Sign Language Translation from Instructional Videos

Laia Tarrés[1,2] Gerard I. Gállego[1] Amanda Duarte[2] Jordi Torres[1,2] Xavier Giró-i-Nieto[3,*]
[1]*Universitat Politècnica de Catalunya* [2]*Barcelona Supercomputing Center* [3]*Amazon*
https://imatge-upc.github.io/slt_how2sign_wicv2023

Abstract

The advances in automatic sign language translation (SLT) to spoken languages have been mostly benchmarked with datasets of limited size and restricted domains. Our work advances the state of the art by providing the first baseline results on How2Sign, a large and broad dataset.

We train a Transformer over I3D video features, using the reduced BLEU as a reference metric for validation, instead of the widely used BLEU score. We report a result of 8.03 on the BLEU score, and publish the open-source implementation to promote further advances.

1. Introduction

Sign language translation (SLT) is the task of translating continuous sign language videos into spoken language sentences. SLT is a challenging multimodal problem that requires both a precise understanding of the signer's pose and the generation of a textual transcription. The current state of the art for automatic SLT is still far away from considering the problem solved [8, 12, 14, 45, 47, 48].

Recent advances in SLT have followed a trajectory similar to other computer vision and natural language processing problems: training deep neural networks on large-scale datasets. However, the availability of public sign language datasets is limited and especially reduced when considering parallel corpus of videos and their textual translations. Up to date, the most used dataset to assess the progress in SLT is PHOENIX-2014-T [20], with only 9.2 hours of video recordings on the restricted weather forecasts domain.

In this work, we consider a much larger and complex dataset, How2Sign [19], which contains almost 80 hours of instructional videos from 10 different topics. In addition, we explore and suggest using an alternative metric [17], *reduced BLEU* (rBLEU) to better characterize the performance and choose better checkpoints during training.

We provide open code and models which allows reproducibility and adaptation to other datasets.[1]

*Work done outside of Amazon
[1]https://github.com/imatge-upc/slt_how2sign_wicv2023

2. Related Work

Sign language video understanding has been addressed from a variety of tasks: sign language recognition (SLR) over isolated or continuous signs [1, 15, 21, 22, 32, 34, 36], sign language translation (SLT) [7,13,20,25], sign language production (SLP) [38–42] or retrieval [18]. Our work focuses on sign language translation.

Table 1 shows the current state of the art in terms of the BLEU metric for different SLT benchmarks. Reasonable scores in the range between 29 and 60 BLEU have been reported in three datasets of limited vocabulary size: KETI [26], PHOENIX-2014T [6], and CSL Daily [48].

Our work aims at the more open domain of instructional videos across 10 different topics, to set the first SLT baselines on the How2Sign [19] dataset. This dataset has been used for other sign language-related tasks [5, 18], but never for SLT.

While the scores are not directly comparable, our baselines are similar to OpenASL [43]. Other works on alternative datasets of large scale obtained very poor BLEU scores: 1.0 in BOBSL [3], 0.4 in SWISSTXT-NEWS [9], 0.4 in VRT-NEWS [9], or 0.37 in SRF [44] and 0.84 in Focus-News [17] in the WMT shared task on sign language translation 2022 [33].

3. Data Preprocessing

One of the main challenges in SLT is the variability and complexity of sign languages, which can be influenced by a variety of factors such as the signer's background, context, and appearance. Therefore, it is important to preprocess the data to reduce this variability. This includes techniques such as visual feature extraction and normalization, as well as standardizing the format of the target data.

