ISSN 1213-6972 (print)
ISSN 1213-6964 (online)
Journal of WSCG
http://www.wscg.eu
Vol.33, No-1-2, 2025

Training Strategies for Isolated Sign Language Recognition

Karina Kvanchiani
SaluteDevices
karinakvanciani@gmail.com

Roman Kraynov
SaluteDevices
ranakraynov@gmail.com

Elizaveta Petrova
SaluteDevices
kleinsbotle@gmail.com

Petr Surovcev
SaluteDevices
petr.surovcev@gmail.com

Aleksandr Nagaev
SaluteDevices
sashanagaev1111@gmail.com

Alexander Kapitanov
SaluteDevices
kapitanovalexander@gmail.com

ABSTRACT

Accurate recognition and interpretation of sign language are crucial for enhancing communication accessibility for deaf and hard of hearing individuals. However, current approaches of Isolated Sign Language Recognition (ISLR) often face challenges such as low data quality and variability in gesturing speed. This paper introduces a comprehensive model training pipeline for ISLR designed to accommodate the distinctive characteristics and constraints of the Sign Language (SL) domain. The constructed pipeline incorporates carefully selected image and video augmentations to tackle the challenges of low data quality and varying sign speeds. Including an additional regression head combined with IoU-balanced classification loss enhances the model's awareness of the gesture and simplifies capturing temporal information. Extensive experiments demonstrate that the developed training pipeline easily adapts to different datasets and architectures. Additionally, the ablation study shows that each proposed component expands the potential to consider ISLR task specifics. The presented strategies enhance recognition performance across various ISLR benchmarks and achieve state-of-the-art results on the WLASL and Slovo datasets.

Keywords

Isolated Sign Language Recognition, ISLR Dataset, Computer Vision

1 INTRODUCTION

Sign Languages are the primary means of communication for many deaf and hard of hearing individuals. According to data from the All-Russian Society of the Deaf (VOG[1]), there were more than 150,000 native speakers of Russian Sign Language (RSL) in 2019. Such languages do not replicate spoken language but possess their own lexicon and unique grammatical rules. Due to the language gap, deaf and hard of hearing people may experience prejudice in finding employment, pursuing academic education, or accessing medical services. Furthermore, learning Sign Language (SL) is challenging due to the limited number of teachers and native speakers.

However, developing robust Sign Language Recognition (SLR) systems remains challenging. Variations in gesturing speed, complex spatial-temporal features, and real-world capturing conditions (e.g., diverse backgrounds or lighting) can notably affect recognition accuracy. Furthermore, large-scale, high-quality datasets are still limited.

Sign language understanding is covered by three primary tasks: ISLR (Isolated Sign Language Recognition), CSLR (Continuous Sign Language Recognition), and SLT (Sign Language Translation). This research focuses on ISLR, where videos are classified into individual sign categories. One of the beneficial applications of this task is to create an automatic SL trainer, where users are shown a video example of a sign, and the recognition system assesses the quality of human gesturing. This advancement could make the learning of SL more accessible.

The development of automatic SLR systems is becoming widespread [16, 8] to facilitate communication between deaf and hearing individuals. SLR is a com-

