Logos as a Well-Tempered Pre-train for Sign Language Recognition

Ilya Ovodov, Petr Surovtsev, Karina Kvanchiani,
Alexander Kapitanov, Alexander Nagaev
SberAI
{iovodov, petr.surovcev, karinakvanciani, kapitanovalexander, sashanagaev1111}@gmail.com

Abstract

This paper examines two aspects of the isolated sign language recognition (ISLR) task. First, although a certain number of datasets is available, the data for individual sign languages is limited. It poses the challenge of cross-language ISLR model training, including transfer learning. Second, similar signs can have different semantic meanings. It leads to ambiguity in dataset labeling and raises the question of the best policy for annotating such signs. To address these issues, this study presents Logos, a novel Russian Sign Language (RSL) dataset, the most extensive available ISLR dataset by the number of signers, one of the most extensive datasets in size and vocabulary, and the largest RSL dataset. It is shown that a model, pre-trained on the Logos dataset can be used as a universal encoder for other language SLR tasks, including few-shot learning. We explore cross-language transfer learning approaches and find that joint training using multiple classification heads benefits accuracy for the target low-resource datasets the most. The key feature of the Logos dataset is explicitly annotated visually similar sign groups. We show that explicitly labeling visually similar signs improves trained model quality as a visual encoder for downstream tasks. Based on the proposed contributions, we outperform current state-of-the-art results for the WLASL dataset and get competitive results for the AUTSL dataset, with a single stream model processing solely RGB video. The source code, dataset, and pre-trained models are publicly available.

1 Introduction

Sign languages (SL) are visual-spatial signals for communication among deaf communities. Although national sign languages are mostly associated with national spoken languages, they are distinct languages with their own grammar and vocabulary. Primarily, signs are expressed by hand shape and motion (manual components of sign), but also with a great aid of motion of mouth, head, eyes, and the body (non-manual components). The term "gloss" is used to refer to the word that signifies the sign. Generally, this word represents the sign's meaning. Still, one sign can be translated by several words and vice versa, so glosses should be considered just as word labels for signs.


(a) Летать (to fly)

(b) Крыло (wing)

(c) Овца (sheep)

(d) Разгневанный (angry)

Figure 1: Sample frames from Russian Sign Language dataset Logos: (a,b) and (c,d) are visually similar signs (VSSigns).

The problem of computer sign language recognition and translation has a practical application with significant social impact because it can help deaf and hearing people communicate. On the other hand, it is a challenging scientific problem located at the junction of computer vision and natural language processing areas.

The presented work deals with the isolated sign

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24352–24365
November 4-9, 2025

language recognition (ISLR) problem, i.e., the classification of videos that contain only one sign each. The ISLR task has not only independent significance but is also important for building a more practical continuous sign language translation (CSLT) solution (Chen et al., 2022; Wei and Chen, 2023; Zuo et al., 2024).

A significant obstacle to building SLR solutions is a shortage of training data (Gokul et al., 2022; Papadimitriou and Potamianos, 2023). While a number of annotated SL datasets exist, they represent different sign languages (Table 1), and dataset corpora for many individual languages are insufficient. It highlights the task of using cross-lingual data. Some researchers utilize cross-lingual training either by combining two or more rather small datasets or using a multilingual dataset. The presented study examines the ability of a model trained on an extensive SL dataset of one language to serve as an encoder for SL tasks for other sign languages, and compares different approaches to it.

This paper presents an extensive Russian Sign Language (RSL) dataset, Logos, one of the largest existing sign language datasets in terms of volume and vocabulary size and the largest in terms of the number of signers. We show that a model pre-trained on the Logos dataset can be successfully transferred to another language SLR tasks, including few-shot learning. The dataset size is critical, and the effect degrades if a smaller dataset is used for pre-train. Next, we compare transfer learning methods and find that simultaneous training with the large dataset using multiple classification heads for different languages benefits the target language SLR models the most, compared to other transfer learning methods.

Another problem with SLR is that signs with similar handshapes and motions can have various semantic meanings. Such signs can be either strictly indistinguishable (the cases of polysemy or homonymy) or only distinguishable by their constituent non-manual features (Zuo et al., 2023; Hu et al., 2021b), see Figure 1. In the latter case, such signs can be viewed as minimal pairs for SL. The difference between the individual signers' manners blurs the boundary between non-manual features and makes such signs practically indistinguishable out of context. For this reason, while some researchers consider such signs to be separate, others consider them to be the same (Ebling et al., 2015). This paper calls such hardly distinguishable signs "*visually similar signs*" (*VSSigns*). This concept includes both polysemy and minimal pair cases. While they differ from a linguistic perspective, it makes sense to combine them for machine learning purposes.

Different SL datasets have VSSigns annotated with either different or similar labels. To the best of our knowledge, no studies have examined the impact of the VSSigns annotation approach on resulting SLR models. We explore its effect in this work using the Logos dataset, which has both *ungrouped gloss* and *grouped VSSign* annotations. We find that VSSigns grouping benefits the SLR model.

The key contributions of this work are:

- We present Logos, a new publicly available Russian Sign Language ISLR dataset. It is the most extensive available ISLR dataset by the number of signers and one of the largest datasets while also the largest RSL dataset in size and vocabulary. The dataset's key feature is an explicit annotation of visually similar sign (VSSign) groups.

- Using the Logos dataset, we show that explicitly grouping VSSign labels benefits trained model quality as a video encoder for downstream tasks like transfer learning to other sign languages.

- We show that a model, pre-trained on the proposed Logos dataset can be transferred to another language SLR tasks, including few-shot learning. We compare transfer learning methods and demonstrate that the method of cross-lingual multi-dataset co-training with multiple language-specific classification heads improves SL models for low-resource datasets the most, compared to the conventional "pre-train and finetune" method.

- Based on the described contributions, we obtain recognition accuracy for the American Sign Language dataset WLASL, superior to state-of-the-art (SOTA), with a single stream model processing solely RGB video.

The research was conducted in cooperation with the "All-Russian Society of the Deaf" (VOG). VOG experts and professional sign language interpreters participated at every stage of the Logos dataset creation. We also engaged deaf consultants in developing training strategies to apply considerations to specific solutions. Additionally, some of our

researchers completed formal courses on RSL to enhance their knowledge in this domain.

The source code, dataset, and pre-trained models are publicly available[1].

