Topic Detection in Continuous Sign Language Videos

Álvaro Budria[1] Laia Tarrés[1] Gerard I. Gállego[1]
Francesc Moreno-Noguer[2] Jordi Torres[1,3] Xavier Giró-i-Nieto[1,2]
[1]*Universitat Politècnica de Catalunya* [2]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC* [3]*Barcelona Supercomputing Center*

Abstract

Significant progress has been made recently on challenging tasks in automatic sign language understanding, such as sign language recognition, translation and production. However, these works have focused on datasets with relatively few samples, short recordings and limited vocabulary and signing space. In this work, we introduce the novel task of sign language topic detection. We base our experiments on How2Sign [13], a large-scale video dataset spanning multiple semantic domains. We provide strong baselines for the task of topic detection, and present a comparison between different visual features commonly used in the domain of sign language.

1. Introduction

Sign languages are the native languages and primary means of communication for millions of Deaf and hard-of-hearing people worldwide. Sign languages utilize multiple complementary channels to convey information, including manual features, such as shape, movement and pose, as well as non-manual features, such as facial expressions and movement of head, shoulders and torso.

Tasks of diverse complexity have been addressed in the literature: from the simpler sign language recognition [2,11,14,17,24,27,29] over isolated signs, to the much more challenging ones of sign language translation [5,6,9,19] and production [32–36]. While some methods for translation and production have shown very good results on smaller datasets [5, 19, 20, 22, 38], they have not been proven to produce satisfactory results yet on larger ones containing a wider signing space with longer video sequences.

In this work, we propose the novel task of sign language topic detection, that is, classifying sign language video recordings into one of several categories, as depicted in Figure 1. This task has been broadly explored for spoken languages [25], but not for sign languages.

We believe our work on topic detection in sign language videos could help in the design of more inclusive online experiences for the Deaf and hard-of-hearing. We tackle topic detection with three different neural architectures and four different kinds of visual features, and evaluate their strengths and shortcomings through a set of experiments.

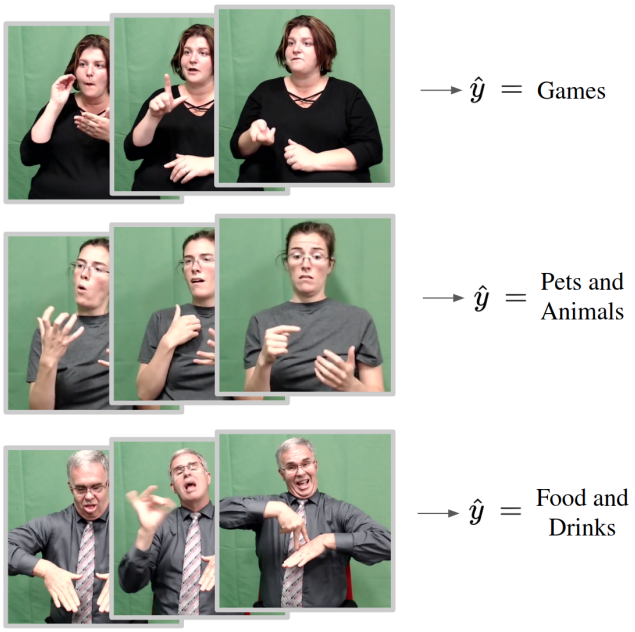

Figure 1. Topic detection in sign language videos is the task of producing a label that describes the semantic content of a signer's discourse.

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, we provide the first study of sign language topic detection.
- We thoroughly measure the performance of three deep learning architectures (LSTM [15], Transformer [39] and PerceiverIO [16]) in combination with four different visual features that are commonly employed for sign language understanding (3D Cartesian body poses, 3D angular body poses, I3D features and sign-spotting annotations).

2. Related Work

In this paper, we address the task of Sign Language (SL) topic detection, which we define as the task of producing a label that semantically describes the content of the discourse being signed, given a sequence of frames.

