Appendix for Auslan-Daily: Australian Sign Language Translation for Daily Communication and News

A Building Auslan-Daily

In this section, we explain the various stages of data processing and labelling in detail for preparing Auslan-Daily, from collecting sources to storing final data.

A.1 Data Curation

Table 1: List all data sources and their official description from which we curate the Auslan-Daily dataset and their hyperlinks.

Sub-dataset	Sources	Description (from website)
Auslan-Daily Communication	*Sally and Possum* G	Sally and Possum is an innovative television series for young children who are deaf or hard of hearing and whose primary language is Auslan.
Auslan-Daily News	*ABC News with Auslan* G	The latest news and information from ABC News. This bulletin will be Auslan interpreted to provide accessible information to keep Australia's deaf community connected and informed.
	Expression Australia G	Expression Australia is a non-profit organisation established in 1884 and is the source of reference, referral, advice and support for people experiencing barriers to participation.
	Lingthusiasm G	Lingthusiasm is a podcast that's enthusiastic about linguistics as a way of understanding the world around us.

The information in [1] shows that the quantity of data available for Australian Sign Language (Auslan) is comparatively modest compared with other nations' sign languages. Furthermore, high-quality Auslan sources with subtitles are relatively scarce. Through our search, we have managed to amass Auslan data from four distinct sources, as shown in Table 1.

A.2 Subtitles Extraction

As delineated in Section 3.1, we employ three distinct operations for subtitle cleaning. Here, we present a few representative examples:

- Incomplete subtitles:
  [00:03:33.23] Why don't we watch some children visiting a farm [00:03:37.15]
  [00:03:37.15] and see how they learn how important water is? [00:03:42.00]
  Revise:
  [00:03:33.23] Why don't we watch some children visiting a farm and see how they learn how important water is? [00:03:42.00]
- Several complete subtitles that appear within a time interval:
  [00:15:24.09] Thanks for watching. See you next time! [00:15:26.04]
  Revise:
  [00:15:24.09] Thanks for watching. [00:15:26.04]
  [00:15:24.09] See you next time! [00:15:26.04]
- Complete sentence that only contains modal particles:
  [00:10:26.02] Mm-hm. Mm-hm. [00:10:29.01]
  Revise:
  Remove this subtitle.

A.3 Long-tailed Words Alignment:

In the translation task, the "long tail problem" typically refers to the translation of specific, unconventional, or uncommon expressions and terms. Since these expressions and terms are used infrequently,

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

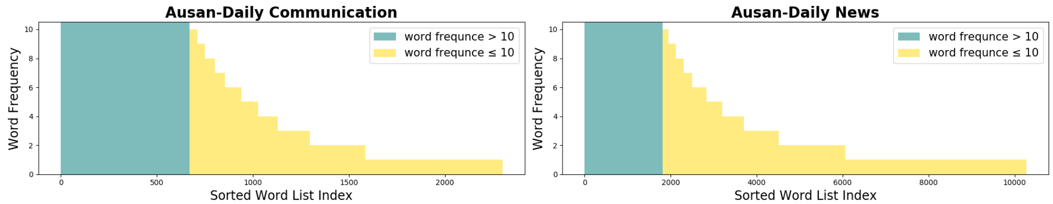

Figure 1: Words Frequency of Auslan-Daily

they are often not included in the training data of most translation systems. This makes handling these "long tail" issues quite challenging. This problem is also significant in the sign language translation dataset. Figure 1 elucidates the severity of the long-tail problem in Auslan-Daily, with the infrequency of certain words being quite pronounced. In Auslan-Daily Communication and News sub-datasets, words appearing less than or equal to 10 times constitute nearly one-third and one-fifth of the total, respectively. This phenomenon is commonly observed across current sign language datasets. For instance, in the PHOENIX-2014T [2] training set, singletons (terms with a word frequency of one) constitute one-third of the set, amounting to 1,077 out of 2,887 total words. Consequently, we solicit expert assistance to annotate the long-tail words exhibiting a word frequency in the range of two to ten. We intend to facilitate further comprehensive research about sign language translation.

