

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

Ilya Ovodov , Petr Surovtsev , Karina Kvanchiani , Alexander Kapitanov and Alexander Nagaev

SaluteDevices

{iovodov, petr.surovcev, karinakvanchiani, kapitanovalexander, sashanagaev1111}@gmail.com

Abstract

This paper examines two aspects of the isolated sign language recognition (ISLR) task. First, despite the availability of a number of datasets, the amount of data for most 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 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. 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. Primarily, information in these languages is expressed by hand shape and their motion (manual components), but also with a great aid of motion of mouth, head, eyes, and the body (non-manual components).

The problem of computer sign language recognition and translation has a practical application with significant social



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

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.

As a rule, a national sign language is associated with a national spoken language. Each sign corresponds to a spoken word named “gloss”, which describes the sign’s meaning. However, sign languages are independent languages with their own vocabulary and rules [Sandler, 2006].

The presented work deals with the isolated sign 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].

The serious 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 cross-lingual use of data. Although some papers utilize cross-lingual training, they are limited to joint use of two or more rather small datasets. To our knowledge, no research investigates the ability of a model trained on an extensive SL dataset to serve as an encoder for SL tasks for other sign languages and compares different approaches to it. While the applicability of a larger dataset of the same language is straightforward, cross-language transfer learning needs an investigation. 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 hand-shapes and motions can have various semantic meanings. Such signs can be either strictly indistinguishable or only distinguishable by their constituent non-manual features [Zuo *et al.*, 2023; Hu *et al.*, 2021b], see Figure 1. The difference between the individual signers’ manner blurs the boundary between non-manual features and makes such signs practically indistinguishable out of context. This paper calls such hardly distinguishable signs “visually similar signs” (VSSigns).

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 trans-

fer 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 available¹.

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; Hu *et al.*, 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].

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

¹<https://github.com/ai-forever/logos>

Dataset	Method	Language	Samples	Signers	Glosses	VSSigns
DEVISIGN-L [Wang <i>et al.</i> , 2016]	lab	Chinese (CSL)	24,000	8	2,000	–
SLR500 [Huang <i>et al.</i> , 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 <i>et al.</i> , 2018]	lab	Swiss German (DSGS)	9,000	30	100	–
BosphorusSign22k [Özdemir <i>et al.</i> , 2020]	lab	Turkish (TSL)	22,542	6	744	grouped
AUTSL [Sincan and Keles, 2020]	lab	Turkish (TSL)	38,336	43	226	–
WLASL [Li <i>et al.</i> , 2020]	web	American (ASL)	21,083	119	2,000	–
BSLDict [Momeni <i>et al.</i> , 2020]	lab	British (BSL)	14,210	28	9,283	addressed
BSL-1K [Albanie <i>et al.</i> , 2020]	TV	British (BSL)	273,000*	40	1,064	–
INCLUDE [Sridhar <i>et al.</i> , 2020]	lab	Indian (ISL)	4,292	7	263	–
NMFs-CSL [Hu <i>et al.</i> , 2021b]	lab	Chinese (CSL)	32,010	10	1,067	addressed
BOBSL [Albanie <i>et al.</i> , 2021]	TV	British (BSL)	452,000*	39	2,281	–
GSL isol. [Adaloglou <i>et al.</i> , 2021]	lab	Greek (GSL)	40,785	7	310	grouped
LSFB-ISOL [Fink <i>et al.</i> , 2021]	lab	Fra/Bel	47,600	100	395	–
CISLR [Joshi <i>et al.</i> , 2022]	web	Indian (ISL)	7,000	71	4,765	–
LSA64 [Ronchetti <i>et al.</i> , 2023]	lab	Argentinian	3,200	10	64	–
ASL Citizen [Desai <i>et al.</i> , 2024]	crowd	American (ASL)	83,399	52	2,731	–
Slovo [Kapitanov <i>et al.</i> , 2023]	crowd	Russian (RSL)	20,000	194	1,000	–
FDMSE-ISL [Patra <i>et al.</i> , 2024]	lab	Indian (ISL)	40,033	20	2,000	–
MM-WLAuslan [Shen <i>et al.</i> , 2024b]	lab	Australian (Auslan)	282,000	76	3,215	–
Logos (Ours)	crowd	Russian (RSL)	200,000	381	2,863/2,004**	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 were assigned a unique label; *addressed* – the authors adopt VSSigns presence in the dataset and propose some methods to tackle them at training time; the dash – VSSigns presence is not discussed. * — these datasets mostly have automatic annotations of isolated glosses. ** — numbers of ungrouped gloss labels and grouped VSSign labels are provided.