SL recognition is perhaps the closest task in the SL literature to that of topic detection. The aim in SL recognition is to tell which sign is being represented, given a short video of a signer producing either an isolated sign or a continuous sequence of signs [9, 20].

The current state-of-the-art in SL recognition is characterized by complex modeling pipelines involving distillation [14], graph neural networks [17], auxiliary losses [24], stochastic labeling [27], or cross-modal alignment [29]. In this work, however, we choose instead simpler pipelines that can act as robust baselines for future work on SL topic detection.

Outside the domain of SL, general video classification and action recognition is the most similar task to SL topic detection. Several methods have been proposed for generic video classification [1, 4, 10, 30, 41–43]. Despite having obtained remarkable results, they are generally unsuitable for the task of topic detection we address here. Their computational requirements are often untenable, due to being designed for dealing with shorter videos of at most a few hundred frames, while SL videos may contain thousands of frames.

3. Methodology

3.1. Dataset

We base our experiments on the recent How2Sign dataset. How2Sign [13] is a large-scale collection of multimodal and multiview SL videos in American Sign Language (ASL) covering over 2500 instructional videos selected from the preexisting How2 dataset [31].

How2Sign consists of more than 80 hours of ASL videos, with sentence-level alignment for more than 35k sentences. It features a vocabulary of 16k English words that represent more than two thousand instructional videos from a broad range of categories. The dataset comes with a rich set of annotations including category labels, text annotations, as well automatically extracted 2D body poses for more than 6M frames.

To the best of the authors' knowledge, How2Sign is currently the only SL dataset containing manually produced per-video category annotations semantically describing a video's content. OpenASL [37], a recent large-scale ASL dataset, features 288 hours of SL video with speech transcriptions, but does not include category labels. Other datasets, such as [3, 5] are restricted to a single topic or semantic domain. Others, like [38], just contain videos of isolated signs, rendering them unsuitable for SL video classification. In this work, we leverage the topic annotations provided at video level (Fig. 2). Each video is associated with one of 10 target labels describing its content, and our aim is to classify videos in their corresponding category.

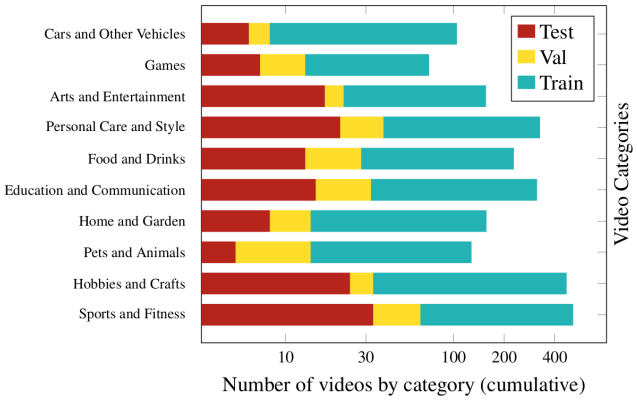

Figure 2. Cumulative topic distribution in the How2Sign dataset (figure from [13])

3.2. Video Features

In our experiments, we train models with five different kinds of data: 3D poses represented by either Cartesian coordinates or joint angles, I3D features [8], sign-spotting annotations obtained with Spot-Align [12], and speech transcriptions. Next, we briefly present each of them.

3D Cartesian poses. Alongside videos, How2Sign also provides body keypoint annotations extracted with OpenPose [7]. In addition, we also extract keypoints using MediaPipe [23], resulting in two sets of body pose annotations.

Figure 3. We train models on body poses extracted with two different pose detectors: MediaPipe (left) and OpenPose (right).

These keypoints provide a light-weight representation of the signer's hands and body that is invariant to the signers's and background's visual characteristics [19]. Nevertheless, this representation is sensitive to occlusions and tends to present a significant amount of noise. OpenPose and Mediapipe produce keypoints for hands, face and body, including arms and legs. We make use of the hands, upper body and arms. Since OpenPose does not produce 3D keypoints directly, we lift them to 3D as described in [44]. Finally, we vectorize the pose for each frame into a vector $v_t = (x_1, y_1, z_1, ..., x_{50}, y_{50}, z_{50})$ of size $50 \times 3 = 150$.