A.4 Pose Extraction

As mentioned in Section 3.2, we use Alphapose [3, 4, 5] to track people in each subtitle-sign video aligned clip and obtain the whole body keypoints. For each frame, we save 136 keypoints for each person – 26 pose landmarks from the body, 68 pose landmarks from the face and 21 additional landmarks for each hand. Alphapose is an accurate multi-person pose estimator, which is the first open-source system. To match poses that correspond to the same person across frames, we also provide an efficient online pose tracker called Pose Flow. It is the first open-source online pose tracker. Figure 2 shows the result of Alphapose in a frame. We employ the ID annotation in the Alphapose output to identify the signer throughout each sign video clip, meanwhile rectifying any anomalies within the results.

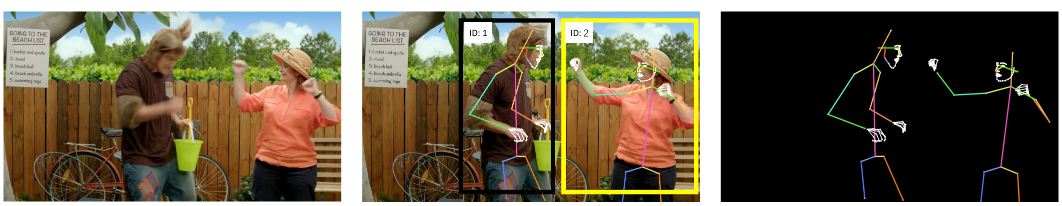

Figure 2: Original image (Left), Keypoints along with image (Middle) and Keypoints (Right)

A.5 Statistics of Data Labelling Procedure

Table 2: Statistics of the data labelling procedure. R/M: Remove/Modify.

	Time (h)	#Annotators	R/M	#Video Clips
Data Download	20	-	- / -	29.7k
Subtitle Cleaning	5	-	2.2k / -	27.5k
Sign Language Alignment	300	5 (experts)	1.4k / 3.2k	26.1k
Pose Extraction	150	-	- / -	26.1k
Signer Verification	60	5	- / 4.8k	26.1k
Total	535	10	3.6k / 8.0k	26.1k

Table 2 shows the time cost, the number of invited annotators, along with the amount of modified data, deleted data, and remaining data in each task of the data labelling procedure. We invite Auslan experts for the sign language alignment task. While this constitutes a time-consuming procedure, it ensures sign data accuracy. The other tasks are Auslan knowledge-free, i.e., they can be fulfilled either by automatic methods or annotators without Auslan knowledge.

A.6 Statistics of Test Set Sentences

We provide the statistics of the sentences in the test set in Table 3. Even though there might be similar English subtitles, every sign video clip in the test sets is still different from those in the training sets. This is because those sign sentences are signed by different signers, captured under different backgrounds or under different camera perspectives. Hence, we can guarantee that test samples are not included in the training set. Moreover, over 80% of video clips in the test set present unique sentences. In other words, these sentences never appear in the training set. Therefore, the robustness of models can be verified by evaluating on the test set.

Table 3: Statistics for Auslan-Daily Communication and Auslan-Daily News

	Auslan-Daily Communication	**Auslan-Daily News**
Num. Sentence in test set	800	700
Num. Unseen Words	10	304
Num. Unseen Sentences	564	662

A.7 Final Dataset Storage

The final datasets, labelled by experts and annotators, are stored in a folder on Google Drive. The structure shows in Figure 3. There are three main sub-folder, *Auslan-Daily Communication*, *Auslan-Daily News* and *Dataset Split Table*, respectively. In *Auslan-Daily Communication and News* sub-folder, including (1) *Signer Only Video Clip*, the video clip crop the signer regions based on the max ground-truth bounding-box. (2) *Multi-Person Video Clip*, the video clip without cropping acting signer. (3) *Whole Video*, original video after downloading. (4) *Pose Annotation*, Keypoints within each video clip are extracted utilizing the Alphapose. In *Dataset Split Table* sub-folder, we furnish partitioned training, validation, and test sets for signer detection, fingerspelling detection, isolated sign language recognition, sign spotting, sign language alignment, and translation tasks. In Table 4, we provide the associated with the split of the sign language translation dataset. It shows recommended data splits used in our experiments for other Auslan sign language-related tasks.