3.1. Video tokenization

We choose I3D features [10] to extract video representations directly from the RGB frames, motivated by their effectiveness in the sign recognition [23,28] and retrieval [18] tasks. I3D features consider not only visual cues, but also temporal information. As a result, they provide a dense and

Dataset	Duration(h)			Vocabulary(K)			BLEU	Domain
	train	val	test	train	val	test		
KETI [26]	20.05	2.24	5.70	←	0.49	→	57.37 [26]	Emergency situations
PHOENIX-2014T [6]	9.2	0.6	0.7	2	0.9	1	25.59 [46]	Weather Forecast
CSL Daily [48]	20.62	1.24	1.41	2	1.3	1.3	23.92 [11]	Daily life
OpenASL [43]	←	288	→	←	33	→	6.72 [43]	Youtube (news + vlogs)
How2Sign [19]	69.6	3.9	5.6	15.6	3.2	3.6	8.03 (Ours)	Instructional

Table 1. Comparison between SLT datasets based on the duration of the videos (in hours), number of unique words (in thousands) in the vocabulary and SOTA on SLT without glosses. ← → indicate that in some cases only statistics on the whole dataset are provided.

reliable source of visual cues as input to our models.

The original I3D network is trained on ImageNet [16] and fine-tuned for action recognition with the Kinetics-400 [24] dataset. As shown in [2, 17, 18, 29, 35, 43], further fine-tuning with sign language data is needed to properly model the temporal and spatial information present in them. We used the I3D features provided in [18].

3.2. Text processing

Text preprocessing is an important step in preparing raw text data into a more suitable format for NLP models.

Similar to NLP pipelines, our system first converts raw text to lowercase. We employ the Sentencepiece tokenizer [27] to segment the lowercase text into sub-word units. Sub-word tokenization requires specifying a fixed vocabulary size, which has trade-offs in terms of better representation and computational efficiency. To ensure a fair assessment of the system's performance, it is necessary to compare the model outputs to the original test set without any prior processing. However, this approach may result in a lower BLEU score, as the model generates text based on preprocessed data. Therefore, we implement a postprocessing step, that involves detokenization and truecasing [30], to restore the original capitalization.

4. Methodology

The building blocks of our implementation are depicted in Figure 1. The input video stream is tokenized with a pretrained I3D feature extractor. These tokens are fed into the encoding Transformer layers. The Transformer decoder operates with lowercase and tokenized textual representations.

4.1. Neural architecture

We use a standard transformer encoder-decoder. We choose an asymmetric encoder-decoder with six encoder layers and three decoder layers, each with four attention heads, we select an embedding dimension of 256 and feed-forward network hidden size of 1024, which corresponds to ID (17) from Table 3.

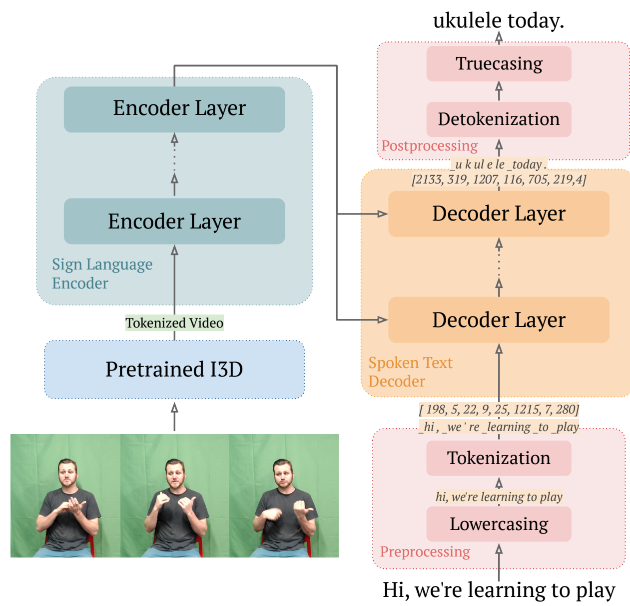

Figure 1. The input video sequence is fed into a Transformer to generate the output text sequence.

4.2. Implementation details

In our implementation, we first preprocess the vocabulary as described in Section 3.2, with a vocabulary size of 7000 subwords.