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

There is no standard approach to annotating visually similar signs (VSSigns). As a result, similar signs for different glosses 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; Albanie *et al.*, 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. Two 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. Tor-

Yin 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. 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, no research has been found that investigates the ability of a model trained on an extensive SL dataset to serve as an encoder for SL tasks for other sign languages and compares different approaches to it. Such a study is another subject of our work.

3 Logos Dataset

3.1 Dataset Characteristics

The Logos dataset contains 199,668 videos, divided into 80.7% in the train and 19.3% in the test sets. 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. 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 2,863 unique gloss classes combined into 2,004 grouped VSSign classes with 35 to 737 samples per class. The dataset was recorded by 381 signers of various age categories, 41% of them are 30-40 years old, and 88% are female (Figure 2g,h). 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. All signers passed an exam confirming their Russian Sign Language proficiency.

The Logos dataset includes the Slovo public dataset [Kapitanov *et al.*, 2023] with the renewed annotations. The pipeline for collecting new data repeats the one used for the Slovo dataset, except for the extended gloss selection stage, the new VSSigns grouping stage, and the train-test split stage.

3.2 Gloss Selection

The Logos vocabulary selection is based on the frequency list of the Russian language corpus². 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 SpreadTheSign³ sign language

dictionary website. We added 1,863 new glosses, bringing the total in the Logos dataset to 2,863 glosses.

3.3 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 templates videos, we identified the 10 most similar templates for each one. Deaf experts compared each of the identified videos to their templates and labeled matching ones 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 samples. If VSSign candidates were found, we consulted deaf experts and grouped the labels additionally.

3.4 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.

4 Experiments setup

4.1 Datasets

In addition to the extensive Logos dataset, we selected two widely used ISLR datasets as examples of low-resource datasets: the Turkish Sign Language (TSL) dataset AUTSL [Sincan and Keles, 2020] and the American Sign Language (ASL) dataset WLASL [Li *et al.*, 2020]. Their key characteristics are shown in Table 1. The WLASL dataset has a large number of glosses but is relatively small with 21,083 samples, averaging about 10 samples per class. The AUTSL dataset contains more samples but has a limited vocabulary and number of signers.

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

²<http://dict.ruslang.ru/freq.php>

³<https://spreadthesign.com/ru.ru/search/>

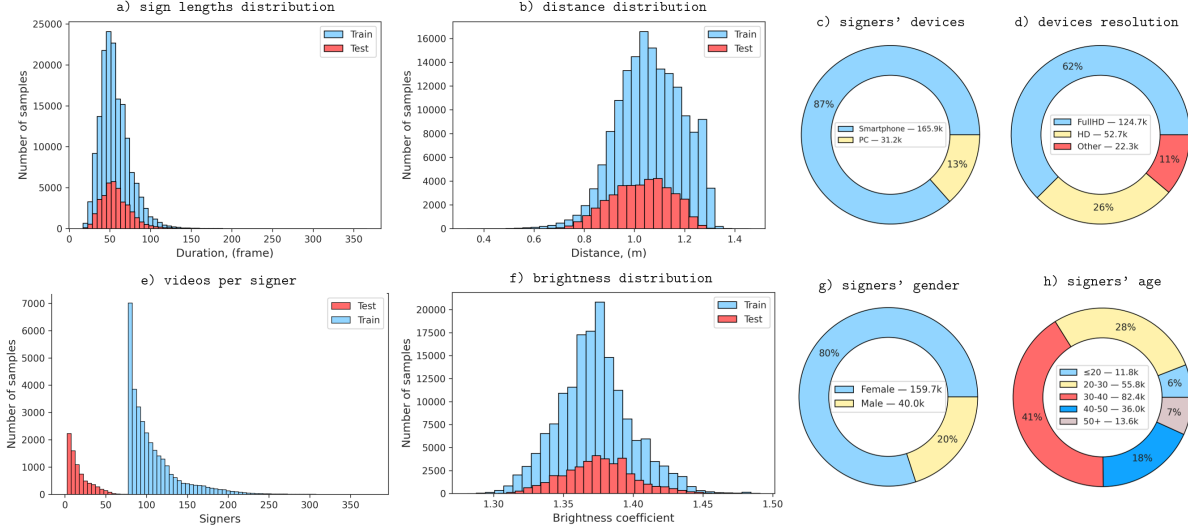


Figure 2: Dataset characteristics and distribution analysis. a) Sign length distribution. b) Distance distribution. The distance (in meters) is approximately estimated based on the length between the left and right shoulders of the signer obtained using MediaPipe [Lugaresi *et al.*, 2019]. c) Signers’ 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].