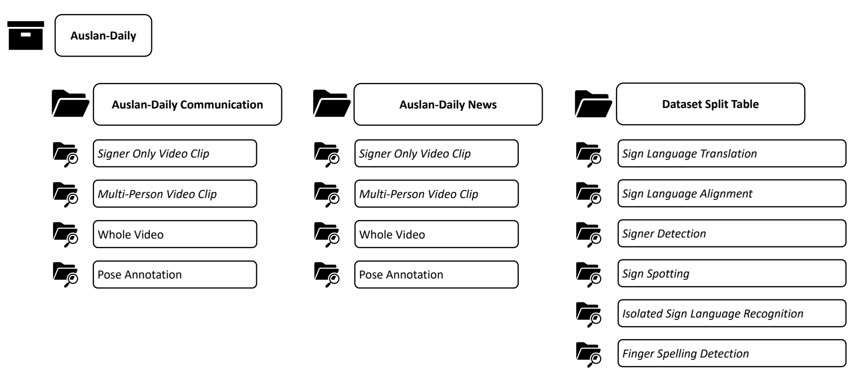

Figure 3: Hierarchical data folders for Auslan-Daily on Google Drive

B More Details for Video Representation

RGB-based: We use the pre-trained I3D model form [6] and features with a window width of 16 and a stride of 2 are extracted:

$$f_t = \text{I3D}(F_{t-\frac{n}{2}} \oplus ... \oplus F_t \oplus ... \oplus F_{t+\frac{n}{2}}), \tag{1}$$

Table 4: Data Split for Auslan-Daily Sub-Tasks

Sub-Task	Train	Dev	Test	Total
Signer Detection	37k people (20k signers)	4.6k people (2.4k signers)	4.8k people (2.7k signers)	46.4k people (25.1k signers)
Sign Language Alignment	120 episodes	17 episodes	20 episodes	157 episodes
Isolated Sign Lanuage Recoginition	1.8k (600 classes)	0.6k (600 classes)	0.6k (600 classes)	3k (600 classes)
Fingersplling Detection	1.7k	0.1k	0.2k	2k
Sign Spotting	0.8k (100 classes)	0.1k (100 classes)	0.1k (100 classes)	1k (100 classes)

where f_t is the representation of the t-th frame, n is the window width, and $\oplus$ denotes the concatenation operation.

Pose-based: Leveraging pose information in action recognition presents significant benefits regarding robustness and semantic representation. We flatten the pose array $A \in R^{T \times N \times 2}$ to $A_f \in R^{T \times 2N}$, where T is the number of frames and N is the number of keypoints. Meanwhile, our experiment results show that using partial body and two hands keypoints will perform better for the sign language translation task. The selected keypoints are shown in Figure 4.

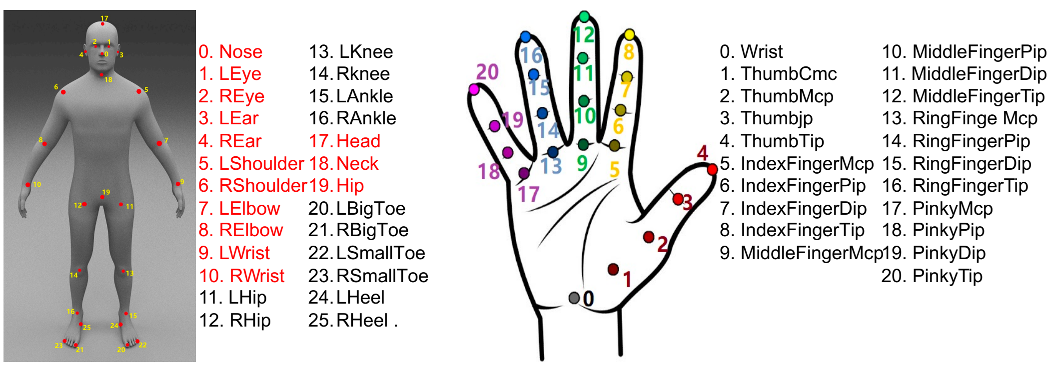

Figure 4: Pose landmarks extracted using Alphapose

C Experimental Settings

C.1 Model Implementation

We mention that all models used in this work are publicly available. Each of the models we use is linked below:

- **Sign Language Translation:** SL-Luong [7], SL-Transf [8], TSPNet [9] and MMTLB [10]
- **Signer Detection and Isolated Sign Language Recognition:** I3D [11] and Pose-TGCN [6]
- **Sign Language Alignment:** SAT [12]
- **Fingerspelling Detection:** Multi-task Model [13]
- **Sign Spotting:** MMSP [14]

We express profound gratitude to the aforementioned authors for their invaluable contributions.