For training, the batch size is set to 32, and we use cross entropy loss with label smoothing of 0.1. We select the Adam optimizer, we warm-up the learning rate for the first 2000 updates, and then we apply a cosine decay from 10^{-3} to 10^{-7} with warm restart every $1.7 \cdot 10^4$ steps. We train the model for 10^5 steps, equivalent to 108 epochs. We perform validation every two epochs. Our training process takes 3.5 hours on a single NVIDIA GeForce RTX 2080 Ti GPU.

For inference, we adopt steps commonly used in machine translation and use beam search algorithm to generate predictions, we choose a beam size of five.

	val					test				
	rBLEU	BLEU-1	BLEU-2	BLEU-3	**BLEU**	**rBLEU**	BLEU-1	BLEU-2	BLEU-3	**BLEU**
Ours.	2.79	35.2	20.62	13.25	8.89	2.21	34.01	19.3	12.18	**8.03**

Table 2. Best scores on How2Sign for Sign Language Translation.

4.3. Evaluation protocol

To measure the performance of our SLT models, we use BLEU score [37] [2].

The difficulty of the SLT task causes a bias in the model prediction towards most statistically frequent patterns, such as Example (2) in Table 4. These patterns can inflate the BLEU scores without actually translating anything meaningful. Inspired by [17] we compute reducedBLEU (rBLEU). This metric consists of removing certain words from the reference and the prediction before computing the BLEU score. We create a blacklist of words that are frequently used in the training data but do not contribute much to the meaning of the sentences. Table 4 shows a comparison between rBLEU and BLEU metrics.

Focusing on concrete examples, row (2) in Table 4, shows that both the prediction and the reference contain the phrase “In this clip I’m going to show you how to”, which is one of the frequent patterns on the instructional dataset. This pattern inflates the BLEU score, while it does not affect the rBLEU score, which is low, suggesting that sentences have different meanings.

Our experimental results indicate that rBLEU is a more reflective indicator of actual performance than traditional BLEU, for low-resource settings that also have repetitive patterns, given that it considers mostly semantically meaningful words. In order to provide comparable results with other works, we also report standard BLEU in our results.

5. Experiments

The performance of our proposed approach is shown in Table 2. We evaluate our models using the metrics described in Section 4.3 and provide examples of generated spoken language translation sentences.

5.1. Quantitative results

Our implementation achieves results reported in Table 2. To the authors’ knowledge, these are the first published results for SLT obtained with the How2Sign dataset. The table displays the results of our best configuration, which provides a baseline from where future work can build upon.

[2]Other SLT papers use BLEU-4 instead of BLEU. It represents the same score, we use BLEU for simplicity.

5.2. Qualitative results

We provide a qualitative assessment of the results in Table 4, showing a few spoken language translations generated by our best-performing model. Words used to compute rBLEU are in bold.

Example (1) demonstrates the ability of our model to provide detailed translations even for complex words like *“women’s self defense”*. Our metrics indicate both high BLEU and rBLEU scores meaning that the model is generating a good translation, considering both full sentences and meaningful words.

However, our results also suggest that this is not always the case. For instance, in Example (2), BLEU is higher than rBLEU. As mentioned in 4.3, BLEU score is high due to the repetitive patterns frequent in the instructional dataset. Given that high BLEU scores can be misleading due to their susceptibility to frequent phrases, we emphasize the importance of using rBLEU instead of BLEU when selecting the best checkpoint.

The provided examples suggest that the models’ performance may depend on the complexity and length of the signed video. We observed that the model was able to provide reasonably accurate translations for short sentences, but not for sentences like Example (4).

Example (5) illustrates the reason behind the disparity between rBLEU and BLEU metrics, explained in Section 4.3. In this case, despite obtaining a high BLEU score and an accurate translation, the corresponding rBLEU score is zero due to the reduced number of remaining words for rBLEU calculation, which is less than four.

Overall, the findings suggest that the model’s quality is still suboptimal, as demonstrated by Example (3), which has comparable metrics to the overall performance.

