Isolated Sign Language Recognition for Icelandic Sign Language: Experiments in a Low-resource Setting

Finnur Ágúst Ingimundarson[1], **Guðný Björk Þorvaldsdóttir**[2]
Mathias Müller[1], **Sarah Ebling**[1]
[1]Department of Computational Linguistics, University of Zurich
[2]Communication Centre for the Deaf and Hard of Hearing in Iceland
finnuragust.ingimundarson@uzh.ch
gudny.bjork.thorvaldsdottir@shh.is
{mmueller,ebling}@cl.uzh.ch

Abstract

We present the first experiments on isolated sign language recognition (ISLR) for Icelandic Sign Language (ÍTM). We use ÍTM SignWiki, a dataset derived from a bilingual Icelandic–ÍTM online dictionary. It is genuinely low-resource: 1,845 videos cover 849 classes, 86% of which have only two examples, making the full task effectively one-shot recognition across signers. We compare two open-source ISLR frameworks, OpenHands and SPOTER, on three tasks of increasing vocabulary size (22, 117 and 849 classes), and evaluate three pose estimators and two forms of cross-lingual transfer. With ÍTM data alone, SPOTER outperforms OpenHands on all three tasks, and MediaPipe poses give better results than AlphaPose or SDPose. Cross-lingual transfer brings the largest gains: pretraining SPOTER on American Sign Language data before fine-tuning on ÍTM raises accuracy by 14–24 percentage points, to 72.7%, 47.9% and 22.6% on the three tasks, and multilingual training with data from six other sign languages lifts OpenHands from 1.41% to 28.86% on the full task. Although far from practical use, the results suggest that transfer from better-resourced sign languages is promising for very low-resource ones. We release our adapted versions of both frameworks.

1 Introduction

Although not always recognized as such, sign languages are full-fledged natural languages that have their own grammar and vocabulary.[1] Contrary to a widespread belief, there is no universal sign language and several hundred sign languages have been documented around the world. The number of deaf people who use them as a primary means of communication worldwide is estimated to be roughly 70 million, and more than half a million in Europe (Way et al., 2024).

[1]This paper is based on Ingimundarson (2026) the first author's master's thesis at the University of Zurich

The field of sign language processing (SLP) is more recent than that of natural language processing (NLP) and has lagged behind, constrained by a lack of technological or computational resources, software, and data (Yin et al., 2021; Müller et al., 2022). This situation has gradually improved but SLP remains underrepresented and fundamental design decisions underexplored (O'Brien et al., 2026). In general, it can also be argued that all sign languages are low-resource languages when compared to spoken languages (Joshi et al., 2020). Nevertheless, larger sign language communities, such as that of American Sign Language (ASL), are comparatively well represented and well resourced, whereas sign languages of smaller communities are low-resource, which restricts the development and deployment of SLP tools based on them.

One such low-resource sign language is Icelandic Sign Language, that will hereafter be abbreviated as ÍTM (*íslenskt táknmál*), which is preferred by the Icelandic Deaf community and ÍTM researchers. ÍTM is now estimated to have about 300 native users, and in total approximately 2,000 users. The total number includes L2 users with late onset of hearing loss, children of deaf adults, foreign L1 signers (ÍTM L2 signers) and hearing signers with various levels of proficiency, the last group comprising somewhere between 1,000 and 1,500 people (Koulidobrova and Sverrisdóttir, 2021).

Automatic processing of ÍTM has not taken place so far and this paper represents one of the first steps in that direction. The task at hand, isolated sign language recognition (ISLR), is not the same as translation, and therefore less practical than a translation system. Nevertheless, for a low-resource sign language such as ÍTM, ISLR is a good starting point. With that in mind, this paper explores how well ISLR can perform in a real low-resource setting, on limited ÍTM data.

The contributions of this paper include:

Proceedings of the Second Workshop on Sign Language Processing, pages 84–95
October 29, 2026 ©2026 Association for Computational Linguistics

1. The first-ever ISLR experiments on ÍTM data
2. Comparison of SPOTER and OpenHands, two open-source frameworks for ISLR
   (a) SPOTER yields significantly better baseline results than OpenHands on all three tasks, outperforming it by 36.36, 14.53 and 6.95 in test accuracy
   (b) Multilingual training in OpenHands improves the test performance by 27.8 and 27.45 percentage points
   (c) Pretraining on ASL data and finetuning on ÍTM data improves SPOTER test performance by 22.73, 23.93 and 14.25 percentage points
   (d) MediaPipe pose estimates in SPOTER performed better than AlphaPose and SDPose pose estimates

2 Related Work

Automatic sign language recognition (SLR) can be split into two tasks, *continuous* or *isolated* SLR (Koller, 2020). In continuous SLR (CSLR), the recognition is performed either directly on the continuous stream or with an intermediate segmentation step, where sign boundaries are identified and the recognition is then performed on isolated signs based on the assumption that a single sign/gloss is contained within the segment boundaries. Isolated SLR (ISLR) is performed on single sign input.

The recognition is then treated as a classification task where information is extracted from signed video input that is processed into a representation suitable for a downstream task (Way et al., 2024). The output labels can be, for instance, words or lemmas in the corresponding spoken language or sign language glosses.

2.1 Video-based Recognition

As observed by Way et al. (2024), in the wake of deep learning, the field of SLR has witnessed a surge of new techniques, models and model architectures. One of the first end-to-end approaches to SLR was the work of Camgoz et al. (2017), who drew on speech recognition and the use of Connectionist Temporal Classification (CTC) algorithms. The novel architecture they proposed for sequence-to-sequence learning, called *SubUNets*, consisted of three layers: a CNN layer to extract spatial features from image input, bidirectional LSTM layers to temporally model those spatial features, and a CTC loss layer on top. This was followed by the *Sign Language Recognition Transformer* (Camgoz et al., 2020), a unified model which was trained to jointly learn CSLR and translation.