All the training and fine-tuning experiments are run on a machine with two NVIDIA GeForce RTX 3090 GPUs. For single-person and end-to-end multi-person SLT models, the batch size, learning rate, and epoch are set to 64, 5e-5, and 200, respectively. We use the default hyperparameters for training the models of other sign language-related tasks.

D The Baseline of Sub-tasks on Auslan-Daily

D.1 More Methods for Mentioned Sign Language-Related Tasks

We further incorporate more methods to evaluate the performance of other sign language-related tasks on our Auslan-Daily. Specifically, we adopt TSN [15], SlowFast [16] and Timesformer [17] to perform isolated sign language recognition (Table 5) and signer detection (Table 6), Bi-LSTM CTC [18] and Modified R-C3D [19] to perform fingerspelling detection (Table 7), HS-I3D [20] and Two-Stage-SP [21] to perform sign spotting (Table 8), and S_{audio} [12], S_{audio+} [12] (whether we shift audio for alignment) and Segment SLV [22] to perform sign language alignment (Table 9).

Table 5: The baseline of Isolated Sign Language Recognition on Auslan-Daily.

Model	Top-1	Top-5
I3D [11]	11.87	20.10
TSN [15]	24.75	**41.31**
SlowFast [16]	23.97	38.25
Timesformer [17]	**27.38**	40.33

Table 6: The baseline of Signer Detection on Auslan-Daily.

Model	Top-1
I3D [11]	89.01
TSN [15]	89.98
SlowFast [16]	90.37
Timesformer [17]	**93.65**

Table 7: The baseline of Fingerspelling Detection on Auslan-Daily.

Model	AP@0.1	AP@0.5
Bi-LSTM CTC [18]	0.25	0.09
Modified R-C3D [19]	0.30	0.18
Multi-FD [40]	**0.33**	**0.21**

Table 8: The baseline of Sign Spotting on Auslan-Daily.

Model	F1 score
HS-I3D [20]	**0.35**
Two-Stage-SP [21]	0.23
Dual-Branch-SP [14]	0.27

Table 9: The baseline of Sign Language Alignment on Auslan-Daily.

Model	F1@.10	F1@.50
S_{audio} [12]	50.68	22.30
S_{audio+} [12]	74.87	29.65
Segment SLV [22]	77.43	41.76
SAT [21]	**82.49**	**60.76**

Table 10: Sign Language Production (SLP) performance of Text2Pose [23] on Auslan-Daily.

Communication		News	
ROUGE	BLEU-4	ROUGE	BLEU-4
16.30	0.61	26.29	0.54

D.2 Sign Language Production

As the reverse task of Sign Language Translation (SLT) [24, 23], Sign Language Production (SLP) undoubtedly holds exceptionally high research value. To evaluate Sign Language Production (SLP), accurate 3D keypoints of signers are often required. However, when we apply state-of-the-art 3D pose estimation methods to our dataset, they all fail to provide precise 3D poses, especially for hand gestures. This is because the resolution of hand areas might not be as high as the datasets specific for 3D hand pose estimation. Moreover, in our Auslan-Daily, signers may not face cameras in a frontal view and there are self-occlusions and occlusions by other objects. These factors impose difficulties in accurate 3D pose estimation. Employing generative adversarial networks for SLP is another option. However, we found the cluttered background in Auslan-Daily significantly impedes network learning and the hand gestures are barely recognisable in the generated videos. To provide a baseline for SLP, we replace the 3D keypoints with 2D keypoints in Text2Pose (T2P) [23]. As indicated in Table 10, T2P achieves 0.61 and 0.54 in BLEU-4 on the communication and news subsets, respectively. Without precise 3D keypoints, the diverse orientation of signers inevitably introduces ambiguity to SLP. Additionally, compared to PHOENIX-2014T [4], the larger vocabulary size of Auslan-Daily further imposes challenges on SLP. Therefore, we reckon the current annotations of Auslan-Daily (missing accurate 3D poses) may not be sufficient for high-quality SLP. In the future, we will consider annotations of 3D poses. Then, SLP can be accurately evaluated on Auslan-Daily.

Table 11: Gloss-free Single-Pre. SLT with different sign language video representations. PD, WS, R and B4 refer to the pre-training dataset, window size, ROUGE score and BLEU-4 score, respectively.