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: $L = L_{cls} + 2.5L_{reg}$. 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 languages contain different signs for different words. Each dataset has its own label space with no common taxonomy. Therefore, they cannot be directly mixed for simultaneous training.

In our pipeline (Figure 3), 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.

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

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. The AUTSL and WLASL datasets were taken as examples of low-resource datasets. 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 en-

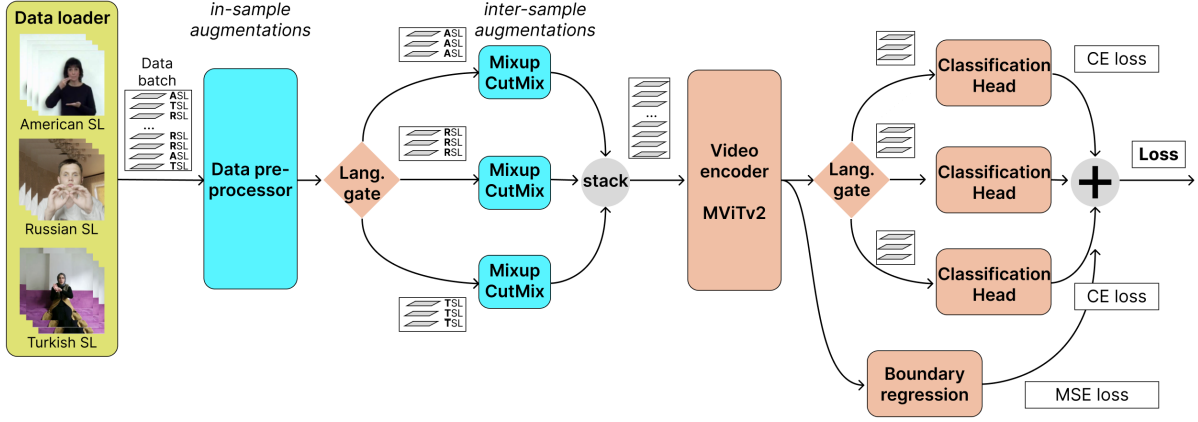


Figure 3: Multi-dataset co-training pipeline. Samples from different languages are processed as a united batch. Before the inter-sample augmentations and the language-specific classification heads, the language-specific gates split the batch into language-specific sub-batches.

Method	Top-1 accuracy	
	AUTSL	WLASL
Full dataset (baseline)	95.25	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.

coder 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. Below, we show that a large Logos dataset size is critical for the encoder quality (Section 5.4) and that VSSigns label grouping improves it (Section 5.5).

5.2 Cross-lingual Multi-dataset Co-training Results

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 on low-resource datasets of the models, separately trained on each dataset from scratch, and also surpasses individual models trained using conventional transfer learning (Table 2). Moreover, results for the

Model	Top-1 accuracy	
	AUTSL	WLASL
BSL-1K [Albanie <i>et al.</i> , 2020]	–	46.9
SignBERT [Hu <i>et al.</i> , 2021a]	–	54.7
SAM-SLR [Jiang <i>et al.</i> , 2021]	98.5	58.7
One Model is Not Enough [Hrúz <i>et al.</i> , 2022]	96.4	–
BEST [Zhao <i>et al.</i> , 2023]	–	54.6
SignBERT+ [Hu <i>et al.</i> , 2023]	–	55.6
NLA-SLR [Zuo <i>et al.</i> , 2023]	–	61.3
SL-GDN [Miah <i>et al.</i> , 2023]	96.5	–
ST-GCN [Papadimitriou and Potamianos, 2023]	96.7	–
Audio-visual ... [Ryumin <i>et al.</i> , 2023]	98.6	–
HWGAT [Patra <i>et al.</i> , 2024]	95.8	48.5
StepNet [Shen <i>et al.</i> , 2024a]	–	61.2
MViTv2 2024 (our baseline)	96.58	60.88
Multi-dataset 2025 (ours)	97.81	66.82

Table 4: Our results compared with SOTA results for the AUTSL and WLASL datasets.