2.2 Pose-based Recognition

Pose estimation provides a lower-dimensional alternative to raw video. Pose estimators extract human skeletal keypoints, creating signer-invariant representations that abstract over clothing, skin color, gender etc., focusing on the most relevant parts of the signal and ignoring irrelevant RGB features in the video. Despite these abstractions, the poses remain human-interpretable, but are not entirely anonymous (Battisti et al., 2024). Previous work has predominantly used two pose estimators: OpenPose (Cao et al., 2021) and MediaPipe (Lugaresi et al., 2019). Neither was developed specifically for SLP, but given their lightweight nature compared to raw videos, they have become widespread in SLP.

A recent and comprehensive comparison study of pose estimators is the work of O'Brien et al. (2026). They compare eight different pose estimators and evaluate them on sign language translation (SLT). They show that four pose estimators achieve higher translation scores than MediaPipe, and suggest that alternatives to MediaPipe should be more strongly considered for pose-based SLT. Well-performing and interesting alternatives include Sapiens (Khirodkar et al., 2024) and SDPose (Liang et al., 2026) – which are the best-performing estimators, but have significant compute requirements – and AlphaPose (Fang et al., 2023), which had the third-best translation performance and the fastest compute speed.

2.3 Publicly Available ISLR Frameworks

SPOTER The *Sign Pose-based Transformer* (Boháček and Hrúz, 2022),[2] proposed using pose estimates from Apple's Vision API to train a word-level SLR model. Along with a novel normalization scheme, they presented new data augmentation methods, namely four different spatial augmentations and adding Gaussian noise, which gave the best results. In addition, they demonstrated how well the SPOTER model performs on limited training data compared to an I3D model in the same setting. Their code and data availability has enabled follow-up works such as Azad and Rahman (2026) who introduce several modifications to the

[2] https://github.com/maty-bohacek/spoter

original SPOTER, use MediaPipe poses instead of Vision API, adapt the implementation to Bengali Sign Language data, and substantially improve the previous results on an existing ISLR dataset.

Boháček et al. (2022) improves the original SPOTER approach by replacing the Vision API poses with MediaPipe Holistic. The poses were adapted to the landmarks used in the original SPOTER, in addition to performing hyperparameter search over augmentation parameters. By doing so, they greatly improved the performance on WLASL100, achieving state-of-the-art results for it at the time. This code is, however, not available.

SPOTER only implements one model architecture (Transformer) and neither pretraining strategies nor multilingual training setups are supported.

OpenHands Another important contribution in terms of open-sourced code is the OpenHands library by Selvaraj et al. (2022),[3] (see also NC et al. (2022)) which aimed to exploit key low-resource language insights from traditional NLP for the benefit of sign languages. The open-source framework they released is no longer maintained, but includes pose estimates of existing datasets for five sign languages and more than 1,000 hours of Indian Sign Language data for self-supervised training.

OpenHands supports four model variants: two sequence-based models, RNN and Transformer, and two graph-based models, a spatio-temporal graph convolutional network (ST-GCN) and a sign language GCN (SL-GCN). Based on the results reported by Selvaraj et al. (2022), the graph-based models outperformed the sequence-based models on all of the datasets where accuracy was reported, with the SL-GCN performing best overall.

The framework also provides a multilingual training configuration that allows combining data from multiple sign languages. Two vocabulary strategies are available: a unified vocabulary, where the original classes from each dataset are "normalized" into English glosses, or a combined vocabulary of the original vocabulary of each dataset.

In its default configuration, OpenHands supports 11 datasets for 7 sign languages (American, Chinese, Indian, Turkish, Greek, Argentinian and German).[4] However, three of these datasets are not publicly available. For the other datasets, the authors provide MediaPipe poses.

[3] https://github.com/AI4Bharat/OpenHands

[4] See https://openhands.ai4bharat.org/en/latest/instructions/datasets.html for an overview of the supported datasets.

Number of videos	1,845
Total length (hours:minutes:seconds)	1:44:29
Number of classes	849
Number of signers	33

Table 1: Overall statistics of ÍTM SignWiki dataset

# Classes	# Samples	Proportion
732	2	86.22%
95	3	11.19%
18	4	2.12%
2	5	0.24%
1	6	0.12%
1	8	0.12%

Table 2: Samples per class distribution in the ÍTM SignWiki dataset. Most classes represented by two samples

3 Dataset

We use the *ÍTM SignWiki* dataset presented in Ingimundarson (2026). It is based on the dictionary component of the Icelandic part of *SignWiki*,[5] a web and mobile platform for sign languages and deaf education. The dictionary is bilingual, where ÍTM sign language videos are mapped to Icelandic words or phrases, and currently contains roughly 13,000 signs and phrases. Although the dictionary is open for public input, the vast majority of the recordings originate from the Communication Centre for the Deaf and Hard of Hearing in Iceland (SHH), contributed by people who have been or are currently employed there, and the recordings date from the early 1990s and up until the present day.

3.1 Dataset Profile

An overview of the dataset is given in Tables 1 and 2. As the numbers illustrate, the dataset is very limited in terms of size. Instead of having many examples on a limited set of classes, there are limited examples for many classes. This is in stark contrast to most ISLR datasets, such as the subsets of the MS-ASL dataset (Joze and Koller, 2019), that range from 100–1,000 classes and have, on average, 25.5–57.4 samples per class (189–222 signers).