		Auslan-Daily Comm.			Auslan-Daily News		
Model	**PD**	**WS**	**R**	**B4**	**WS**	**R**	**B4**
SL-Luong [7]	WLASL [6]	8/12/**16**	13.49	4.66	8/12/**16**	16.14	**2.68**
SL-Transf [8]	WLASL [6]	8/12/**16**	14.97	5.20	8/12/**16**	14.93	2.52
SL-Luong [7]	MSASL [27]	8/**12**/16	**19.10**	**6.94**	8/12/**16**	**16.68**	2.31
SL-Transf [8]	MSASL [27]	8/**12**/16	16.58	4.90	8/**12**/16	15.43	2.45
SL-Luong [7]	BSL [28]	8/12/**16**	12.55	3.54	8/**12**/16	12.95	1.48
SL-Transf [8]	BSL [28]	8/12/**16**	15.98	4.01	**8**/12/16	11.07	1.42
SL-Luong [7]	BSL+WLASL [6]	8/**12**/16	14.71	5.23	8/**12**/16	14.82	1.97
SL-Transf [8]	BSL+WLASL [6]	**8**/12/16	14.18	4.41	**8**/12/16	15.70	2.08
SL-Luong [7]	BSL+MSASL [27]	8/**12**/16	15.21	6.43	8/12/**16**	15.76	2.33
SL-Transf [8]	BSL+MSASL [27]	8/**12**/16	14.56	4.57	8/**12**/16	14.29	1.95

E Ablation Study on Single-Person SLT

E.1 Different visual stream representations.

To dissect the impacts of different visual representations, we evaluate sign language translation methods with different network architectures (RNN-based and Transformer-based models), window sizes and pre-trained backbones. As shown in Table 11, using MSASL [36] and the window size of 12, the SLT model performs the best on Auslan-Daily Communication while using WLASL [35] and a window size of 16, the model SLT performs the best on Auslan-Daily News. These experiments indicate the differences between the communication and News corpora and the challenges inherent in sign language translation in the wild.

E.2 Different pose stream representations.

AlphaPose [3] is the latest and most accurate multi-person pose estimator. It not only provides keypoint extraction but also integrates the pose-tracking function. However, most previous works seem to use HRNet [25] to extract keypoints [26]. We evaluate the SLT performance based on the poses extracted by HRNet, and the results are shown in Table 12. We observe that the SLT performance based on HRNet is lower than that using AlphaPose because AlphaPose achieves higher and more reliable keypoint estimation results, especially hands [3].

Table 12: Gloss-free Single-Pre. SLT with different pose representations.

		Auslan-Daily Communication					Auslan-Daily News				
	Input	R	B1	B2	B3	B4	R	B1	B2	B3	B4
SL-Luong [7]	HRNet [25]	34.86	30.24	14.83	10.14	8.04	**21.57**	20.78	6.73	3.30	2.15
SL-Transf [8]	HRNet [25]	36.23	27.75	14.52	10.57	8.82	19.43	19.27	6.13	3.23	2.00
SL-Luong [7]	**Alphapose [3]**	**37.27**	30.15	**16.26**	**11.67**	**9.45**	20.65	19.84	**7.81**	**4.59**	**2.81**
SL-Transf [8]	Alphapose [3]	35.65	**31.31**	16.17	11.41	9.20	20.25	**21.25**	6.57	3.32	2.11

F Case Study for Auslan-Daily Sign Language Translation

We randomly select 15 examples from the test set of Auslan-Daily Communication and News. We compare the well-performance of Single-Person SLT and Multi-Person SLT models prediction against the ground truth transcription (see Table 13). The precisely correct translations are primarily short and commonly used sentences in Auslan-Daily. The model frequently fails to capture their general meaning for longer and more complex sentences though some keywords can be predicted correctly. Meanwhile, it can be observed that the efficacy of the SD+SLT model aligns closely with the performance demonstrated by the single-person SLT model.

Table 13: Case study. We highlight exactly correct translations in red and semantically correct translations in blue.