2 Related Works

2.1 Isolated Sign Language Recognition

In recent years, a group of approaches for ISLR tasks rely on using RGB input data. Then, either 2D convolutional neural network (CNN) is applied to extract individual frames' features, followed by LSTM for the temporal component processing (Koller et al., 2019), or the spatial and temporal components are simultaneously processed using 3D CNN (Papadimitriou and Potamianos, 2023; Zuo et al., 2023; Albanie et al., 2020; Huang et al., 2018; Li et al., 2020; Joze and Koller, 2018). After the proliferation of transformers, transformer-based image and video processing architectures were applied (Kapitanov et al., 2023; Kvanchiani et al., 2024). In addition to the RGB input, a depth map can be used (Jiang et al., 2021; Zuo et al., 2023).

Another group of approaches utilizes pose (skeleton) keypoints and face landmarks generated by available frameworks (Hrúz et al., 2022; Jiang et al., 2021; Miah et al., 2023; Papadimitriou and Potamianos, 2023; Ryumin et al., 2023). The skeleton keypoints can be represented as a sequence of heatmaps and processed similarly to video data (Zuo et al., 2023). A series of methods build a graph based on physical skeleton connections and explore Graph Convolutional Networks (GCNs) (Hu et al., 2021a, 2023; Patra et al., 2024; Zhao et al., 2023; Jiang et al., 2021).

Most current SOTA SLR models are multi-stream and multi-modal and combine more than one of the methods listed above (Hrúz et al., 2022; Zuo et al., 2023; Jiang et al., 2021; Miah et al., 2023; Papadimitriou and Potamianos, 2023; Ryumin et al., 2023).

Nevertheless, we focus our research on purely frontal RGB video, as this setup is most relevant for practical applications. Many possible real-world scenarios, such as educational tools for RSL and ASL learning, video conferencing, and public kiosk systems in environements like metro stations or airports, typically involve users facing a single RGB camera.

[1]https://github.com/ai-forever/logos

2.2 ISLR Datasets

The ISLR datasets differ in several aspects: language, collection method, size, vocabulary size, number of signers (see Table 1). The most common method of dataset collection is recording invited signers in laboratory conditions. However, this approach generally results in insufficient scene and signer variety, requiring the authors to record each video individually. Web scrapping of SL videos is rather effective and results in more diverse datasets. However, its serious problem is the absence of consent from the video owner and person represented in the video on the usage of the video as a part of the dataset. Albanie et al. (2020; 2021) prepared the British SL datasets using BBC TV programs with SL translation. The datasets are large but have limited scene variety and number of signers, and they mostly only have automatic annotation. Collecting video from SL experts using a web crowdsourcing platform has no problem with signers' consent and provides much more diverse footage. We have used this approach for our work.

Vocabulary size is critical for building a production-quality SLR model. We suppose that practically useful models must recognize over 1,000 glosses. Therefore, a massive number of video samples is needed to simultaneously satisfy both the requirements of a large number of glosses and of samples per gloss. Number of diverse signers is also important. As seen from Table 1, only a few datasets meet these requirements.

2.3 The VSSigns Problem

Some SL signs with different semantic meanings either can be considered as strictly indistinguishable (the cases of polysemy or homonymy) or can have similar handshapes and motions, but can only be distinguished by their constituent non-manual features (Zuo et al., 2023; Hu et al., 2021b), see Figure 1. The latter can be viewed as minimal pairs for SL. The border between polysemy and minimal pair signs can be blurred due to the individual signers' manners, making even signs with different non-manual features practically indistinguishable out of context. For this reason, while some researchers consider such signs to be separate, others consider them to be the same (Ebling et al., 2015). On the contrary, even the same sign can have distinct manual features depending on the context, i.e., a question vs. a statement (Mukushev et al., 2020). It results in ambiguity in the annotation of such

Dataset	Method	Language	Samples	Signers	Glosses	VSSigns
DEVISIGN-L (Wang et al., 2016)	lab	Chinese (CSL)	24,000	8	2,000	–
SLR500 (Huang et al., 2018)	lab	Chinese (CSL)	125,000	50	500	–
MS-ASL (Joze and Koller, 2018)	web	American (ASL)	25,513	222	1,000	grouped
SMILE (Ebling et al., 2018)	lab	Swiss German	9,000	30	100	–
BosphorusSign22k (Özdemir et al., 2020)	lab	Turkish (TİD)	22,542	6	744	grouped
AUTSL (Sincan and Keles, 2020)	lab	Turkish (TİD)	38,336	43	226	–
WLASL (Li et al., 2020)	web	American (ASL)	21,083	119	2,000	–
BSLDict (Momeni et al., 2020)	lab	British (BSL)	14,210	28	**9,283**	addressed
K-RSL (Imashev et al., 2020)	lab	Kazakh-Russian	28,250	10	600	addressed
BSL-1K (Albanie et al., 2020)	TV	British (BSL)	273,000[1]	40	1,064	–
INCLUDE (Sridhar et al., 2020)	lab	Indian (ISL)	4,292	7	263	–
NMFs-CSL (Hu et al., 2021b)	lab	Chinese (CSL)	32,010	10	1,067	addressed
BOBSL (Albanie et al., 2021)	TV	British (BSL)	**452,000**[1]	39	2,281	–
GSL isol. (Adaloglou et al., 2021)	lab	Greek (GSL)	40,785	7	310	grouped
LSFB-ISOL (Fink et al., 2021)	lab	Fra/Bel	47,600	100	395	–
CISLR (Joshi et al., 2022)	web	Indian (ISL)	7,000	71	4,765	–
LSA64 (Ronchetti et al., 2023)	lab	Argentinian	3,200	10	64	–
ASL Citizen (Desai et al., 2024)	crowd	American (ASL)	83,399	52	2,731	–
Slovo (Kapitanov et al., 2023)	crowd	Russian (RSL)	20,000	194	1,000	–
FDMSE-ISL (Patra et al., 2024)	lab	Indian (ISL)	40,033	20	2,000	–
MM-WLAuslan (Shen et al., 2024b)	lab	Australian(Auslan)	282,000[2]	76	3,215	–
Logos (Ours)	crowd	Russian (RSL)	200,000	**381**	2,863[3]	both

Table 1: Summary of existing ISLR datasets. *Method* – the collection method: laboratory, web scrapping, TV programs, crowdsourcing. *VSSigns* column shows if visually similar signs (VSSigns) were considered by the dataset authors: *grouped* – VSSigns groups have common labels; *addressed* – the authors adopt VSSigns presence in the dataset and propose some methods to tackle them at training time; "–" – VSSigns presence is not discussed.
[1] These datasets mostly have automatic annotations of isolated glosses.
[2] Actually, the dataset contain 70730 samples recorderd from four views each.
[3] This number of glosses is grouped into 2,004 VSSign labels.

signs. This paper calls such hardly distinguishable signs "*visually similar signs*" (*VSSigns*). Formally, we define them as signs that have different meanings but have the same manual component. The VSSign concept includes both polysemy and minimal pair cases. While they differ from a linguistic perspective, it makes sense to combine them for machine learning purposes.