As far as the signers in the data are concerned, 84.77% are L1 deaf, 10.30% are hearing and 4.93% are L2 deaf. It is also worth noting that 89.7% of the signers are female and 10.3% are male, and all are white, which gives an indication of the overall composition of the data. As the data has been collected over many years, some of the more fre-

[5] https://is.signwiki.org/

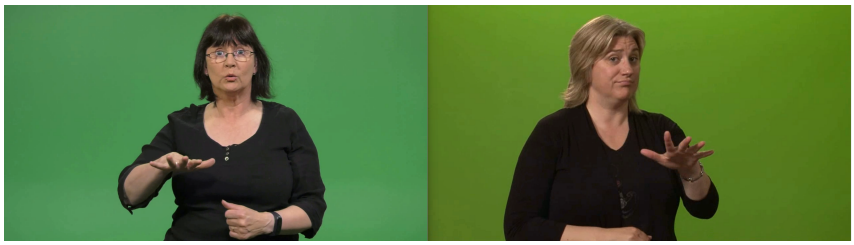

Figure 1: Example of left-/right-handed contrast in the ÍTM SignWiki dataset

quently featured signers have several different appearances, i.e. wearing different clothing, with different hairstyles etc. This would increase the diversity in a video-/RGB-based approach, but is less relevant for pose-based methods.

A further characteristic of the dataset is the fact that the second-most frequently occurring signer in the dataset (Signer 1) is left-handed, the only one out of the 33 signers. The signer appears in 267 out of 1,845 samples. Therefore, many of the classes have a left-/right-hand contrast in addition to the signer difference between training and test samples. In total, 1,075 instances are labeled as right-handed, 202 left-handed, and 568 samples are two-handed symmetrical signs. An example of left-/right-handed contrast is shown in Figure 1.

It is worth noting that a few instances of phrases, as opposed to isolated signs, are included in the data, e.g. *hvað heitir þú* ('what is your name'). The dataset also contains multiple instances of other classes that consist of two or even three signs, such as the word *sérkennari* 'special needs teacher', composed of the signs SÉR ('special'), KENNA ('teach'), and PERSÓNA ('person'). Furthermore, seven of the 849 classes are fingerspelled. It must also be noted that the vocabulary of the dataset is unsystematic and contains, for example, signs for the words *Algeria*, *browned potatoes*, *Lord* (in Christian sense), *feather* and *hippopotamus*.

3.2 Data splits and recognition tasks

Data splits We split the full dataset into a training, validation and test set with 917, 79 and 849 samples, respectively. The difference in size between the validation and test splits is due to dataset constraints. For 38 of the 117 classes that have three samples or more, two samples were retained for training instead of having a validation sample. In many cases, the validation sample is a different recording by the same signer as the training sample, whereas the test sample is drawn from a different signer not represented for that class. Therefore, the performance on the validation set might be misleading with regard to test set performance. This is preferable to validating on an unseen signer while testing on a seen signer for the same class.

Task	Split	# Classes	# Samples
Minimal	Train	22	52
	Validation	22	22
	Test	22	22
Trimmed	Train	117	185
	Validation	79	79
	Test	117	117
Full	Train	849	917
	Validation	79	79
	Test	849	849

Table 3: Recognition tasks based on the ÍTM SignWiki dataset, varying number of classes in the training, validation and test sets

Tasks We define three different recognition tasks with increasing difficulty (varying random chance of success and class balance of training and validation data), see Table 3. The *Minimal* task entails recognizing 22 classes (with ≥ 4 samples per class), the *Trimmed* task has 117 classes (with ≥ 3 samples per class) and the *Full* task has 849 classes (with ≥ 2 samples per class). Moving to the next task means adding more, and more challenging, classes, with the Full task effectively being one-shot recognition over 849 classes.

4 Experiments

We train a series of baselines with OpenHands and SPOTER using only the ÍTM SignWiki dataset (Section 4.1). Then we perform additional experiments on multilingual training (Section 4.2) and a pretraining/finetuning scheme (Section 4.3) where the training data includes other languages as well.

4.1 Baselines

OpenHands We train systems for all four model variants that OpenHands supports: two sequence-

based models, RNN and Transformer (BERT), and two graph-based models, a spatio-temporal graph convolutional network (ST-GCN) and a sign language GCN (SL-GCN) (see Section 2.3).

The framework provides precomputed poses for eight ISLR datasets (listed in Table 8), but also a pipeline to extract MediaPipe poses from video data. We used this pipeline on the ÍTM data. In the native OpenHands setup, the authors then define two different keypoint presets for the pose estimates: *Minimal* (27 2D keypoints of the upper-body, hands and face) or *Top body* (59 keypoints with better coverage of the upper body). Following Selvaraj et al. (2022)'s example scripts we only used the minimal preset in all of our experiments.

In addition to trying different model architectures, we evaluate on all three recognition tasks (see Section 3.2) and either disable or enable data augmentation. Taken together, we train $4 \times 3 \times 2 = 24$ baseline OpenHands models.

In the native OpenHands setup, the validation accuracy is monitored for early stopping, and we followed this example (max epochs=500/1000, patience=80 mode=max). Further hyperparameters were the following: CosineAnnealingLR scheduler with Adam, lr = $1e-3$, and batch size 4/8/16 for *Minimal*/*Trimmed*/*Full*. Models were trained on an A100 or H100 GPU (this is true for all models in our experiments).