5.3. Hyperparameter search

Transformer under low-resource conditions is highly dependent on hyperparameter settings [4]. Our experiments show that using an optimized Transformer improves the translation quality over 3.47 BLEU points and 1.8 reduced BLEU points compared to the default hyperparameters for SLT.

Table 3 shows the hyperparameters that we optimize. Default hyperparameters for SLT come from [8].

A current observation in Transformers is that increasing the number of parameters will improve the performance.

	Values
Text preprocessing	{**yes**, no}
Vocabulary size	{1k, 4k, **7k**}
Batch size	{**32**, 64}
Learning Rate (LR)	{5e-2, **1e-3**, 5e-3}
LR scheduler	{**cosine**, inv_sqrt}
Warm-up steps	{0, **2k**, 4k}
Warm restarts period	{0,**17k**, 22k}
Weight Decay	{1e-3, 1e-2, **1e-1**}
Label Smoothing	{0, **0.1**}
Dropout	{0, 0.1, 0.2, **0.3**}
# Layers (encoder-decoder)	{2-2, 3-3, 4-2, **6-3**}
Embed dim	{**256**, 512}
FFN dim	{512, **1024**, 2048}
# Attention heads	{**4**, 8}

Table 3. Hyperparameters search space. In bold are the optimal ones that we found during validation, and underlined are defaults.

ID		rBLEU	BLEU
(1)	*Ref:* And that's a **great vital point technique** for **women's self defense**. *Pred:* It's really a **great point** for **women's self defense**.	30.29	38.25
(2)	*Ref:* In this **clip** I'm going to **show** you how to **tape** your **cables** down. *Pred:* In this **clip** I'm going to **show** you how to **improve push ups**.	24.88	64.53
(3)	*Ref:* You are **dancing**, and now you are going to **need** the **veil** and you are going to just **grab** the **veil** as **far** as **possible**. *Pred:* So, **once** you're **belly dancing**, **once** you've **got** to have the **strap**, you're going to **need** to **grab** the **thumb**, and **try** to **avoid** it.	4.93	8.04
(4)	*Ref:* But if you have to **setup** a **new campfire**, **there's two ways** to do it in a very **low impact**; one is with a **mound fire**, which we should in the **campfire segment earlier** and the other way to **setup** a **low impact campfire** is to have a **fire pan**, which is just a **steel pan** like the **top** of a **trash** can. *Pred:* And other thing I'm going to **talk** to you is a little bit more **space**, a **space** that's what it's going to do, it's **kind** of a **quick**, and then I don't want to take a **spray skirt** off, and then I don't want it to take it to the **top** of it.	0.85	3.79
(5)	*Ref:* So, this is a very **important part** of the **process**. *Pred:* It's a very **important part** of the **process**.	0.0	61.86

Table 4. Qualitative examples from our best-performing model. In bold the words remaining to compute rBLEU. Corresponding input frames from examples can be found in Appendix A.2

However, in low-resource languages, increasing the number of model parameters can hinder performance [47]. We study this effect by changing the number of layers in the encoder and decoder, the number of attention heads, the feed-forward layer dimension, and embedding dimensions.

Since the optimization of the learning rate (LR) is dependent on the number of parameters of the model, we tune it together with other hyperparameters related to the architecture size. Furthermore, we introduce the use of LR scheduling of cosine with warm restarts [31], which has been shown to perform better than alternatives.

Our experiments point to the direction that smaller models obtain better results, for example using an encoder-decoder configuration of 2-2, with an embedding dimension of 256, feed-forward dimension of 512 and 4 heads, the model achieves 1.37 rBLEU. Due to the fact that the input data is by far more complex than the output, we choose to carry out further experiments with both the best symmetric model and the asymmetric, but a priori we do not observe much improvement.

Given the observed overfitting on bigger models, we add regularization by adding dropout, weight decay, and label smoothing. We observe that adding regularization to big models outperforms the rest of configurations. For our final model, with the parameters highlighted in 3 we see substantial improvement by using dropout of 0.3, weight decay of 0.1 and label smoothing of 0.1, obtaining a val rBLEU of 2.79, which improves the best model without regularization by 1.85 rBLEU.

