of one sign and the start of the

	train	val	test
avg #frames per sign	11.3 \pm 8.5	11.3 \pm 7.9	11.5 \pm 8.6
avg #frames per video	80.8 \pm 33.3	81.2 \pm 33.3	80.8 \pm 32.8
avg #glosses per video	6.8 \pm 1.8	7.1 \pm 1.9	6.9 \pm 1.8
total #videos	5413	763	703
total #signers	157	20	21
total #unique glosses	969	671	620

Table 1: Statistics for our signer-independent split of a subset of BSLCORPUS for which dense gloss annotations are available.

next sign are also assigned as boundary. In Sec. 4, we conduct an empirical evaluation of a number of variants of this model, including: the number of stages and the importance of finetuning the I3D backbone.

4. EXPERIMENTS

In this section, we describe the datasets used in our experiments (Sec. 4.1), present our ablations to assess different components of our approach (Sec. 4.2), compare to prior work (Sec. 4.3), test the generalisation capability on different datasets (Sec. 4.4), and provide qualitative results (Sec. 4.5).

4.1. Datasets and evaluation metrics

BSLCORPUS [27, 28] is a BSL linguistic corpus that provides various types of manual annotations, of which approximately 72K are gloss annotations, i.e., individual signs with their sign categories and temporal boundaries. We use the subset of videos where such gloss annotation is available, and we split it into train/validation/test sets as in Tab. 1. The gloss annotations contain separate labels for the right and left hand, which we merge with priority for the dominant hand. We define sign categories according to several rules, e.g., assigning lexical variants of the same word to one class, filtering classes with less than 10 occurrences. Resulting annotations are used to cut the video into shorter clips with at least three consecutive signs. We use this data for training and evaluation.

BSL-1K [1] is a recently collected large-scale dataset of BSL signs for which sparse annotations are obtained using a mouthing-based visual keyword spotting model [23, 31]. This dataset does not provide precise start/end times of signs, but rather an approximate position of the sign. We run our segmentation model to complement the dataset with automatic temporal annotations, which we quantitatively evaluate on a small manually annotated subset.

RWTH-PHOENIX-Weather-2014 [20] (PHOENIX14) is a standard benchmark in computer vision with dense gloss annotations without timings for German Sign Language (DGS) signs. The work of [21] provides automatically generated alignments for the training set, making use of ground-truth gloss information, against which we compare our boundary estimations. Since automatic timing annotations are not available on the validation or test sets, we randomly partition the training set into training and test splits following a 4:1 ratio.

	mF1B	mF1S
Uniform baseline (using GT #signs)	41.37	34.89
BSL-1K [1]	58.48 \pm 1.4	33.66 \pm 2.3
BSLCORPUS (class-labels)	68.68\pm0.6	47.71\pm0.8
BSL-1K \rightarrow BSLCORPUS (class-labels)	66.17 \pm 0.5	44.44 \pm 1.0
BSL-1K \rightarrow BSLCORPUS (class-agnostic)	68.23 \pm 1.8	44.36 \pm 3.3

Table 2: The influence of I3D training data: We observe that finetuning the BSL-1K classification model [1] on BSLCORPUS brings a significant boost in performance. However, the proposed method does not require category information on BSLCORPUS to be effective—class-agnostic training suffices.

Evaluation metrics. To assess the sign segmentation performance, we measure both the capability to predict the position of the boundary and the extent of the sign segments. We define a boundary prediction to be correct, if its distance to a ground-truth boundary is lower than a given threshold. The distance is measured from the mean position of the predicted and the ground-truth boundary, wherein a boundary refers to a series of contiguous 1s in y . We calculate the F1 score for the boundaries as the harmonic mean of precision and recall, given this definition of a correct boundary detection. We use all integer-valued thresholds that fall within the closed interval [1, 4] and report the mean across thresholds, which we refer to as mF1B. The quality of the sign segments is also evaluated with the F1 score, where sign segments with an IoU higher than a given threshold are defined as correct. We average the results for thresholds from 0.4 to 0.75 with a step size of 0.05 and report the result as mF1S. We conduct each experiment with three different random seeds and report the mean and standard deviation. For further interpretability of the results, we provide additional metrics, such as the mean boundary width, on our project page [30].

4.2. Ablation study

We first perform several ablations on BSLCORPUS.

Uniform baseline. We measure the performance for a simple baseline that uniformly splits a video into temporal segments given the true number of signs. Note that automatically counting the signs occurring in a video is an unsolved problem, therefore this baseline does *not* represent a lower-bound, but is instead used to provide indicative results for the metrics. As can be seen in Tab. 2, such a uniform baseline is suboptimal (41.37 mF1B) even if it uses ground-truth information.