SPOTER The best results with SPOTER were achieved with MediaPipe poses as the input representation (see Section 2.3). For this paper, we adapted the SPOTER code to the binary pose format developed by Moryossef et al. (2021a). Inspired by O'Brien et al. (2026) we compare three different pose estimators: AlphaPose (Fang et al., 2023), MediaPipe (Lugaresi et al., 2019) and SDPose (Liang et al., 2026), instead of assuming MediaPipe as a fixture of the experiment.

As described in Section 2, MediaPipe and OpenPose have been the two predominant pose estimators used in SLP in recent years. AlphaPose and SDPose are lesser known in an SLP context. In O'Brien et al. (2026)'s comparison of pose estimators, evaluated on SLT, AlphaPose ranked third, and was found to be more computationally efficient than the better-performing estimators. It is primarily the low-latency benefit of AlphaPose that makes it appealing for the purposes of this project. While working with a low-resource language does not necessarily entail limited computational resources, a lightweight solution with competitive performance would be advantageous.

To extract the poses, we used the *video-to-pose* repository by O'Brien et al. (2026),[6] which at the time of writing supports eight pose estimators in total. In the SPOTER implementation, a total of 54 body landmarks are extracted, including five head landmarks (eyes, ears, and nose) and 21 body landmarks that represent body joints. This results in 54 2D points and an 108 dimensional feature vector for each frame. Following Boháček and Hrúz (2022) we train for 350 epochs and add Gaussian noise to the training set (along with other data augmentation techniques the framework provides).

To summarize, we trained SPOTER baselines for all three recognition tasks (see Section 3.2) and for three different pose estimators ($3 \times 3 = 9$ baseline SPOTER models).

4.2 Multilingual training

We train additional OpenHands models on more datasets in other sign languages (see Table 8 in Appendix A for the full list of datasets).

Vocabulary strategies We experiment with either keeping the vocabularies of all datasets separate (*original*) or unifying them into a single, normalized vocabulary (*unified*). In the first setting, the language code (ISO) of each sign language included is prepended to the class label as a one-hot encoded vector, i.e. *ice__* for ÍTM.

In the second setting, the vocabulary of each non-English/non-ASL dataset is normalized to English glosses. Here, the ÍTM subset has only 817 classes as the normalization allows to combine different sign variants of the same sign, of which there are several examples in the dataset, with the same normalized gloss. See Appendix A for a more detailed explanation.

This results in a single normalized vocabulary and predictions are therefore made with normalized glosses, which can then be mapped back to the original vocabulary, but not to separate variants. As the unified vocabulary experiment effectively is a different version of the *Full* task, we keep those results separate and present them in Appendix A.

These experiments use the multilingual training feature of OpenHands, described in NC et al. (2022), but only the unified vocabulary, but the experiments were inspired by example configs

[6] https://github.com/ZurichNLP/video-to-pose

for multilingual training provided with the framework.[7] We used the following hyperparameters: CosineAnnealingLR scheduler with Adam, learning rate $1e-3$, batch size 64, max epochs 200, with early stopping on validation accuracy (max, patience 30).

Due to time and resource constraints, we evaluate these multilingual models only on the *Full* recognition task.

4.3 Pretraining/finetuning scheme

SPOTER does not support pretraining out-of-the-box. For this paper, we adapted the framework to support pretraining (either with the encoder frozen for a specified number of epochs or unfrozen from the start) and finetuning.

For this experiment, a SPOTER model was pretrained on the ASL Citizen dataset (Desai et al., 2023). It is a community-sourced dataset for ASL that has nearly 84,000 videos filmed by 52 signers and 2,731 classes. One motivation for using it was that MediaPipe poses in pose format for the dataset were already available, though any comparable ISLR dataset for a higher-resource sign language would have been a viable alternative. As MediaPipe outperformed the other two pose estimators in the baseline experiments (see Section 5.1), this setup was exclusively tested with MediaPipe poses. As the aim here was simply for the encoder to learn general representations of ASL and not to achieve the best results on the dataset, the training was limited to 30 epochs (test accuracy 40.28).

For the finetuning we tested two approaches: one where the encoder was frozen for the first 30 epochs and only the decoder and head trained, and another where the full model was finetuned from the beginning.

Thus the model is pre-trained on ASL data and finetuned on ÍTM, with the encoder either frozen from the start or unfrozen; exclusively with MediaPipe poses (3×2 experiments).

4.4 Evaluation Metrics

We report test accuracy and validation accuracy, given that the validation set is somewhat particular and limited in two ways (see Section 3.2): on the one hand it covers only 79 classes (compared to the total of 849) and on the other hand, the validation samples are mostly with the same signer as (one of) the training sample(s), whereas the signer in the test set is always unseen for that particular sign.

[7] See https://github.com/AI4Bharat/OpenHands/tree/main/-examples/configs/multilingual

Task	Model	Augmentation	Validation	Test
Minimal	LSTM	✓	18.20	9.09
		-	9.1	0.00
	Transformer	✓	22.70	9.09
		-	31.8	9.09
	ST-GCN	✓	13.60	9.09
		-	40.9	**13.64**
	SL-GCN	✓	9.10	4.54
		-	54.5	9.09
Trimmed	LSTM	✓	3.80	0.00
		-	2.5	0.85
	Transformer	✓	3.80	1.70
		-	7.6	2.56
	ST-GCN	✓	21.5	2.56
		-	51.9	3.42
	SL-GCN	✓	10.10	4.27
		-	59.5	**9.40**
Full	LSTM	✓	7.60	0.00
		-	10.1	0.23
	Transformer	✓	2.50	0.00
		-	3.8	0.11
	ST-GCN	✓	22.8	0.71
		-	26.6	1.06
	SL-GCN	✓	17.70	**1.41**
		-	30.4	0.82

Table 4: Accuracy of OpenHands baseline models trained on ÍTM data only (✓=with augmentation, -=only normalization)

5 Results

This section presents the initial ISLR experiments we performed with the two frameworks. An overview of the best results is shown in Table 7 and an additional table for the multilingual results is included in Appendix A.