There is no standard approach to annotating VSSigns: they can be annotated with either different or similar labels. In this paper, we call it *ungrouped gloss* and *grouped VSSign* annotations. The datasets collected for the most common words of spoken language (Sincan and Keles, 2020; Kapitanov et al., 2023), typical continuous phrases (Albanie et al., 2020, 2021; Adaloglou et al., 2021), or based on an SL dictionary (Patra et al., 2024) primarily have different (*ungrouped*) labels for similar signs. For instance, according to (Zuo et al., 2023), among 2,000 classes of widely used WLASL dataset (Li et al., 2020), 334 classes form groups of VSSigns. Additional efforts are needed to merge similar VSSigns and assign unique *grouped VSSign* labels to them.

Among the reviewed datasets, three papers state that VSSigns were grouped. Three papers confirm the presence of ungrouped VSSigns in the presented datasets and propose some techniques to distinguish them (Table 1). To improve VSSigns classification, Hu et al. (2021b) deform a feature map, stretching more informative areas to emphasize non-manual features. Zuo et al. (2023) propose label smoothing depending on their semantic difference and a common latent space for gloss embeddings and vision features to maximize the separability of confusing signs. Other works do not mention any steps to handle VSSigns in the proposed datasets. To our knowledge, no research has examined the impact of VSSigns on the quality of the resulting encoder for downstream tasks. Such a study is one of the topics of this work, using transfer learning to another language as a downstream task example.

2.4 Multi-Dataset Training

Although researchers complain about insufficient SLR training data (Gokul et al., 2022; Papadimitriou and Potamianos, 2023), the topic of cross-

language dataset sharing is poorly exploited. Gokul et al. (2022) implemented a multilingual SLR model for 11 sign languages by simply translating the labels of all the languages into English. The authors themselves admit that their model of combining different languages is primitive and does not make progress for some datasets. Tornay et al. (2020) train a unified hand movement model using 3 different sign language resources. Then, they optimize the classifier using the target sign language data. However, their cross-lingual model falls short of the monolingual reference. Yin et al. (2022) propose the MLSLT translation network as a single model for multilingual translation. Their work is limited to their rather small datasets and doesn't address leveraging large SL datasets to improve the model quality. SignCLIP (Jiang et al., 2024) utilizes a multilingual sign language dictionary, SpreadTheSign.com, for training. It contains mainly a single sample per concept per language, and the SignCLIP authors translate all concepts into English, similar to (Gokul et al., 2022), so knowledge transfer between different sign languages is not investigated. Hu et al. (2022) introduced an additional shared module that learns knowledge from two languages. It improved accuracy for Chinese CSL-Daily (Zhou et al., 2021) and German Phoenix-14 (Koller et al., 2015) datasets. Wei et al. (2023) also benefit from the joint using the same datasets by creating a gloss translation map based on the visual similarity of signs, rather than their meanings. Authors train the model for the German language using both datasets and replace gloss labels in Chinese videos with German labels using this map, treating Chinese signs as German. As shown below, this mapping method does not give optimal results (see Section 5.3). However, as far as we know, the ability of a model trained on an extensive SL dataset to serve as an encoder for SL tasks for other sign languages has not been explored enough. Such a study, along with the comparison of different approaches to it, is another subject of our work.

3 Logos Dataset

3.1 Dataset Characteristics

The Logos dataset contains 199,668 videos recorded by 381 signers (deaf individuals, professional interpreters, sign language teachers, and CODA/SODA[2]). The total duration of the dataset video is 221.4 hours, with 104.7 hours representing the demonstration of signs themselves and the rest being fragments before and after the sign demonstration. The dataset contains signs for 2,863 of the most commonly used lemmas in the Russian general vocabulary, combined into 2,004 grouped VSSign classes with 35 to 737 samples per class.

The Logos dataset includes the Slovo public dataset (Kapitanov et al., 2023) with the renewed annotations, amended with VSSign classes.

More details on the dataset's characteristics are provided in Appendix A.1.

3.2 Gloss Selection

The Logos vocabulary selection is based on the frequency list of the Russian language corpus[3]. We have (1) selected the top 3,000 lemmas, except for prepositions, conjunctions, particles, and interjections, (2) removed lemmas that present in the Slovo dataset, and (3) selected glosses as lemmas for which sample video present on the SpreadTheSign[4] sign language dictionary website. We added 1,863 new glosses, bringing the total in the Logos dataset to 2,863 glosses.

3.3 Data Collection

The Logos data collection pipeline includes signer selection, video collection, video validation, and sign time interval annotation stages that coincide with the ones of the Slovo(Kapitanov et al., 2023) pipeline. See Appendix A.2 for details.

3.4 VSSigns Grouping

We grouped visually similar signs based solely on their manual components through two stages.

First, we trained a baseline model on the dataset with ungrouped glosses and processed 2,863 sign template videos with the model. Using confidence of prediction classes for the template videos, we identified the 10 most similar templates for each one. Deaf experts compared each template video with its 10 counterparts and marked pairs that differ only in non-manual components. Sign pairs matched by the majority of 5 experts were annotated as VSSigns.

Next, we applied three rounds of additional verification. In each round, the model was trained on the currently grouped labels. Based on the classification results, we identified the most confusing class pairs and visually inspected misclassified

[2]Child of Deaf Adult, Sibling of Deaf Adult/Deaf person

[3]http://dict.ruslang.ru/freq.php

[4]https://spreadthesign.com/ru.ru/search/

samples. If VSSign candidates were found, we consulted deaf experts and grouped the labels additionally.