Sign recognition pretraining on the target data. Next, we compare the performance when using three different versions of input features for MS-TCN training: pretraining the I3D model (i) on BSL-1K (ii) on BSLCORPUS and (iii) on BSL-1K and finetuning on BSLCORPUS. In all cases, I3D is initially pretrained on Kinetics [17]. For (iii), we consider two variants: (1) finetuning using semantic *class-labels* for recognition and (2) *class-agnostic* finetuning in which the model is trained directly for class-agnostic boundary classification (and does not make use of the sign labels themselves). Tab. 2 summarises the results of retraining MS-TCN with each of these I3D input features. Strikingly, while there is a clear ben-

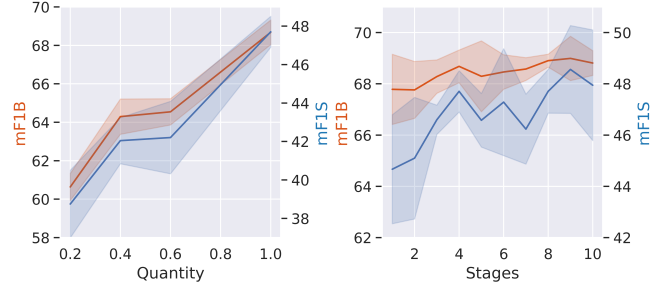


Fig. 3: Ablations: We study (left) the quantity of training data available to the model, and (right) the number of refinement stages in the MS-TCN architecture.

efit to finetuning on BSLCORPUS, the model does not require the pretraining step on BSL-1K. Furthermore, class labels for the I3D training are not essential to achieving good segmentation performance.

Quantity of training data. To investigate whether the quantity of available training data represents a limiting factor for model performance, we plot segmentation metrics against the quantity of data used for training both the I3D and the MS-TCN models. As can be seen in Fig. 3, training data availability represents a significant bottleneck. Exploring the use of automatically segmented videos, which is made possible with our approach, could be a way to mitigate the lack of training data. We leave this to future work.

Number of refinement stages. We next ablate the MS-TCN architecture by changing the number of refinement stages. The results in Fig. 3 suggest that one and two stages exhibit inferior performance, likely due to lacking sufficient access to context. The model has diminishing benefits from the addition of a large number of stages. In the rest of our experiments, we employ a total of four stages as in [13].

4.3. Comparison to prior work

In this section, we compare our model with the method introduced in [12], which uses hand-crafted geometric features computed on 3D body pose keypoints in combination with a Random Forest classifier. Since the DJS LC dataset [5] used by [12] is not publicly available to facilitate a comparison, we turn to the publicly available BSLCORPUS with boundary annotations and compare to our re-implementation of their approach. In contrast to [12], which assumes 3D skeletal information given by motion capture, we estimate the 3D pose coordinates with the recently proposed monocular DOPE model [35]. While the performance is bounded by the quality of the pose estimation, this ensures that the method is applicable to unconstrained sign language videos. For the geometric features of [12], we calculate angular and distance measurements for different joint pair combinations. We compute the Laplacian kernel matrix as described in [12], and concatenate the flattened upper triangle of this matrix with the raw geometric features for a given window. We refer the reader to [12] for further details. To identify the respective influence of the features and the classifiers, in Tab. 3, we re-

Method	mF1B	mF1S
Geometric features + RF [12]	50.49 \pm 0.1	37.46 \pm 0.1
Geometric features + MS-TCN	60.77 \pm 2.3	36.25 \pm 2.0
I3D + MS-TCN (proposed)	68.68\pm0.6	47.71\pm0.8

Table 3: State of the art comparison: We compare to the geometric features proposed by [12] on **BSLCORPUS**. We show that our I3D features, combined with the MS-TCN segmentation model, significantly outperforms [12] which uses Random Forest (RF) classifiers.

Method	mF1B	mF1S
Geometric features + RF [12]	51.26 \pm 0.5	34.28 \pm 1.0
I3D + MS-TCN	61.12\pm0.9	49.96\pm0.6

Table 4: Generalisation to BSL-1K: We report the results of applying the BSLCORPUS-trained segmentation model on a small fraction of BSL-1K which we manually annotated.

port the performance of the geometric features [12] with both Random Forest and MS-TCN classifiers. A first improvement over [12] can be attributed to the use of the MS-TCN model. We further significantly improve over the geometric features with our I3D finetuning on BSLCORPUS.

4.4. Generalisation to other sign language datasets

Here, we evaluate the capability of our method to generalise on different datasets.