5.1 Baselines (ÍTM Data Only)

OpenHands Table 4 shows the performance of all OpenHands baselines. The graph-based models generally outperform the LSTM and Transformer models, with a GCN variant achieving the best test accuracy of 13.64% / 9.40% / 1.41% on the *Minimal* / *Trimmed* / *Full* tasks respectively. Only one configuration yielded more than 10% accuracy on the *Minimal* task; an ST-GCN model without any data augmentation. For reference, the best OpenHands baseline model is repeated in Table 7.

Across the baselines, disabling augmentation yields slightly better test accuracy in 8 of 12 pairs and in general higher validation accuracy. However, since the test sets are rather small (22, 117 and 849 examples) these are in fact minor differences.

Task	Poses	Validation	Test
Minimal	MediaPipe	72.72	**50.00**
	AlphaPose	63.63	45.45
	SDPose	63.63	31.82
Trimmed	MediaPipe	72.15	**23.93**
	AlphaPose	59.49	18.80
	SDPose	59.49	17.09
Full	MediaPipe	67.09	**8.36**
	AlphaPose	53.16	3.06
	SDPose	55.70	4.00

Table 5: Accuracy of SPOTER baseline models trained on ÍTM data only, varying the pose estimation system

Task	Setting	Val	Test	ΔBaseline
Minimal	Frozen	95.45	**72.73**	+22.73
	Unfrozen	95.45	59.09	+9.09
Trimmed	Frozen	82.27	**47.86**	+23.93
	Unfrozen	67.09	26.50	+2.57
Full	Frozen	78.48	**22.61**	+14.25
	Unfrozen	68.35	13.43	+5.07

Table 6: Test accuracy of a SPOTER model pre-trained on ASL and finetuned on ÍTM compared to the baseline in Table 5 (Frozen=Pre-trained encoder frozen for first 30 epochs, Unfrozen=Full model finetuned from start)

SPOTER Table 5 shows the performance of all SPOTER baselines. MediaPipe poses consistently outperforms other estimators, for example the test accuracy of MediaPipe is roughly 5 percentage points higher than AlphaPose on all three tasks. In general, the SPOTER baselines show considerably higher accuracy on the test set than comparable OpenHands baselines (see above). For reference, the best SPOTER baseline model (only MediaPipe results) is repeated in Table 7.

5.2 Multilingual Training

Multilingual training results with the original vocabulary are shown in Table 7, for a direct comparison with baseline scores. Individual per-dataset scores are reported in Table 8 in Appendix A. Multilingual training results only concern the *Full* task. Adding multilingual training outperforms the best OpenHands and SPOTER baselines. For example, the test accuracy of the best SPOTER baseline is 8.36, while the test accuracy of the best multilingual model is 28.86. Unifying the multilingual vocabulary (as opposed to keeping separate vocabularies) yields higher accuracy, 29.21, but on a slightly smaller vocabulary (see Section A.1).

5.3 Pretraining / finetuning scheme

Pretraining / finetuning results are shown in full in Table 6. Pretraining on ASL data also outperforms the best baselines by at least 10 percentage points in test accuracy. For instance, on the *Full* task, the best SPOTER baseline achieves 8.36 test accuracy, while the best finetuned model achieves 22.61 accuracy. Furthermore the results demonstrate that freezing the pre-trained encoder at the beginning of finetuning increases the test accuracy by at least 10 percentage points.

6 Discussion

Baselines: SPOTER vs. OpenHands When only using ÍTM data, SPOTER clearly outperforms Openhands on all three tasks: 50.00 vs. 13.64, 23.93 vs. 9.40 and 8.36 vs. 1.41 test accuracy (see Section 5.1). These margins should be read with the size of the test sets in mind. On the *Minimal* task each prediction is worth $1/22 = 4.54$ percentage points, and the gap amounts to 11 vs. 3 correct predictions; on *Trimmed* it is 28 vs. 11 out of 117. The *Full* task, with 71 vs. 12 correct predictions out of 849, therefore carries most of the evidence, and the smaller tasks should not be over-interpreted.

We emphasize that this is a comparison of two frameworks as they are distributed, not of two architectures. Input representation, normalization, augmentation, optimization and checkpoint selection all differ at the same time, and our experiments do not isolate these factors. We therefore offer hypotheses rather than explanations.

For example, OpenHands and SPOTER reduce the full set of MediaPipe keypoints in different ways. OpenHands offers 27-point and 59-point presets, but our OpenHands models use only the 27 point preset, following Selvaraj et al. (2022). SPOTER, on the other hand, uses 54 2D keypoints. This difference in keypoint resolution may in part explain the difference in performance between the frameworks: a coarser hand representation could plausibly mean a disadvantage for OpenHands. Rerunning OpenHands with the 59-point preset would test this directly.