WLASL dataset are far above existing SOTA metrics⁴, see Table 4. As for the AUTSL dataset, note that not only models with accuracy better than ours [Ryumin *et al.*, 2023; Jiang *et al.*, 2021], but other leading models use ensembling, pose recognition, depth map (or some of the above). In contrast, our model uses a single stream that takes only RGB input. It confirms that the co-training method is promising for low-resource sign languages.

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 from the train set. 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

⁴according to <https://paperswithcode.com/> and other papers referring to the datasets in question

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	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	67.9	—	—

Table 6: The importance of the pre-train dataset size for cross-language transfer learning. Results for both whole and truncated versions of the AUTSL and WLASL datasets using pre-training on the Logos dataset and more low-resource WLASL and AUTSL ones.

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.

5.4 The Importance of the Dataset Size

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.

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 2,880 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 that such a model has worse quality than the baseline model, trained on grouped VSSign labels. Notably, the degradation is observed even for signs that are not VSSigns and have unique grouped labels. Furthermore, Table 8 shows that VSSigns grouping results in more effective transfer learning to other sign languages.

Train	Test	Top-1 accuracy		
		Whole	non-VSSigns	VSSigns
grouped	grouped	97.90	97.49	98.33
not grouped	grouped	97.44	97.10	97.79
not grouped	not grouped	87.02	97.10	76.51

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

Logos pre-train	Top-1 accuracy			
	AUTSL	AUTSL, 3-shot	WLASL	WLASL, 3-shot
Grouped VSSigns	97.25	83.99	62.44	54.10
No grouping	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.

6 Limitations

This work is limited by MvIT baseline architecture and ISLR cross-language transfer learning as a downstream task. Additional research on other large-scale pre-train datasets, low-resource target datasets, and other downstream tasks, including continuous SL, is needed to generalize the conclusions. We consider it as a perspective.

The Logos dataset has a biased age and gender distribution due to the demographics of the project participants and race distribution due to the natural bias of RSL native users. It can limit its applicability.

7 Conclusions

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.

Ethical Statement

All crowdworkers provided informed consent, authorizing the processing and publication of the collected data. 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 the registration, in compliance with the Civil Code of the Russian Federation⁵. 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

- [Adaloglou et al., 2021] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakis, Dimitris Papazachariou, and Petros Daras. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, 24:1750–1762, 2021.
- [Albanie et al., 2020] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer, 2020.
- [Albanie et al., 2021] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.
- [Chen et al., 2022] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5120–5130, 2022.
- [Desai et al., 2024] Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. Asl citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36, 2024.
- [Ebling et al., 2018] Sarah Ebling, Necati Cihan Camgöz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, et al. 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), 2018.
- [Fink et al., 2021] Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 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, 2021.
- [Gokul et al., 2022] NC Gokul, Ladi Manideep, Negi Sumit, Selvaraj Prem, Kumar Pratyush, and Khapra Mitesh. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. *Advances in Neural Information Processing Systems*, 35:36202–36215, 2022.
- [Hrůz et al., 2022] Marek Hřůz, Ivan Gruber, Jakub Kanis, Matyáš Boháček, Miroslav Hlaváč, and Zdeněk Krňoul. One model is not enough: Ensembles for isolated sign language recognition. *Sensors*, 22(13):5043, 2022.
- [Hu et al., 2021a] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 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, 2021.
- [Hu et al., 2021b] Hezhen Hu, Wengang Zhou, Junfu Pu, and Houqiang Li. Global-local enhancement network for nmf-aware sign language recognition. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(3):1–19, 2021.
- [Hu et al., 2022] Hezhen Hu, Junfu Pu, Wengang Zhou, and Houqiang Li. Collaborative multilingual continuous sign language recognition: A unified framework. *IEEE Transactions on Multimedia*, 25:7559–7570, 2022.
- [Hu et al., 2023] Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239, 2023.
- [Huang et al., 2018] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2822–2832, 2018.
- [Jiang et al., 2021] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3413–3423, 2021.
- [Joshi et al., 2022] Abhinav Joshi, Ashwani Bhat, S Pradeep, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. Cisl: Corpus for indian sign language recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10357–10366, 2022.
- [Joze and Koller, 2018] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.
- [Kapitanov et al., 2023] Alexander Kapitanov, Kvanchiani Karina, Alexander Nagaev, and Petrova Elizaveta. Slovo: Russian sign language dataset. In *International Conference on Computer Vision Systems*, pages 63–73. Springer, 2023.
- [Koller et al., 2015] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.
- [Koller et al., 2019] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, and Richard Bowden. 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, 2019.
- [Kuprashevich and Tolstykh, 2023] Maksim Kuprashevich and Irina Tolstykh. Mivolo: Multi-input transformer for age and