3.5 Train-test Split

We aim to maintain an 80/20 ratio for the train and test data split applied to both the number of signers and the number of samples for each sign. Given that the number of signs recorded by different signers differs, the dataset split confirming all these requirements hardly has a strict resolution. We applied a dynamic programming algorithm to find the best approximation. See Appendix A.3.

4 Experiments setup

4.1 Datasets

In addition to the proposed large-scale Logos dataset, we selected two widely used ISLR benchmarks as target datasets for transfer learning: the Turkish Sign Language (TİD) dataset AUTSL (Sincan and Keles, 2020) and the American Sign Language (ASL) dataset WLASL (Li et al., 2020). The WLASL dataset offers a large vocabulary but is relatively small in size, comprising an average of approximately 10 samples per class. The AUTSL dataset provides more samples but includes a more limited vocabulary and fewer signers. We also include MM-WLAuslan (Shen et al., 2024b), one of the most extensive publicly available ISLR datasets, as an additional pre-training baseline to compare against Logos. Key characteristics of these datasets are summarized in Table 1.

4.2 Sign Language Recognition Pipeline

Our experimental setup is based on (Kvanchiani et al., 2024). The authors explore various training aspects to propose the optimal ISLR pipeline. They use MViTv2-S (Li et al., 2022) as a backbone, a fully connected (FC) layer for classification, a cross-entropy classification loss with label smoothing, and sign timeline boundary regression as an auxiliary task. The backbone was initialized with Kinetics-400 pre-train. The pipeline processes $32 \times 224 \times 224$ frame chains, randomly sampled from the input video with a step of 2 frames. We implement an auxiliary boundary regression task as follows. The sign's ground truth boundary timestamps are rescaled relative to the sampled clip: the clip length is set as 1, the clip start is set as 0 for the sign start point, and the clip end is set as 0 for the sign endpoint. Alongside the classification heads, we add an extra FC layer with two output channels for the sign start and end regression. Its output and scaled ground truth values are mapped to $(-1, 1)$ using the formula $y = 2\sigma(x) - 1$, where $\sigma(x)$ is the sigmoid function, to diminish the influence of sign boundaries that are outside the clip. We use mean squared error loss to train this regression head. The total loss function is calculated as a weighted sum of the classification and regression losses: $L = L_{cls} + 2.5L_{regr}$. We evaluate the model using a top-1 instance-based accuracy metric: the ratio of the correctly classified samples to the total samples number.

4.3 Multi-dataset Co-training Method

Different national sign language datasets have their own label spaces with no common taxonomy. As a result, they cannot be directly combined for joint training.

In our pipeline (Figure 2), we mark each sample with its language tag. During training, we form batches containing a mix of sign languages. After processing the mixed batch by the common visual encoder, we apply the language-specific gate, which splits the batch into language-specific sub-batches using the language tag and processes each sub-batch by the language-specific classification head. Loss functions from each classification head were weighted proportionally to the number of appropriate language samples in the mixed batch.

At the training stage, we use CutMix (Yun et al., 2019) and Mixup (Zhang, 2017) inter-sample regularization strategies. They can not be applied to the mixed batch because labels of different languages cannot be mixed. We use the same language-specific gate to split the mixed batch into language-specific sub-batches before applying these augmentations and then merge the resulting samples back into one batch.

5 Experiment Results and Ablation Study

5.1 Transfer learning experiments

The presented extensive Logos dataset was used as a pre-train for transfer learning tasks. We used relatively small AUTSL and WLASL datasets as modeling examples of low-resource datasets for transfer learning. These datasets were also selected because benchmark results are available for comparison. Additionally, we created reduced versions of each dataset, limited to 10, 3, and 1 sample per class, for more challenging low-data experiments.

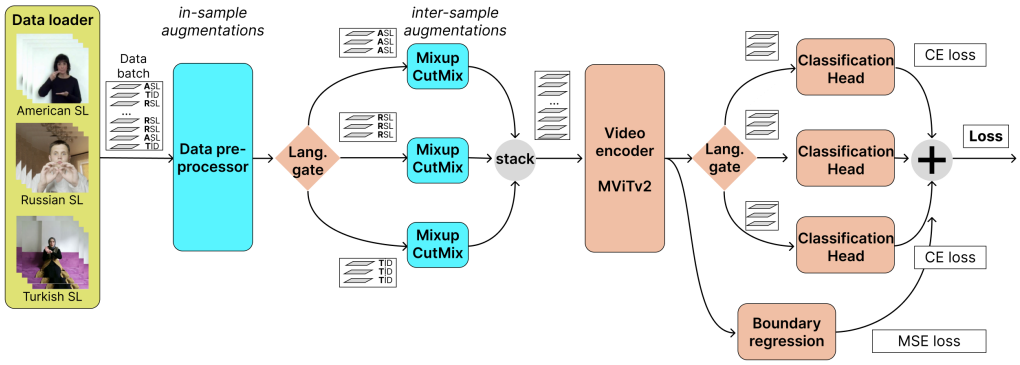

Figure 2: Multi-dataset co-training pipeline. Samples from different languages are processed as a united batch. Before the inter-sample augmentations and the language-specific classification heads, the language-specific gates split the batch into language-specific sub-batches.

Method	**Top-1 accuracy**		
	Logos	AUTSL	WLASL
Separate training (baseline)	97.90	96.58	60.88
Transfer learning from Logos dataset:			
Encoder is frozen	–	97.25	62.44
Encoder is being trained	–	97.73	65.57
Multi-dataset co-training with Logos dataset:			
Logos + AUTSL	97.92	**97.83**	–
Logos + WLASL	**97.93**	–	65.74
Logos + AUTSL + WLASL	97.92	97.81	**66.82**

Table 2: Baseline, transfer learning, and multi-dataset co-training with the Logos dataset. Transfer learning and Multi-dataset co-training experiments use the encoder, initialized from the Logos pre-train.

Method	**Top-1 accuracy**	
	AUTSL	WLASL
Full dataset (baseline)	97.25	62.44
10-shot (10 samples per class)	90.16	61.12
3-shot (3 samples per class)	83.99	54.10
one-shot (1 sample per class)	82.44	37.07

Table 3: Few-shot and one-shot transfer learning with frozen Logos pre-trained encoder.