Second, the frameworks normalize differently. OpenHands applies a single shoulder-referenced centering and scaling to the whole skeleton, so hand landmarks occupy a small region of the normalized space, whereas SPOTER normalizes body and hands separately, distorting hand keypoints to

Task	Random	Framework	Model	Setting	Val	Test	Time
Minimal	4.54	OpenHands	ST-GCN	Baseline (No Augmentation)	40.9	13.64	00:01:54
		SPOTER	Transformer	Baseline (MP)	72.72	50.00	00:02:18
		SPOTER	Transformer	ASL-Finetuned (MP)-Frozen	95.45	**72.73**	*05:22:24
Trimmed	0.85	OpenHands	SL-GCN	Baseline (No Augmentation)	59.5	9.40	00:07:39
		SPOTER	Transformer	Baseline (MP)	72.15	23.93	00:18:06
		SPOTER	Transformer	ASL-Finetuned (MP)-Frozen	82.27	**47.86**	*05:25:13
Full	0.12	OpenHands	SL-GCN	Baseline (Augmentation)	17.70	1.41	01:14:32
		OpenHands	SL-GCN	Multilingual Original	68.2	**28.86**	08:15:16
		SPOTER	Transformer	Baseline (MP)	67.09	8.36	01:31:25
		SPOTER	Transformer	ASL-Finetuned (MP)-Frozen	78.48	22.61	*06:07:20

Table 7: Validation and test accuracy on the three ÍTM recognition tasks for OpenHands and SPOTER. (Baseline=best baseline score, Random=Random classification accuracy, MP=MediaPipe Holistic poses, Time=Training time as HH:MM:SS, *=Combined training time (ASL Citizen pretraining took 05:21:11 hours)

a lesser degree. Normalization is intimately tied to generalization; because the test signer is always unseen for a given class (Section 3.2), our test set specifically measures signer-invariant generalization. Consistent with this, SPOTER retains a much larger share of its validation accuracy on the test set (69%/33%/12% across the three tasks) than the best OpenHands models do (33%/16%/8%). Both frameworks overfit to the seen signer; OpenHands does so considerably more. The importance of preprocessing choices of this kind for pose-based SLP has been noted before (Coster et al., 2023; O'Brien et al., 2026).

Third, the training and model selection protocols differ. SPOTER trains for a fixed number of epochs, saves the two checkpoints with highest training and validation accuracy every ten epochs, and evaluates over all of them. Our OpenHands runs use early stopping and checkpoint selection on validation accuracy, which on the *Full* task is computed over 79 of 849 classes, and a single checkpoint was then chosen manually from k saved checkpoints. Model selection is thus both noisier and more weakly related to the target task for OpenHands, and we did not evaluate all saved checkpoints for comparison.

Fourth, and in our view most informative, the graph-based models appear to be data-starved rather than unsuited to the task. The same SL-GCN that reaches 1.41 on the *Full* task with ÍTM data alone reaches 28.86 once other sign languages are added (Section 5.2). Selvaraj et al. (2022) report their graph models performing best on datasets with tens of samples per class; ÍTM SignWiki offers one or two. In this respect our results are in line with Boháček and Hrúz (2022), who show SPOTER learning effectively from small training sets, although their comparison was against a video-based I3D model that must first learn general properties of human motion, whereas both frameworks compared here operate on poses. Our results are therefore consistent with their claim but do not test it under the same conditions.

Finally, although both frameworks use MediaPipe, the poses were extracted with different pipelines: the extraction script shipped with OpenHands in one case and *video-to-pose* (O'Brien et al., 2026) in the other. Differences in version or configuration cannot be ruled out as a contributing factor.

SPOTER: Choice of pose estimator MediaPipe clearly outperformed AlphaPose and SDPose. It is worth reiterating that this was not a comparison of the full pose estimates but of the customized (reduced) SPOTER format. However, the format was the same for all three estimators and the training schemes identical. As mentioned in Section 2, a comparison study of pose estimators performed by O'Brien et al. (2026) revealed that three other pose estimators performed better than MediaPipe on SLT. In our experiments, the downstream task is different, recognition instead of translation, and the set of keypoints is reduced. Nevertheless, MediaPipe appears better suited to ISLR. This is in line with findings of the comparative studies of Coster et al. (2023) and Moryossef et al. (2021b) of pose estimators for SLR, where MediaPipe outperformed OpenPose and MMPose. Our results are, to the best of our knowledge, the first comparison of MediaPipe, AlphaPose and SDPose for SLR.

Cross-lingual transfer effects Both multilingual training and fine-tuning a pretrained model improved the performance, meaning that both con-

stitute genuine cross-lingual transfer effects from higher-resourced sign languages to a low-resource one. Multilingual training raises the best OpenHands result on the *Full* task from 1.41 to 28.86, a twentyfold increase, while ASL pretraining raises the best SPOTER result from 8.36 to 22.61, and by 14–24 percentage points across the three tasks. On the *Full* task the multilingual model is ahead (245 vs. 192 correct predictions of 849).

These two numbers should not be read as a ranking of the two strategies. They come from different frameworks, whose baselines already differ by a factor of six; from different source data, eight datasets in six sign languages in one case and a single ASL dataset in the other; and from different label spaces at inference. The multilingual model predicts over all 5,732 classes of the concatenated dataset, and 137 of its 849 test predictions fall on classes belonging to other sign languages (see Appendix B); these are incorrect by construction, so its ÍTM accuracy is measured under a handicap that the pretrained model does not face. A controlled comparison would require training SPOTER jointly on ASL Citizen and ÍTM and evaluating both schemes within one framework, and the multilingual configuration on the two smaller tasks, neither of which we were able to do.