BSL-1K. We apply our segmentation model on the BSL-1K dataset and complement the sparse sign annotations provided by [1] with timing information for start and end frames. While both our training data BSLCORPUS and test data BSL-1K include the same sign language, there is a considerable domain gap in appearance as shown in Fig. 2. To enable a quantitative performance measure, we manually annotated a 2-minute sequence with exhaustive segmentation labels, resulting in 177 sign segments. In Tab. 4, we observe a trend consistent with the previous experiments: the proposed approach outperforms the prior work of [12] by a sizeable margin.

PHOENIX14. Next, we test the limits of our approach on German Sign Language (DGS). Note that the timing annotations are noisy due to automatic forced-alignment [21]. Due to the challenging domain gap both in the visual appearance and in the sign languages (BSL and DGS), we obtain a limited but reasonable cross-lingual generalisation performance (see Tab. 5), suggesting that the model learns to exploit some common visual cues for segmenting the different sign languages. These common visual cues contain for example significant changes in direction or orientation of the hands.

4.5. Qualitative analysis

In Fig. 4, we provide qualitative results on all three datasets: BSLCORPUS, BSL-1K, and PHOENIX14. The examples enclosed in the orange box show two different failure categories on BSLCORPUS. In one case, our model over-segments a fingerspelled word by predicting the boundaries of individual letters. We also observe that the model has difficulty with

I3D training data	MS-TCN training data	mF1B	mF1S
BSLCORPUS	BSLCORPUS	46.75 \pm 1.2	32.29 \pm 0.3
BSLCORPUS	PHOENIX14	65.06 \pm 0.5	44.42 \pm 2.0
PHOENIX14	PHOENIX14	71.50 \pm 0.2	52.78 \pm 1.6

Table 5: Generalisation to PHOENIX14 German Sign Language: We evaluate our method against the automatic labels provided in PHOENIX14. The model which is only trained on BSLCORPUS shows limited generalisation (46.75 mF1B) to a different sign language. The control experiment provides an approximate upper bound by training both the I3D and MS-TCN models using the PHOENIX14 automatic labels (71.50 mF1B) .

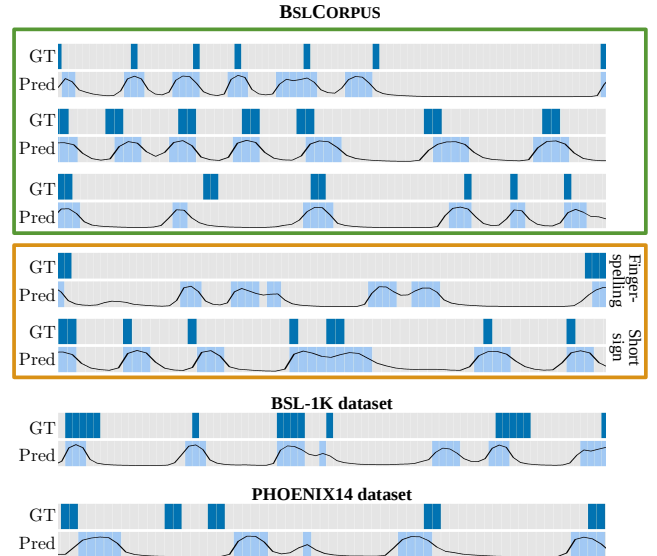


Fig. 4: Qualitative analysis: Results on BSLCORPUS (top), BSL-1K and PHOENIX14 (bottom) datasets. We illustrate success (green) and failure cases (orange) on BSLCORPUS, with the latter arising from fingerspelling and a very short sign.

very short signs. For more qualitative analysis and the corresponding videos, we refer the reader to our project page [30].

5. CONCLUSION

In this paper we addressed the problem of temporal sign language segmentation. We employed temporal convolutions and formulated the problem as sign boundary detection. We provided a comprehensive study to analyse our various components. We further reported the results on three datasets with varying properties showcasing our method’s generalisation capabilities. Future directions include extending the amount of training data by exploring automatic annotations.

Acknowledgements: This work was supported by EPSRC grant ExTol. KR was supported by the German Academic Scholarship Foundation. The authors would like to express their gratitude to C. Camgoz for the help with the BSLCORPUS data preparation, N. Fox for his invaluable assistance in providing an annotation sequence for evaluation, L. Momeni and A. Braffort for valuable feedback. SA would like to acknowledge the support of Z. Novak and S. Carlson in enabling his contribution to this research.