Auslan-Daily Communication Translation Results	
GT	**it is delicious .**
SL-Luong + Pose	this is delicious
SD + SL-Luong + Pose	this is delicious
GT	**well our time is up .**
SL-Luong + Pose	well our time is up .
SD + SL-Luong + Pose	well our time is up .
GT	**i know possum would love to help with this .**
SL-Luong + Pose	i know possum would love to help .
SD + SL-Luong + Pose	i think possum would love to help .
GT	**why do not we watch some child learn about the weather .**
SL-Luong + Pose	why do not we watch some child learn about them .
SD + SL-Luong + Pose	well why do not we watch some child learn about fraction .
GT	**yes possum that is right .**
SL-Luong + Pose	yes you is right possum .
SD + SL-Luong + Pose	yes .
GT	**i know possum would love to help .**
SL-Luong + Pose	i know possum would love to help .
SD + SL-Luong + Pose	well i am go to go outside and ask him.
GT	**the instruction on how to make it be in the drawer .**
SL-Luong + Pose	there are some instruction on how to make it .
SD + SL-Luong + Pose	can you please get the recipe out of your tree .
GT	**sally do you know what else tree be good for .**
SL-Luong + Pose	sally can we make something like that.
SD + SL-Luong + Pose	i am go to the top of my tree .
GT	**would you like to make some .**
SL-Luong + Pose	would you like to make something possum .
SD + SL-Luong + Pose	would you like to make it for me .
GT	**sally what do it say .**
SL-Luong + Pose	sally what do you want to do .
SD + SL-Luong + Pose	what are you doing .
Auslan-Daily News Translation Results	
GT	**hello and welcome to abc news .**
SL-Luong + Pose	hello and welcome to abc news .
SD + SL-Luong + Pose	hello and welcome to abc news .
GT	**that is the late from abc news .**
SL-Luong + Pose	that is the late from abc news .
SD + SL-Luong + Pose	that is the late from abc news .
GT	**the prime minister has again insisted his government is not ...**
SL-Luong + Pose	the prime minister anthony albanese have claimed the government ...
SD + SL-Luong + Pose	the pandemic hit a high help with the air at the moment and that he ...
GT	**but that is not how leeanne caton remembers things play out .**
SL-Luong + Pose	but he says that is not important enough .
SD + SL-Luong + Pose	but he says it is not important.
GT	**the new south wale teacher federation have rubbish ...**
SL-Luong + Pose	the new south wale prime minister anthony albanese have ...
SD + SL-Luong + Pose	a new south wale and the federal government is plan ...

References

[1] Gokul NC, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh Khapra. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. In *NeurIPS*, 2022.

[2] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7784–7793. Computer Vision Foundation / IEEE Computer Society, 2018.

[3] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[5] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[6] 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 *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[7] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics, 2015.

[8] 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 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE, 2020.

[9] Dongxu Li, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, and Hongdong Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[10] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5110–5120. IEEE, 2022.

[11] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017.

[12] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning subtitles in sign language videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11532–11541. IEEE, 2021.

[13] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Fingerspelling detection in american sign language. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4166–4175. Computer Vision Foundation / IEEE, 2021.

[14] Hongyu Fu, Chen Liu, Xingqun Qi, Beibei Lin, Lincheng Li, Li Zhang, and Xin Yu. Sign spotting via multi-modal fusion and testing time transferring. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13808 of *Lecture Notes in Computer Science*, pages 271–287. Springer, 2022.

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 20–36. Springer, 2016.

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019.

[17] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021.

[18] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

[19] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: region convolutional 3d network for temporal activity detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5794–5803. IEEE Computer Society, 2017.

[20] Ryan Wong, Necati Cihan Camgöz, and Richard Bowden. Hierarchical I3D for sign spotting. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13808 of *Lecture Notes in Computer Science*, pages 243–255. Springer, 2022.

[21] Manuel Vázquez-Enríquez, José Luis Alba-Castro, Laura Docío Fernández, Júlio C. S. Jacques Júnior, and Sergio Escalera. ECCV 2022 sign spotting challenge: Dataset, design and results. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13808 of *Lecture Notes in Computer Science*, pages 225–242. Springer, 2022.

[22] Hannah Bull, Michèle Gouiffès, and Annelies Braffort. Automatic segmentation of sign language into subtitle-units. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12536 of *Lecture Notes in Computer Science*, pages 186–198. Springer, 2020.

[23] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 687–705. Springer, 2020.

[24] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5131–5141. IEEE, 2022.

[25] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, 2021.

[26] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. In *NeurIPS*, 2022.

[27] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding american sign language. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 100. BMVA Press, 2019.

[28] 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.