⁵<https://ihl-databases.icrc.org/en/national-practice/federal-law-no-152-fz-personal-data-2006>

- gender estimation. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 212–226. Springer, 2023.
- [Kvanchiani *et al.*, 2024] Karina Kvanchiani, Roman Kraynov, Elizaveta Petrova, Petr Surovcev, Aleksandr Nagaev, and Alexander Kapitanov. Training strategies for isolated sign language recognition. *arXiv preprint arXiv:2412.11553*, 2024.
- [Li *et al.*, 2020] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020.
- [Li *et al.*, 2022] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Kartikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 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, 2022.
- [Lugaresi *et al.*, 2019] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.
- [Miah *et al.*, 2023] Abu Saleh Musa Miah, Md Al Mehedi Hasan, Si-Woong Jang, Hyoun-Sup Lee, and Jungpil Shin. Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition. *Electronics*, 12(13):2841, 2023.
- [Momeni *et al.*, 2020] Liliane Momeni, Gul Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [Özdemir *et al.*, 2020] Oğulcan Özdemir, Ahmet Alp Kindiroğlu, Necati Cihan Camgöz, and Lale Akarun. Bosphorussign22k sign language recognition dataset. *arXiv preprint arXiv:2004.01283*, 2020.
- [Papadimitriou and Potamianos, 2023] Katerina Papadimitriou and Gerasimos Potamianos. 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, 2023.
- [Patra *et al.*, 2024] Suvajit Patra, Arkadip Maitra, Megha Tiwari, K Kumaran, Swathy Prabhu, Swami Punyeshwarananda, and Soumitra Samanta. Hierarchical windowed graph attention network and a large scale dataset for isolated indian sign language recognition. *arXiv preprint arXiv:2407.14224*, 2024.
- [Ronchetti *et al.*, 2023] Franco Ronchetti, Facundo Manuel Quiroga, César Estrebow, Laura Lanzarini, and Alejandro Rosete. Lsa64: an argentinian sign language dataset. *arXiv preprint arXiv:2310.17429*, 2023.
- [Ryumin *et al.*, 2023] Dmitry Ryumin, Denis Ivanko, and Elena Ryumina. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors*, 23(4):2284, 2023.
- [Sandler, 2006] Wendy Sandler. Sign language and linguistic universals. *Cambridge University*, 2006.
- [Shen *et al.*, 2024a] Xiaolong Shen, Zhedong Zheng, and Yi Yang. Stepnet: Spatial-temporal part-aware network for isolated sign language recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(7):1–19, 2024.
- [Shen *et al.*, 2024b] 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. MM-WLAuslan: Multi-View Multi-Modal Word-Level Australian Sign Language Recognition Dataset. *arXiv preprint arXiv:2410.19488*, 2024.
- [Sincan and Keles, 2020] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE access*, 8:181340–181355, 2020.
- [Sridhar *et al.*, 2020] Advait Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1366–1375, 2020.
- [Tornay *et al.*, 2020] Sandrine Tornay, Marzieh Razavi, and Mathew Magimai Doss. Towards multilingual sign language recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6309–6313. IEEE, 2020.
- [Wang *et al.*, 2016] Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. Isolated sign language recognition with grassmann covariance matrices. *ACM Transactions on Accessible Computing (TACCESS)*, 8(4):1–21, 2016.
- [Wei and Chen, 2023] Fangyun Wei and Yutong Chen. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621, 2023.
- [Yin *et al.*, 2022] Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlsit: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5109–5119, 2022.
- [Yun *et al.*, 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.
- [Zhang, 2017] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- [Zhao *et al.*, 2023] Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. Best: Bert pre-training for sign language recognition with coupling tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3597–3605, 2023.
- [Zhou *et al.*, 2021] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 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, 2021.
- [Zuo *et al.*, 2023] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14890–14900, 2023.
- [Zuo *et al.*, 2024] Ronglai Zuo, Fangyun Wei, and Brian Mak. Towards online sign language recognition and translation. *arXiv preprint arXiv:2401.05336*, 2024.