	LSTM	Transformer	PerceiverIO	Majority
Cartesian (OP)	30.35 ± 3.01	34.02 ± 0.33	30.34 ± 2.58	
Cartesian (MP)	29.43 ± 1.17	33.10 ± 2.58	33.56 ± 1.30	
Angular (OP)	31.95 ± 1.05	29.66 ± 0.12	31.49 ± 2.34	
Angular (MP)	32.64 ± 0.32	34.71 ± 1.42	30.80 ± 1.97	25
I3D	45.75 ± 1.59	46.26 ± 1.30	48.27 ± 0.33	
Spotted signs	$\mathbf{58.03 \pm 1.40}$	$\mathbf{53.33 \pm 2.18}$	$\mathbf{52.88 \pm 0.32}$	
Transcriptions *(upper bound)*	70.35 ± 4.50	75.38 ± 0.75	75.90 ± 2.48	

Table 1. Test accuracy obtained for each model and data type. We report the average and standard deviation over three runs. *OP* stands for OpenPose and *MP* for MediaPipe. The majority classifier's performance is reported under column "Majority".

3D angular poses. Although this 3D Cartesian representation allows handling occlusions and different camera angles much more effectively, it suffers from sensitivity to scale and length of the speaker's limbs. For this reason, we decided to follow [26] and [40], by converting the Cartesian coordinates to an angular representation [48]. In essence, this means that a vector θ_j associated with bone j encodes the relative rotation of j w.r.t its parent bones. For each frame, we vectorize the rotation vectors of each of the joints into a single array of size $48 \times 6 = 288$.

I3D features from BOBSL. As a third visual representation, we choose I3D [8] features obtained from a feature extractor pretrained on the large-scale BBC-Oxford British Sign Language Dataset (BOBSL) [3]. I3D features take into account not only visual cues but also temporal information. As a result, they provide a dense and reliable source of visual cues as input to our models. Each frame is encoded as a 1024-dimensional feature vector.

Signs spotted with SPOT-ALIGN . SPOT-ALIGN [12] is a framework for spotting signs in continuous sign language videos, that is, detecting individual signs in videos and annotating their corresponding English translation. SPOT-ALIGN alternates between repeated sign spotting (to obtain more annotations) and jointly training a visual and a textual encoder on the resulting annotations together with dictionary exemplars, to obtain better features for spotting. For each video, we train on a list of words corresponding to the signs spotted by SPOT-ALIGN . We associate to each word a trainable vector embedding of size 256.

Speech transcriptions. How2Sign provides English speech transcriptions (also called *English translation*) for each of its videos. These transcriptions were manually produced and originate from the How2 [31] dataset, which How2Sign is based upon. English translations were manually time-aligned at sentence-level with the How2Sign sign language videos. With the aim of obtaining a pseudo-upperbound on performance for our topic detection models, we also train them on these speech transcriptions. As with spotted signs, we embed each token into a 256-dimensional trainable vector.

3.3. Neural Architectures

We test three different architectures that stand behind some of the most notable successes in video analytics: the LSTM [15] with attention [47], the Transformer [39] and the PerceiverIO [16]. These three architectures represent different trends for processing sequential inputs. The LSTM treats samples in a sequential manner, while the Transformer and the PerceiverIO process them in parallel via self-attention. PerceiverIO is specifically designed to handle extremely long input sequences, while the Transformer scales poorly with respect to input sequence length.

LSTM with attention. LSTM is still one of the go-to architectures for dealing with sequential data. It processes an input video sequentially frame-by-frame, which allows it to scale linearly in terms of computational complexity with respect to the input length. In order to boost the performance of the LSTM, we use a bidirectional configuration, and we add an attention mechanism over the hidden states, as described in [47].