7 Conclusion

In this paper we explore the performance of two publicly available SLR frameworks in a low-resource setting, using the first ISLR dataset for ÍTM. The dataset is very limited with regard to the intended use in a machine learning task; it has many classes (849) and few samples per class, only two samples for the majority of classes (732). The experiments therefore explored how well SLR can perform in a real low-resource setting.

We demonstrate that multilingual learning and cross-lingual transfer can benefit lower-resource sign languages. In the OpenHands experiments, the multilingual training scheme achieved the best results on the full dataset. In the case of SPOTER, we show that SPOTER learns effectively from very small training sets, corroborating the findings of Boháček and Hrúz (2022). Additionally, finetuning on ÍTM after pretraining on ASL Citizen data yielded the best overall results on the original ÍTM vocabulary on two out of three tasks and demonstrates the effectiveness of a strong pre-trained checkpoint that can be finetuned to different tasks. The multilingual model with original vocabulary in OpenHands yields the best performance on the full (most challenging) task.

The repositories of the adapted versions of the two frameworks have been made publicly accessible on GitHub, in the hope that they may prove useful to other researchers,[8,9] and the dataset will soon be published as well.

8 Limitations

Framework choice Both of these frameworks are from 2022 and comparison with at least one newer ISLR method would have been preferable, but no more recent, publicly available code could be found, except extensions of SPOTER.

Error analysis Beyond general accuracy measures, we present no error analysis. Given the composition of the dataset, it could for instance be insightful to analyze errors based on sign types, e.g. one-handed vs. two-handed, fingerspelled signs and multi-sign compounds (see Section 3.1), or per-signer performance.

Baseline tuning The discrepancy in performance in the OpenHands baselines with regard to data augmentation, where the performance without augmentation was in general better, was unexpected and would need further inspection. The augmentation techniques used followed the examples of the OpenHands authors, but could perhaps be adjusted better to the ÍTM dataset.

Variability We only present one single run for each training configuration. Training several models with different random seeds would make our results and conclusions drawn from them more robust.

Applicability Although the experiments yielded meaningful results it should nonetheless be stressed that the results do not yet constitute practically applicable performance. ISLR systems can be applied to tasks such as looking for signs in videos or dictionaries, but none of the models we trained could be immediately deployed in such a setting. This would require substantially greater amounts of training data as well as a larger vocabulary.

[8] https://github.com/ZurichNLP/openhands-itm
[9] https://github.com/ZurichNLP/spoter-pose

9 Acknowledgements

We would like to thank three anonymous reviewers for useful comments. MM received funding from the SIGMA project (grant no. G-95017-01-07), supported by the Digital Society Initiative (DSI) at the University of Zurich.

The language of the abstract and discussion sections was refined with a Claude agent.

References

Sayad Ibna Azad and Md. Atiqur Rahman. 2026. *BdSL-SPOTER: A Transformer-Based Framework for Bengali Sign Language Recognition with Cultural Adaptation*, page 304–315. Springer Nature Switzerland.

Alessia Battisti, Emma van den Bold, Anne Göhring, Franz Holzknecht, and Sarah Ebling. 2024. Person Identification from Pose Estimates in Sign Language. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 13–25, Torino, Italia. ELRA and ICCL.

Matyáš Boháček, Zhuo Cao, and Marek Hrúz. 2022. Combining Efficient and Precise Sign Language Recognition: Good pose estimation library is all you need. *Preprint*, arXiv:2210.00893.

Matyáš Boháček and Marek Hrúz. 2022. Sign Pose-based Transformer for Word-level Sign Language Recognition. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 182–191.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. *Preprint*, arXiv:2003.13830.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.

Mathieu De Coster, Ellen Rushe, Ruth Holmes, Anthony Ventresque, and Joni Dambre. 2023. Towards the extraction of robust sign embeddings for low resource sign language recognition. *Preprint*, arXiv:2306.17558.

Aashaka Desai, Lauren Berger, Fyodor O. Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé, Alex X. Lu, Naomi Caselli, and Danielle Bragg. 2023. ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2023. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7157–7173.

Finnur Ágúst Ingimundarson. 2026. Isolated Sign Language Recognition for Icelandic Sign Language (ITM): Experiments in a Low-resource Setting. Master's thesis, University of Zurich.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Hamid Reza Vaezi Joze and Oscar Koller. 2019. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. *Preprint*, arXiv:1812.01053.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2024. Sapiens: Foundation for Human Vision Models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part IV*, page 206–228, Berlin, Heidelberg. Springer-Verlag.

Oscar Koller. 2020. Quantitative Survey of the State of the Art in Sign Language Recognition. *Preprint*, arXiv:2008.09918.

Elena Koulidobrova and Rannveig Sverrisdóttir. 2021. How to Ensure Bilingualism/Biliteracy in an Indigenous Context: The Case of Icelandic Sign Language. *Languages*, 6(2).

Shuang Liang, Jing He, Chuanmeizhi Wang, Lejun Liao, Guo Zhang, Yingcong Chen, and Yuan Yuan. 2026. SDPose: Exploiting Diffusion Priors for Out-of-Domain and Robust Pose Estimation. *Preprint*, arXiv:2509.24980.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *Preprint*, arXiv:1906.08172.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2021a. pose-format: Library for viewing, augmenting, and handling .pose files. `https://github.com/sign-language-processing/pose`.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgöz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling. 2021b. Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition. *Preprint*, arXiv:2104.10166.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. 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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gokul NC, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh M Khapra. 2022. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Catherine O'Brien, Gerard Sant, and Mathias Müller. 2026. Convenience code for installing and using several pose estimation systems. `https://github.com/ZurichNLP/video-to-pose`.