First, we trained the separate baseline models on the Logos, AUTSL, and WLASL datasets using the same setup (Section 4). Then, we examined the applicability of the Logos pre-trained model for transfer learning to smaller AUTSL and WLASL datasets. With the model backbone initialized from the Logos pre-train, we evaluated two transfer learning strategies: (a) training all model weights and (b) freezing the pre-trained encoder and training only the classification head. The Logos pre-train substantially improves the model accuracy compared to training from scratch (Table 2).

Next, we explored the potential of a Logos pre-trained encoder for few-shot learning on other sign languages. We limited train sets of AUTSL and WLASL datasets to the randomly selected 10, 3, and 1 samples per class. Then, we applied transfer learning with a frozen encoder to these truncated datasets. The test part of the datasets was left intact. Although truncated datasets produce worse models, training even on 1 sample per class still keeps the models working, at least for the AUTSL dataset, which has a smaller vocabulary (Table 3).

These experiments demonstrate the possibility of transfer learning from the extensive Logos dataset to other sign languages with only a limited amount of training data.

5.2 Cross-lingual Multi-dataset Co-training

We investigated the described multi-dataset co-training method using the pairs Logos and AUTSL, Logos and WLASL, and all three datasets combined. The encoder and the Logos classifier were initialized from the Logos baseline model for all experiments.

A single model, produced by a multi-dataset co-training, far surpasses the accuracy of the models, separately trained on low-resource datasets from scratch, and also surpasses individual models trained using conventional transfer learning (Table 2). Moreover, results for the WLASL dataset are far above existing SOTA metrics[5], see Table 4. As for the AUTSL dataset, note that all leading models use ensembling, pose recognition, and depth maps (or some of the above). In contrast,

[5]according to https://paperswithcode.com/ and other papers referring to the datasets in question

Model	Top-1 accuracy	
	AUTSL	WLASL
BSL-1K (Albanie et al., 2020)	–	46.9
SignBERT (Hu et al., 2021a)	–	54.7
SAM-SLR (Jiang et al., 2021)	98.5	58.7
One Model is Not Enough[1]	96.4	–
ZBEST (Zhao et al., 2023)	–	54.6
SignBERT+ (Hu et al., 2023)	–	55.6
NLA-SLR (Zuo et al., 2023)	–	61.3
SL-GDN (Miah et al., 2023)	96.5	–
ST-GCN[2]	96.7	–
Audio-visual ... (Ryumin et al., 2023)	**98.6**	–
HWGAT (Patra et al., 2024)	95.8	48.5
StepNet (Shen et al., 2024a)	–	61.2
Uni-Sign (Li et al., 2025)	–	**63.5**
MViTv2 (our baseline)	96.58	60.88
Multi-dataset with MM-WLAuslan	97.43	64.59
Multi-dataset with Logos (ours)	**97.81**	**66.82**

Table 4: Our results compared with SOTA results for the AUTSL and WLASL datasets.
[1] (Hrúz et al., 2022)
[2] (Papadimitriou and Potamianos, 2023)

our model uses a single stream that takes only RGB input.

5.3 The Encoder Generalization Ability Check

We examined the hypothesis that an encoder pre-trained on the Logos dataset does not produce universal sign features but can only recognize the signs of the pre-train language. When applied to another language, the model maps these signs to the most similar target language signs, as in the approach of (Wei and Chen, 2023). To emulate this hypothesis, we processed the WLASL train set with the Logos pre-trained model and built the map by associating the assigned Logos labels with the most frequent WLASL ground truth labels. Then, we applied the same model to the WLASL test set and substituted the resulting Logos labels with WLASL labels using the map instead of training a target language classification head. We repeated the same experiment with the AUTSL dataset.

The results in Table 5 show that although this label mapping method works, it is significantly inferior to the trained classifier for the Logos pre-trained encoder. It confirms that the Logos pre-trained encoder produces universal sign embeddings that can encode new, unseen signs from another language.

Method	Top-1 accuracy	
	AUTSL	WLASL
Transfer learning	**97.25**	**62.44**
Map labels to target language	65.78	23.63

Table 5: Transfer learning with frozen encoder compared to label mapping from Logos to other language datasets.

Pre-train dataset	Top-1 accuracy			
	AUTSL	AUTSL, 3-shot	WLASL	WLASL, 3-shot
Logos	**97.25**	**83.99**	**62.44**	**54.10**
AUTSL	–	–	28.46	18.76
WLASL	93.16	67.90	–	–
MM-WLAuslan	93.64	69.79	45.21	33.01

Table 6: The importance of the pre-train dataset for cross-language transfer learning. Results for both whole and truncated versions of the AUTSL and WLASL datasets using pre-training on the Logos, WLASL, AUTSL, and MM-WLAuslan datasets.

5.4 The Importance of the Pre-train Dataset

Table 6 demonstrates that extensive dataset size is critical for training a powerful encoder for cross-language transfer learning. We repeated transfer learning experiments using pre-train on smaller AUTSL and WLASL datasets. One can see that the resulting accuracy degrades substantially compared to Logos pre-train.

We also compared the impact of the Logos dataset and the large-scale MM-WLAuslan dataset on multi-dataset co-training and transfer learning (Tables 4,6). The results obtained with Logos are substantially better than those with MM-WLAuslan for both tasks.

5.5 The Effect of VSSigns Grouping

We investigated the contribution of our approach with grouping labels of visually similar signs in obtaining a high-quality encoder. We trained the classifier on the Logos dataset, using unique pairs of ungrouped and grouped labels as classes. It formed 2,863 ungrouped gloss classes instead of 2,004 grouped VSSign classes in the baseline Logos annotation. Each ungrouped label has a unique associated grouped label, so the model, trained on the ungrouped labels, can be evaluated on grouped labels.

Table 7 shows the accuracy for models trained on 2,004 VSSign and 2,863 ungrouped classes. Quite predictably, the last model yields lower accuracy,

VSSigns		Top-1 accuracy		
train	**test**	Whole	non-VSSigns	VSSigns
Yes	Yes	**97.90**	**97.49**	**98.33**
No	Yes	97.44	97.10	97.79
No	No	87.02	97.10	76.51

Table 7: Comparison of training using grouped VSSigns annotation (baseline) and annotation without grouping.