Transformer. Since its appearance [39], the Transformer has dominated the NLP landscape and has also recently become prominent on several image processing tasks [46]. The core component of the Transformer is the self-attention module which performs a comparison of each of the input tokens against the rest. One advantage of the Transformer over the LSTM, is that the Transformer allows processing input tokens in parallel in a non-sequential fashion, thus reducing training time. However, Transformer incurs in a quadratic computational cost with respect to the input length.

PerceiverIO. A recent trend in the machine learning literature is to design deep learning architectures that overcome the quadratic cost of the self-attention mechanism. One line of work has focused on projecting the inputs to a lower dimensional latent space. PerceiverIO [16] leverages a cross-attention module at the beginning of the architecture which maps an input array of length T to a latent array of length N, with $N \ll T$.

	LSTM			Transformer			PerceiverIO		
	Params.	**FLOPs**	**Ratio**	**Params.**	**FLOPs**	**Ratio**	**Params.**	**FLOPs**	**Ratio**
Cartesian (OP)	2.6M	49.1B	$5.3 \cdot 10^{-5}$	9.8M	10.1B	$9.7 \cdot 10^{-4}$	15M	7.85B	$1.9 \cdot 10^{-3}$
Cartesian (MP)	2.6M	49.1B	$5.3 \cdot 10^{-5}$	3.2M	4.03B	$7.9 \cdot 10^{-4}$	5.9M	2.60B	$2.2 \cdot 10^{-3}$
Angular (OP)	2.8M	146B	$1.9 \cdot 10^{-5}$	1.0M	2.15B	$4.7 \cdot 10^{-4}$	6.0M	3.13B	$1.9 \cdot 10^{-3}$
Angular (MP)	0.5M	23.7B	$2.1 \cdot 10^{-5}$	3.9M	7.32B	$5.3 \cdot 10^{-4}$	5.8M	3.03B	$1.9 \cdot 10^{-3}$
I3D	4.3M	167B	$2.6 \cdot 10^{-5}$	11M	12.2B	$9.0 \cdot 10^{-4}$	7.4M	3.80B	$1.9 \cdot 10^{-3}$
Spotted signs	2.5M	10.5B	$2.4 \cdot 10^{-4}$	2.9M	3.11B	$9.3 \cdot 10^{-4}$	3.1M	1.95B	$1.6 \cdot 10^{-3}$
Transcriptions	2.5M	10.5B	$2.4 \cdot 10^{-4}$	2.9M	3.11B	$9.3 \cdot 10^{-4}$	3.1M	1.95B	$1.6 \cdot 10^{-3}$

Table 2. Number of parameters and FLOPs for each model and data type. *OP* stands for OpenPose and *MP* for MediaPipe.

4. Experiments

A suite of models are trained across several architectures and feature types, with the aim of introducing strong baseline models with different characteristics for the task of sign language topic detection. All possible combinations between the different architectures and features mentioned above are explored, as depicted in Table 1. For each pair of architecture and data type, we perform a grid search in order to select the most adequate hyperparameters.

We train all of our models on a single GeForce RTX 3090 GPU. We run training until validation accuracy stops decreasing and use early stopping on validation accuracy to select the best checkpoint. As optimizer, we utilize Adam [18] with a learning rate scheduler having a decrease factor of 0.5 per 8 epochs of non-decreasing validation loss. We leave the learning rate as a hyperparameter to be determined for each model.

We use the SentencePiece [21] tokenizer with a dictionary size of 8000 for speech transcriptions and 1470 for spotted signs, and input embeddings of size 256 for all models using one of these two input types. We implement our models and training pipelines on the Fairseq[1] library [28], which runs on PyTorch and is designed to perform translation, summarization and other spoken language tasks.

In Table 1, we report average accuracy over three runs. We also report theoretical FLOPs (Table 2) to compare the models in terms of their computational demands.

As expected, we find that classifying the English translations in the form of text, which is a standard task in NLP, can be addressed with a fair amount of success with all three architectures. All models trained on speech transcriptions obtain a test accuracy of over 70%, with PerceiverIO obtaining the highest score.