References

- [1] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, “BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues,” in *ECCV*, 2020. 2, 3, 4
- [2] A. Braffort and L. Boutora, “DEGELS2012 annotation challenge: Segmentation [in French],” in *JEP-TALN-RECITAL, DEGELS: Gestures and Sign Language Challenge*, 2012. 1
- [3] D. Bragg, O. Koller *et al.*, “Sign language recognition, generation, and translation: An interdisciplinary perspective,” in *ACM SIGACCESS*, 2019. 1
- [4] D. Brentari, “Effects of language modality on word segmentation: An experimental study of phonological factors in a sign language,” *Papers in laboratory phonology*, vol. 8, pp. 155–164, 2009. 1
- [5] H. Brock and K. Nakadai, “Deep JSLC: A multimodal corpus collection for data-driven generation of Japanese Sign Language expressions,” in *LREC*, 2018. 2, 3
- [6] H. Bull, M. Gouiffès, and A. Braffort, “Automatic segmentation of sign language into subtitle-units,” in *ECCVW (SLRTP)*, 2020. 2
- [7] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in *CVPR*, 2017. 2
- [8] H. Cooper and R. Bowden, “Learning signs from subtitles: A weakly supervised approach to sign language recognition,” *CVPR*, 2009. 2
- [9] K. Cormier and J. Fenlon, “BSL corpus annotation guidelines,” 2014. 1, 2
- [10] O. Crasborn and I. Zwitterlood, “Annotation of the video data in the corpus NGT,” *Radboud University Nijmegen*, 2008. 1
- [11] P. Dreuw and H. Ney, “Towards automatic sign language annotation for the ELAN tool,” in *LREC Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 2008. 1
- [12] I. Farag and H. Brock, “Learning motion disfluencies for automatic sign language segmentation,” in *ICASSP*, 2019. 2, 3, 4
- [13] Y. A. Farha and J. Gall, “MS-TCN: Multi-stage temporal convolutional network for action segmentation,” in *CVPR*, 2019. 2, 3
- [14] M. Gonzalez, “Computer vision methods for unconstrained gesture recognition in the context of sign language annotation,” Ph.D. dissertation, Université de Toulouse, 2012. 1
- [15] T. Hanke, S. Matthes, A. Regen, and S. Wörseck, “Where does a sign start and end? segmentation of continuous signing,” in *LREC Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, 2012. 1, 2
- [16] H. R. V. Joze and O. Koller, “MS-ASL: A large-scale data set and benchmark for understanding american sign language,” in *BMVC*, 2019. 2
- [17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The Kinetics human action video dataset,” *arXiv*, 2017. 3
- [18] O. Koller, N. C. Camgoz, H. Ney, and 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*, 2020. 2
- [19] O. Koller, “Quantitative survey of the state of the art in sign language recognition,” *arXiv:2008.09918*, 2020. 1
- [20] 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, 2015. 2
- [21] O. Koller, S. Zargaran, and H. Ney, “Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs,” in *CVPR*, 2017. 2, 4
- [22] D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in *WACV*, 2020. 2
- [23] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman, “Seeing wake words: Audio-visual keyword spotting,” *BMVC*, 2020. 2
- [24] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman, “Watch, read and lookup: learning to spot signs from multiple supervisors,” in *ACCV*, 2020. 2
- [25] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, “Real-Time Sign Language Detection using Human Pose Estimation,” in *ECCVW, Sign Language Recognition, Translation and Production (SLRTP)*, 2020. 2
- [26] P. Santemiz, O. Aran, M. Saraclar, and L. Akarun, “Automatic sign segmentation from continuous signing via multiple sequence alignment,” in *ICCVW*, 2009. 2
- [27] A. Schembri, J. Fenlon, R. Rentelis, and K. Cormier, “British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition),” 2017. [Online]. Available: <http://www.bslcorpusproject.org> 2
- [28] A. Schembri, J. Fenlon, R. Rentelis, S. Reynolds, and K. Cormier, “Building the British Sign Language Corpus,” *Language Documentation & Conservation*, vol. 7, pp. 136–154, 2013. 2
- [29] D. Shao, Y. Zhao, B. Dai, and D. Lin, “FineGym: A hierarchical video dataset for fine-grained action understanding,” in *CVPR*, 2020. 1
- [30] “Sign segmentation project page,” <https://www.robots.ox.ac.uk/~vgg/research/signsegmentation/>. 3, 4
- [31] T. Stafylakis and G. Tzimiropoulos, “Zero-shot keyword spotting for visual speech recognition in-the-wild,” in *ECCV*, 2018. 2
- [32] R. Sutton-Spence and B. Woll, *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press, 1999. 1
- [33] S. Theodorakis, V. Pitsikalis, and P. Maragos, “Dynamic-static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition,” *Image and Vision Computing*, vol. 32, no. 8, pp. 533–549, 2014. 2
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017. 1
- [35] P. Weinzaepfel, R. Brégier, H. Combaluzier, V. Leroy, and G. Rogez, “DOPE: Distillation of part experts for whole-body 3D pose estimation in the wild,” in *ECCV*, 2020. 3