Logos pre-train on VSSigns	Top-1 accuracy			
	AUTSL	AUTSL, 3-shot	WLASL	WLASL, 3-shot
Yes	**97.25**	**83.99**	**62.44**	**54.10**
No	96.79	82.38	60.74	51.60

Table 8: The effect of VSSigns grouping on transfer learning. Results for WLASL and AUTSL (whole and truncated to 3 samples per class) trained from Logos pre-train on grouped VSSigns annotation (baseline) and annotation without grouping.

primarily due to confusion of VSSigns. However, it is essential that it achieves lower accuracy on grouped VSSigns classes (the classes on which the 1st model was trained). Notably, the degradation is observed even for signs that are not VSSigns whose labels do not differ in all cases.. Furthermore, Table 8 shows that VSSigns grouping results in more effective transfer learning to other sign languages.

6 Conclusion

The paper examines two aspects of the isolated sign language recognition (ISLR) task: cross-language SL model training, including transfer learning, and approaches to handling visually similar signs (VSSigns). To explore these issues, this work presents Logos, a new publicly available Russian Sign Language dataset, the most extensive ISLR dataset by the number of signers and one of the largest available datasets while also the largest RSL dataset in size and vocabulary. It is shown that a model, pre-trained on the Logos dataset can be used as a universal encoder for other language SLR tasks, including few-shot learning. The cross-language transfer learning methods are evaluated, and it is demonstrated that the method of multi-dataset co-training with multiple language-specific classification heads improves SL models for low-resource datasets the most, compared to the conventional "pre-train and finetune" method. The key feature of the Logos dataset is the explicit annotation of visually similar sign groups. With its use, we show that explicitly grouping VSSign labels benefits trained model quality as a video encoder for downstream tasks, such as transfer learning to other sign languages. Based on the proposed contributions, we outperform current state-of-the-art results for the WLASL dataset and get competitive results for the AUTSL dataset, with a single stream model processing solely RGB video.

Limitations

This work is limited by using the MViT baseline architecture and focusing on cross-language transfer learning in ISLR as the downstream task. To generalize the conclusions, further research is needed involving diverse model architectures, low-resource target datasets, and a broader range of downstream tasks, including continuous sign language recognition. We consider this a promising direction for future work.

The Logos dataset reflects the demographics of the participants involved in its collection, resulting in an unbalanced distribution in terms of age and gender. Additionally, the dataset focuses exclusively on RSL, which limits its direct applicability to more diverse settings. Nonetheless, our findings suggest it can still be effectively leveraged in broader sign language recognition tasks.

Ethical Statement

Legal and ethical aspects were reviewed and approved by our institution's legal team. All crowdworkers provided informed consent, authorizing the processing and publication of the collected data. Informed consent was provided via a textual form on the crowdsourcing platform. Since all participants were fluent in written Russian, no interpreter translation was required. To save contributors' privacy, we use anonymized user hash IDs. We do not restrict the participation of signers under 18, provided parental consent was obtained during registration, in compliance with the Civil Code of the Russian Federation. Participation was voluntary. Compensation for completed tasks was aligned with the average salary of a sign language interpreter, proportionate to the time invested. We have verified that the Slovo dataset, incorporated into Logos, adheres to these ethical standards. The dataset is made available exclusively for research purposes. Nonetheless, we acknowledge the potential misuse, such as identifying individuals or enabling large-scale surveillance.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24:1750–1762.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and 1 others. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. 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.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and 1 others. 2024. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36.

Sarah Ebling, Necati Cihan Camgöz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, and 1 others. 2018. Smile swiss german sign language dataset. In *Proceedings of the 11th international conference on language resources and evaluation (LREC) 2018*. The European Language Resources Association (ELRA).

Sarah Ebling, Reiner Konrad, Penny Boyes Braem, and Gabriele Langer. 2015. Factors to consider when making lexical comparisons of sign languages: Notes from an ongoing comparison of german sign language and swiss german sign language. *Sign Language Studies*, 16(1):30–56.

Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 2021. Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

NC Gokul, Ladi Manideep, Negi Sumit, Selvaraj Prem, Kumar Pratyush, and Khapra Mitesh. 2022. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. *Advances in Neural Information Processing Systems*, 35:36202–36215.

Marek Hrúz, Ivan Gruber, Jakub Kanis, Matyáš Boháček, Miroslav Hlaváč, and Zdeněk Krňoul. 2022. One model is not enough: Ensembles for isolated sign language recognition. *Sensors*, 22(13):5043.

Hezhen Hu, Junfu Pu, Wengang Zhou, and Houqiang Li. 2022. Collaborative multilingual continuous sign language recognition: A unified framework. *IEEE Transactions on Multimedia*, 25:7559–7570.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.

Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021a. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11087–11096.

Hezhen Hu, Wengang Zhou, Junfu Pu, and Houqiang Li. 2021b. Global-local enhancement network for nmf-aware sign language recognition. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(3):1–19.

Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2018. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2822–2832.

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. K-rsl: a corpus for linguistic understanding, visual evaluation, and recognition of sign languages. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3413–3423.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Signclip: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193.

Abhinav Joshi, Ashwani Bhat, S Pradeep, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. Cislr: Corpus for indian sign language recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10357–10366.

Hamid Reza Vaezi Joze and Oscar Koller. 2018. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*.

Alexander Kapitanov, Kvanchiani Karina, Alexander Nagaev, and Petrova Elizaveta. 2023. Slovo: Russian sign language dataset. In *International Conference on Computer Vision Systems*, pages 63–73. Springer.

Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. 2019. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2306–2320.

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

Maksim Kuprashevich and Irina Tolstykh. 2023. Mivolo: Multi-input transformer for age and gender estimation. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 212–226. Springer.

Karina Kvanchiani, Roman Kraynov, Elizaveta Petrova, Petr Surovcev, Aleksandr Nagaev, and Alexander Kapitanov. 2024. Training strategies for isolated sign language recognition. *arXiv preprint arXiv:2412.11553*.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. 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.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4804–4814.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. *arXiv preprint arXiv:2501.15187*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, and 1 others. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Abu Saleh Musa Miah, Md Al Mehedi Hasan, Si-Woong Jang, Hyoun-Sup Lee, and Jungpil Shin. 2023. Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition. *Electronics*, 12(13):2841.

Liliane Momeni, Gul Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2020. Watch, read and lookup: learning to spot signs from multiple supervisors. In *Proceedings of the Asian Conference on Computer Vision*.

Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishibay, Vadim Kimmelman, and Anara Sandygulova. 2020. Evaluation of manual and non-manual components for sign language recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Oğulcan Özdemir, Ahmet Alp Kındıroğlu, Necati Cihan Camgöz, and Lale Akarun. 2020. Bosphorussign22k sign language recognition dataset. *arXiv preprint arXiv:2004.01283*.

Katerina Papadimitriou and Gerasimos Potamianos. 2023. Sign language recognition via deformable 3d convolutions and modulated graph convolutional networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Suvajit Patra, Arkadip Maitra, Megha Tiwari, K Kumaran, Swathy Prabhu, Swami Punyeshwarananda, and Soumitra Samanta. 2024. Hierarchical windowed graph attention network and a large scale dataset for isolated indian sign language recognition. *arXiv preprint arXiv:2407.14224*.

Franco Ronchetti, Facundo Manuel Quiroga, César Estrebou, Laura Lanzarini, and Alejandro Rosete. 2023. Lsa64: an argentinian sign language dataset. *arXiv preprint arXiv:2310.17429*.

Dmitry Ryumin, Denis Ivanko, and Elena Ryumina. 2023. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors*, 23(4):2284.

Xiaolong Shen, Zhedong Zheng, and Yi Yang. 2024a. Stepnet: Spatial-temporal part-aware network for isolated sign language recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(7):1–19.

Xin Shen, Heming Du, Hongwei Sheng, Shuyun Wang, Hui Chen, Huiqiang Chen, Zhuojie Wu, Xiaobiao Du, Jiaying Ying, Ruihan Lu, Qingzheng Xu, and Xin Yu. 2024b. MM-WLAuslan: Multi-View Multi-Modal Word-Level Australian Sign Language Recognition Dataset. *arXiv preprint arXiv:2410.19488*.

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

Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1366–1375.

Sandrine Tornay, Marzieh Razavi, and Mathew Magimai Doss. 2020. Towards multilingual sign language recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6309–6313. IEEE.

Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. 2016. Isolated sign language recognition with grassmann covariance matrices. *ACM Transactions on Accessible Computing (TACCESS)*, 8(4):1–21.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5109–5119.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Hongyi Zhang. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. Best: Bert pre-training for sign language recognition with coupling tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3597–3605.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14890–14900.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2024. Towards online sign language recognition and translation. *arXiv preprint arXiv:2401.05336*.

A Logos Dataset

This appendix contains more detailed characteristics of the Logos dataset and additional technical details for some production steps.

A.1 Dataset Characteristics

Videos have a resolution of at least 720 pixels on the minimum side at a 30 FPS frame rate. About 62% of videos are in FullHD format. Distributions of some dataset characteristics are represented in Figure 3. Among the 381 dataset signers, 41% are 30-40 years old, and 88% are female. We do not limit crowdsourcers by age and gender, and such an uneven distribution reflects the demographics of signers who wish to participate in the project. The participants include deaf individuals, professional interpreters, sign language teachers, and CODA/SODA. Since the data collection was conducted via a crowdsourcing platform, we do not have distribution statistics for these groups. However, all participants passed an RSL proficiency test before contributing. The dataset is divided into 80.7% in the train and 19.3% in the test sets.

A.2 Data Collection

Logos data collection pipeline follows the Slovo(Kapitanov et al., 2023) pipeline in signer selection, video collection, video validation, and sign time interval annotation stages. The dataset was collected on the crowdsourcing platforms ABC Elementary and Yandex Toloka.

Signers Selection All project signers confirmed their Russian Sign Language (RSL) proficiency by passing an exam on the crowdsourcing platform ABC Elementary[6] or Yandex Toloka[7]. The exam involved watching a video demonstrating an RSL sign and selecting the correct translation into Russian. Successful completion required correctly answering at least 17 out of 20 questions.

Video Collection. Signers watched a video of a correctly performed sign (video template) taken from SpreadTheSign[8] and then recorded a video replicating that sign. Before starting the tasks, signers reviewed the rules for video recording: 1) the gesture must match the example sign; 2) hands must remain within the frame; 3) only one person may appear in the video; 4) video does not shake; 5) video without processing; 6) video must have a short side of at least 720 pixels. Signers could record videos using a smartphone or webcam or upload pre-recorded videos from memory, watch the video, and overwrite the sign.

[6] https://elementary.activebc.ru/
[7] https://platform.toloka.ai/
[8] https://spreadthesign.com/ru.ru/search/

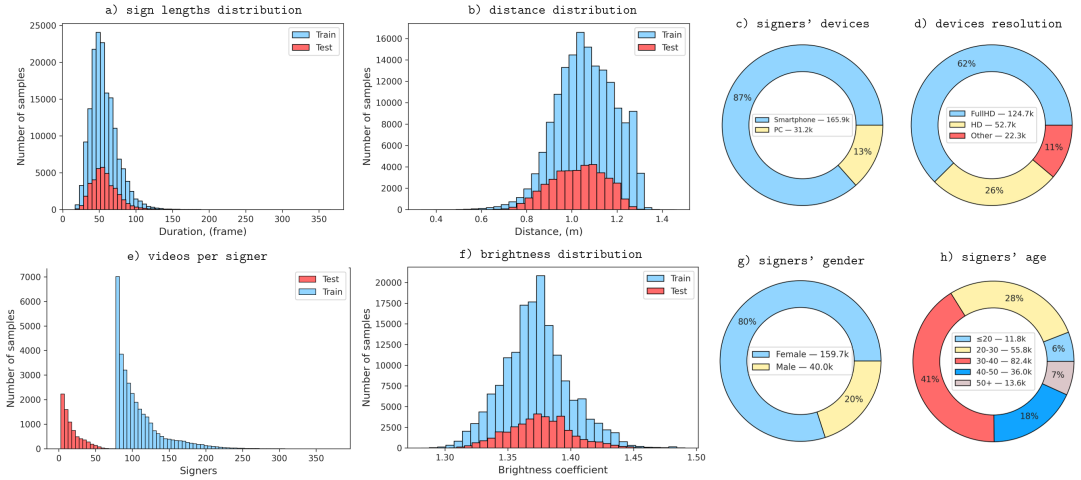

Figure 3: Dataset characteristics and distribution analysis. a) Sign length distribution. b) Distance distribution. The distance (in meters) is approximately estimated based on the length between the left and right shoulders of the signer obtained using MediaPipe (Lugaresi et al., 2019). c) Signers' devices. d) Devices resolution. e) Number of videos per signer. f) Brightness distribution. The sample brightness is the mean pixel brightness of grayscaled video frames. g) Signers' gender; h) Signers' age. The age is determined by the MiVOLO model (Kuprashevich and Tolstykh, 2023).