Catherine O'Brien, Gerard Sant, Mathias Müller, and Sarah Ebling. 2026. Evaluation of Pose Estimation Systems for Sign Language Translation. In *Proceedings of the LREC 2026 12th Workshop on the Representation and Processing of Sign Languages: Language in Motion*, pages 371–386, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.

Andy Way, Lorraine Leeson, and Dimitar Shterionov. 2024. *Sign Language Machine Translation*. Springer Nature Switzerland.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

A Multilingual Results

A.1 Multilingual Training

In Table 8, we present the results of the test set evaluation for both vocabulary approaches, unified and original, of the multilingual training, with two different architectures, along with a comparison of the results reported by Selvaraj et al. (2022) and NC et al. (2022).

The multilingual training outperforms the best accuracy reported by Selvaraj et al. (2022) and NC et al. (2022) on five datasets, ASLLVD, GSL, INCLUDE, LSA64 and WLASL, and in four cases out of five with the original vocabulary and an SL-GCN. The discrepancy in performance on the RWTH-PHOENIX dataset is notable. This data differs from the other datasets, as the poses have been extracted from low-resolution image files instead of videos that are furthermore from continuous data.

Our setup is not completely comparable to the work of NC et al. (2022), where the unified vocabulary approach is used. We include the ÍTM dataset, whereas they include – in addition to the datasets we use – one Turkish Sign Language and two Chinese Sign Language datasets, that were not available to us. They explore different pretraining and finetuning options: one where the unified vocabulary (referred to as *MultiSign-ISLR dataset* is pretrained on Indian Sign Language data and finetuned individually, another model pretrained on *SignCorpus*, a large pretraining dataset of 4.6K hours of data from 10 different sign languages, and finetuned individually, and lastly pretrained on *SignCorpus* and finetuned jointly on all of the sign languages.

Original vs. unified vocabulary As an alternative to a simple concatenation of all sign classes from all languages, we consider a **unified** vocabulary. As an example from our dataset, two different classes for sign variants of *allt fínt* 'all good' are mapped to the same normalized gloss, ALL_GOOD. Including ÍTM in this experiment is admittedly only synthetic, as the original data is not glossed nor are there English glosses available. To synthesize this, the classes (words) in the ÍTM data were machine-translated to English, manually reviewed, and compared to normalized glosses for the other datasets – with 40.8% overlap between the translated ÍTM glosses and the other normalized glosses (see NC et al. (2022) for their approach to the normalization).

	Original		Unified		OpenHands	
Dataset	**ST-GCN**	**SL-GCN**	**ST-GCN**	**SL-GCN**	Selvaraj et al. (2022)	NC et al. (2022)
ASLLVD	50.92	**55.64**	49.49	55.38	-	50.1
AUTSL	90.89	91.42	90.65	91.47	91.9	**92.1**
GSL	94.94	**96.08**	94.40	95.62	95.4	93.6
INCLUDE	96.07	96.57	94.85	**97.55**	93.5	96.3
LSA64	96.56	**98.75**	96.56	96.87	97.8	97.5
MSASL	62.20	65.89	61.48	66.44	-	**67.4**
RWTH	0.41	1.03	0.20	0.82	-	**48.5**
WLASL	44.27	**47.43**	44.92	47.01	30.6	46.6
ÍTM (*Full*)	20.84	28.86	22.12	**29.21**	-	-

Table 8: Per-dataset test accuracy for two multilingual models, 1) original vocabulary (5,732 classes) and 2) unified vocabulary (4,261 classes), along with the best standalone results (all SL-GCN) reported by Selvaraj et al. (2022)

Our results demonstrate how the unified vocabulary outperforms the original vocabulary and demonstrate the effectiveness of crosslingual transfer. It must, however, be stressed that the unified vocabulary has fewer labels than the original vocabulary since several classes that have variants in the dataset are collapsed into a single class in the unified vocabulary. The evaluation is therefore done on a test set with 817 classes instead of the full 849, as in the original vocabulary. And consequently, the results of the two vocabulary approaches are not fully comparable and the difficulty of the tasks is a confounding variable.

In a similar vein, regarding the unified vocabulary approach, we emphasize that these results are based on synthetic ÍTM glosses that have not been verified by an ÍTM expert. Therefore, the results are only an indication of the benefits of this approach, whereas the results with the original vocabulary provide more direct evidence of the benefits of multilingual training.

B Multilingual Prediction Analysis

If the original vocabulary is preserved in the multilingual OpenHands model, then the recognition is performed over a large multilingual label space, and predictions in other languages are possible. On the full test set there are 137 cases of another sign language being predicted and the overall distribution is shown in Table 9. This distribution is in line with the number of classes in the included datasets, and ASL has the largest part of the vocabulary in the concatenated dataset. It is worth considering whether these foreign predictions are in fact correct for the other sign languages. As an example of this, one might for instance look at the prediction for the ÍTM class *frumskógur* 'jungle', which is the plural form of tree in ASL (ase__tree_pl), or the ASL prediction *boy* for the ÍTM class 'man'. The similarity at the level of the word label or gloss alone suggests that the ASL predictions may be correct. However, this might also simply be a coincidence and there are multiple examples of foreign predictions with no label similarity to the ÍTM class, although the underlying signs may nonetheless be phonologically similar. This would require a closer inspection and comparison of the videos.

ISO	SL	Count
ase	American	88
tsm	Turkish	16
gsg	German	1
ins	Indian	20
gss	Greek	9
aed	Argentinian	3
icl	Icelandic	712

Table 9: Multilingual test prediction distribution with original vocabulary (ST-GCN)