6. Conclusions

In this work, we made an open-source implementation that can serve as a first baseline for Sign Language Translation on the How2Sign dataset. We achieved BLEU score of 8.03, which indicates a certain degree of understanding of the signed utterances. This value is on-par with the best results reported for OpenASL [43], the most similar publicly available dataset of comparable complexity.

Furthermore, we have done an extensive hyperparameter search and shown that tuning is necessary to obtain the best set of results. The best results are obtained with a bigger than baseline Transformer trained with great amounts of regularization.

Our quantitative and qualitative evaluations have led us to conclude that rBLEU is a suitable evaluation metric for similar benchmarks, particularly in cases where datasets are low-resource with frequent repetitive patterns. In contrast to the traditional BLEU score, which may be inflated due to these patterns, rBLEU provides a more accurate evaluation that better reflects the model's performance.

Acknowledgements

This research was partially supported by research grant Adavoice PID2019-107579RB-I00 / AEI / 10.13039/501100011033, research grants PRE2020-094223, PID2021-126248OB-I00 and PID2019-107255GB-C21 and by Generalitat de Catalunya (AGAUR) under grant agreement 2021-SGR-00478.

References

[1] 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

[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 (ECCV)*, 2020. 2

[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. 1

[4] Ali Araabi and Christof Monz. Optimizing transformer for low-resource neural machine translation. *Proceedings of the 28th International Conference on Computational Linguistics*, 2020. 3

[5] Alvaro Budria, Laia Tarres, Gerard I Gallego, Francesc Moreno-Noguer, Jordi Torres, and Xavier Giro-i Nieto. Topic detection in continuous sign language videos. *arXiv preprint arXiv:2209.02402*, 2022. 1

[6] 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

[7] 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*. Springer, 2020. 1

[8] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[9] Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021. 1

[10] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[11] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5130, 2022. 2

[12] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. In *Advances In Neural Information Processing Systems (NeurIPS)*, 2022. 1

[13] 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. 1

[14] Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. Frozen pretrained transformers for neural sign language translation. In *18th Biennial Machine Translation Summit (MT Summit 2021)*, pages 88–97. Association for Machine Translation in the Americas, 2021. 1

[15] 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

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[17] Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, and Oscar Koller. Clean text and full-body transformer: Microsoft's submission to the wmt22 shared task on sign language translation. *Proceedings of the Seventh Conference on Machine Translation*, 2022. 1, 2, 3

[18] Amanda Duarte, Samuel Albanie, Xavier Giro i Nieto, and Gul Varol. Sign language video retrieval with free-form textual queries. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[20] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, 2014. 1

[21] 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)*, 2021. 1

[22] 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

[23] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. 2019. 1

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[25] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 2019. 1

[26] 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

[27] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Conference on Empirical Methods in Natural Language Processing*, 2018. 2

[28] 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

[29] 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. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020. 2

[30] Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. tRuEcasIng. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 152–159, Sapporo, Japan, July 2003. Association for Computational Linguistics. 2

[31] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 4

[32] 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

[33] Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, et al. Findings of the first wmt shared task on sign language translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, 2022. 1

[34] 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

[35] Alptekin Orbay and Lale Akarun. Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 222–228. IEEE, 2020. 2

[36] 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

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 3

[38] 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

[39] 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

[40] 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

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

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

[43] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 1, 2, 4

[44] Laia Tarrés, Gerard I Gállego, Xavier Giró-i Nieto, and Jordi Torres. Tackling low-resourced sign language translation: Upc at wmt-slt 22. *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022. 1

[45] Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955, 2021. 1

[46] Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11946–11955, 2021. 2

[47] Kayo Yin and Jesse Read. Better sign language translation with stmc-transformer. *Proceedings of the 28th International Conference on Computational Linguistics*, 2020. 1, 4

[48] Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, 2021. 1, 2