Video Validation. At least three validators (SL experts: teachers, professional interpreters) who have successfully completed the validation training and exam will review the video to ensure compliance with the video recording rules. Honeypots (predefined videos with known answers to the customer) were deliberately added to the validation project to identify and eliminate unscrupulous validators.

Time Interval Annotation. Time interval annotations were necessary to identify the start and the end of the sign dynamics and to exclude unrelated signers' movements. Each video was pre-processed to a frame rate of 30 fps, enabling the linkage of time intervals to frame numbers. All annotators have completed training and exams on time interval annotation. Honeypots were incorporated into the main tasks to identify and exclude dishonest annotators. Three annotators annotated each video, and the results were aggregated by open-source AggMe framework[9].

A.3 Train-test Split

Given that the number of signs recorded by different signers differs, we applied a dynamic programming algorithm to find the approximate solution that has non-intersecting signers in train and test subsets and simultaneously provides approximately 20/80 test/train split ratio both for signers and for each sign samples.

The main idea is that we always select strictly 20% signers and form the test set by the samples they demonstrated. We start with randomly selected 20% signers. Then, in a cycle, we find a sign with a test/train balance the most different from the target 20/80 ratio and try to swap a signer from the test set and a signer not in a test set to reduce the worst disbalance, starting from the test set signers who recorded the largest number of the scoped sign samples. So, we iteratively minimize the maximum deviation from the target test/train sign samples split ratio, keeping the test/train split ratio for signers always equal to the required value. See Algorithm 1 for more details.

B Sign Language Recognition Pipeline

This appendix provides additional details of the training pipeline used in the work.

B.1 Data Pre-processing

We randomly sample a 32-frame chain from a video with a step of 2 frames. If a sign on a video is longer than 63 frames, the sample is randomly selected within the sign duration. If extra frames are present on the video before and/or after the sign, we extend the range for sample selection up to 5 frames before and after the sign boundaries. For signs shorter than 63 frames, the sample is selected from the sign start

[9] https://github.com/ai-forever/aggme

Algorithm 1 Balanced train-test split

#Notation:
$\{G\}$: all gloss labels
$\{S\}$: all signers
$\{T\} \subset \{S\}$: test set signers
N_g : samples number of gloss g
$N_{g,s}$: samples number of g from signer s
$R_{g,s} = N_{g,s}/N_g$
#Initialization:
$p \leftarrow 0.2$ ▷ target test ratio
$\{T\} \leftarrow$ random p from $\{S\}$ ▷ test set signers
#Optimization:
repeat
  $\forall g : D_g \leftarrow \sum_{s \in T} R_{g,s}$ ▷ test gloss ratios
  $d_{wst} \leftarrow \max_g |D_g - p|$ ▷ worst deviation
  $g_{wst} \leftarrow \text{argmax}_g |D_g - p|$ ▷ worst gloss
  #Build sorted list of
  #test signer candidates for replacement:
  if $D_{g_{wst}} > p$ **then**
    $U \leftarrow \text{sorted}_{\text{desc}}\, T$ by $R_{g_{wst},s \in T}$
  else
    $U \leftarrow \text{sorted}_{\text{asc}}\, T$ by $R_{g_{wst},s \in T}$
  end if
  for s' in U **do**
    #Try to replace $s\prime$
    #with signer not in test set:
    for s'' in $S \setminus T$ **do**
      $\forall g : D'_g \leftarrow D_g - R_{g,s'} + R_{g,s''}$
      $d'_{wst} \leftarrow \max_g \left|D'_g - p\right|$
      if $d'_{wst} < d_{wst}$ **then**
        #replace the s' signer in $\{T\}$
        $\{T\} \leftarrow \{T\} \cup \{s''\} \setminus \{s'\}$
        break to outer Repeat
      end if
    end for
  end for
until converge
{test video samples} $\leftarrow$ {video: signer $\in T$}
Output: {test video samples}

and padded at the end by the last frame.

At the training stage *Speed Up&Slow Down* and *Random Add&Random Drop* frame sampling augmentations from Kvanchiani et al. pipeline (2024) are applied. With probability $p = 0.25$, video is accelerated twice or slowed twice with the same p. With $p = 0.5$ we randomly drop 10% of frames. With $p = 0.25$ we truncate a sampled frame chain by 30% and stretch it to the original size by random repeat of remaining frames.

Sampled frames are resized to 300 pix over the longest side and randomly cropped to 224×224 with square padding if needed. Frames are augmented with ColorJitter, RandomNoise, Sharpness, Flip, RandomErasing, and ImageCompression image augmentations. Augmentation parameters are set the same for every frame in a video sample.

Also we use CutMix (Yun et al., 2019) and Mixup (Zhang, 2017) inter-sample regularization strategies.

B.2 Training Schedule

Training on the Logos dataset was performed on 4 Tesla H100s with 80GB RAM using batch size 16 per GPU for 50 epochs, which took about 40 hours.

For the first 5 epochs, the learning rate linearly increases from 8e-6 to 4.8e-3. Then, a cosine scheduler is used for epochs 6 to 40, reducing LR to 8e-5. Then, LR remains constant.

We use the AdamW optimizer with weight decay=0.05.

When datasets other than Logos are used, or in case of multi-dataset training, to maintain comparable training conditions, including training time, we scale training epochs number and LR schedule to keep the same number of iterations. For instance, for training on a 50% subset of Logos, we train for 100 epochs, with cosine LR annealing from epoch 11 to epoch 80.

For tasks of transfer learning with a frozen encoder, we use a faster training schedule with reduced maximum LR: 15 total epochs with linear LR warm-up from 8e-6 to 8e-4 for 5 times shorter period, 3.5 times shorter cosine annealing from 8e-4 to 8e-5, total training iteration number – 30% of initial training duration. The specified number of epochs is calculated for the target dataset as described above.