Among the visual features, the signs spotted with SPOTALIGN produce the highest test accuracy across all architectures, with the LSTM beating the other two. With a slightly worse performance than spotted signs, I3D features are far more adequate than both Cartesian and rotational body poses.

[1] github.com/facebookresearch/fairseq

All body pose features yield similar results, with no clear indication that angular poses might be more suitable than Cartesian ones, or that either one of the pose extractors (OpenPose and MediaPipe) is more adequate than the other. Nevertheless, representing the signers' bodies with keypoints tends to give poorer results. We consider that this gap in performance is related to the limitations of the body pose estimators (OpenPose and MediaPipe) against the challenges presented by sign language videos, which contain fast motion and self-occlusions of the hands. Moreover, none of the three studied architectures are specifically designed for processing body pose inputs, which can be naturally described in the form of a graph, rather than an array of values, as we do in this work. It is likely that adopting architectures with more specific inductive biases, such as Graph Neural Networks [45], or adapting methods currently used in the domain of action recognition [1, 4, 10, 30, 41–43] could lead to better results when training on body poses.

Lastly, PerceiverIO exhibits the highest ratio between number of parameters and FLOPs, so in terms of computational efficiency, it outcompetes both the LSTM and the Transformer.

5. Conclusions

In this work, we present the task of sign language topic detection for the first time in the literature. We provide baseline models for topic detection in sign language videos, which will contribute to the design of more inclusive experiences for the Deaf and hard-of-hearing.

We have compared in terms of accuracy and computational efficiency four different visual features that are commonly used in the sign language understanding literature. We show that spotted signs lead to better performance than I3D features and body poses, with the latter lagging far behind the rest of features.

More work needs to be done in order to improve the quality of the keypoints provided by body pose extractors. Finally, other neural architectures with more suitable inductive biases for dealing with body pose keypoints should be explored.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 4

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*, pages 35–53. Springer, 2020. 1

[3] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset. 2021. 2, 3

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2, 4

[5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. 1, 2

[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319. Springer, 2020. 1

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 3

[9] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030, 2020. 1, 2

[10] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 4

[11] Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Isolated sign recognition from rgb video using pose flow and self-attention. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3436–3445, 2021. 1

[12] Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. Sign language video retrieval with free-form textual queries. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[13] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[14] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11283–11292, 2021. 1, 2

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 1, 3

[16] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 3

[17] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Challenge on Large Scale Signer Independent Isolated Sign Language Recognition (CVPR)*, 2021. 1, 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[19] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683, 2019. 1, 2

[20] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, Dec. 2015. 1, 2

[21] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. 4

[22] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 1

[23] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2

[24] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11542–11551, 2021. 1, 2

[25] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021. 1

[26] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11865–11874, 2021. 3

[27] Zhe Niu and Brian Mak. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 172–186, Cham, 2020. Springer International Publishing. 1, 2

[28] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 4

[29] Ilias Papastratis, Kosmas Dimitropoulos, Dimitrios Konstantinidis, and Petros Daras. Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space. *IEEE Access*, 8:91170–91180, 2020. 1, 2

[30] Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2, 4

[31] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018. 2, 3

[32] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020. 1

[33] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *International journal of computer vision*, 129(7):2113–2135, 2021. 1

[34] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929, 2021. 1

[35] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Skeletal graph self-attention: Embedding a skeleton inductive bias into sign language production. *arXiv preprint arXiv:2112.05277*, 2021. 1

[36] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. *arXiv preprint arXiv:2203.15354*, 2022. 1

[37] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*, 2022. 2

[38] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020. 1, 2

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3

[40] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 3

[41] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[42] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. UNIK: A unified framework for real-world skeleton-based action recognition. In *British Machine Vision Conference (BMVC)*, 2021. 2, 4

[43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2, 4

[44] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3384–3392, 2020. 2

[45] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. 4

[46] Ming Zhou, Nan Duan, Shujie Liu, and Heung-Yeung Shum. Progress in neural nlp: Modeling, learning, and reasoning. *Engineering*, 6(3):275–290, 2020. 3

[47] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 3

[48] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2019. 3