[1] https://voginfo.ru/all-russian-society-of-the-deaf/

http://www.doi.org/10.24132/JWSCG.2025-3

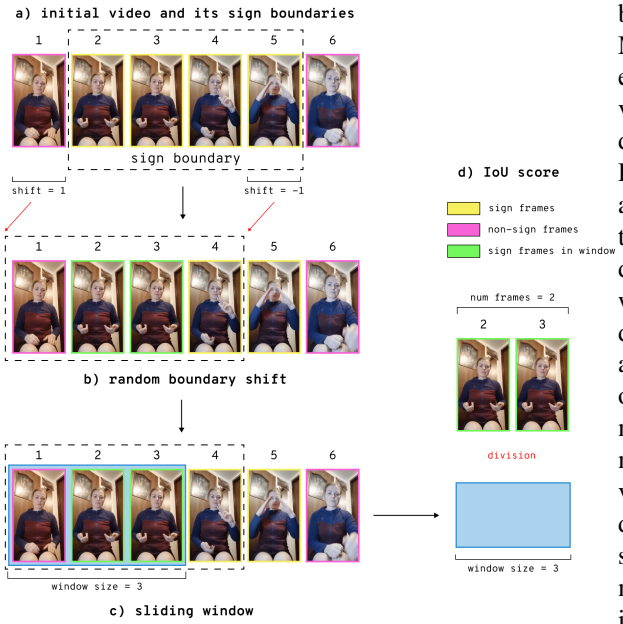

Figure 1: The process of receiving IoU scores. (a) Consider the initial video, where the sign is between frames 2 and 5. (b) Randomly shift the sign's boundaries by one frame to enhance model robustness (see Section 3.1 for details). (c) Collect frames by sliding a window of size 3 across the video (window size of 3 chosen only for illustration). (d) Calculate IoU scores by dividing the number of sign frames in the window by the window size and adjust the classification scores. Note that the window size of 3 in this figure was selected purely as an example and was not used in our experiments.

plex challenge due to the impact of hand movement, location, orientation, shape, and facial expressions on sign meaning [33]. The SLR system should function in real-world settings (schools, hospitals, or train stations). Its complexity is due to variations in sign display speeds, low video quality and resolution, diverse backgrounds, and varying lighting conditions. These factors adversely affect the system, which needs to operate with real-time response and maintain high quality to minimize errors. Besides real-world limitations, SLR faces domain-specific challenges such as limited data availability and complex temporal dependencies. Current methodologies often do not fully address these issues, limiting the development of models that generalize across diverse signing styles, varying execution speeds, and other conditions. To address these problems, we explore novel training strategies and propose a robust pipeline designed to handle the aforementioned challenges in the ISLR domain.

The basic approach to the ISLR task involves a classification neural network, designed to process RGB video data. The model takes the sequence of frames sampled from the video as input and predicts a text label corresponding to the sign depicted in the video. Many existing methodologies within this field provide either a solution at the model level [6, 44, 13], involving alterations to the model architecture, or at the data level [46], focusing on expanding the training data. Due to unaddressed domain-specific challenges, these approaches may not fully leverage the models' potential for ISLR. Our work introduces a training pipeline designed for real-world SLR. It boosts performance without altering the underlying model architecture or dataset. To address both domain-specific challenges and practical deployment constraints, we apply a series of modifications to the basic approach. These enhancements improve accuracy across multiple datasets and model types, including transformers [24, 37] and convolutional neural networks (CNNs) [4]: (1) video-level data augmentations to simulate the different gesturing speeds (shown in Figure 3); (2) image-level data augmentations to reproduce low video quality; (3) an auxiliary sign boundary regression head to direct the network to focus more on the frames containing signs; and (4) 1D Intersection-over-Union-balanced CrossEntropy (IoU-balanced CE) loss to enhance the model's capability to understand signs (see Figure 1). Considering the necessity for real-time response, we focus on modifying strategies solely based on RGB data without additional modalities like hand keypoints [29] and depth [19, 29] information.

The contribution of this paper is fourfold:

- We propose a versatile and scalable training pipeline for ISLR models that considers domain-specific constraints, including data quality, varying sign execution speeds, and sign boundary awareness.
- We present a novel large-scale dataset, SlovoExt, which combines the Slovo [21] dataset with our newly assembled Russian Isolated Sign Language Dataset; we also release[2] the code and pre-trained models to facilitate further research. The SlovoExt dataset will be available as part of a larger RSL dataset in future work.
- We illustrate how the proposed training strategies enable superior performance compared to existing solutions, bridging the gap between data-level and model-level improvements in SLR tasks.

2 RELATED WORK

Sign Language Recognition Training Heuristics. Many works in the SLR field have implemented model architecture-level alterations to improve performance. In [12], self-mutual distillation was employed to

[2] https://github.com/ai-forever/TrainingStrategiesISLR

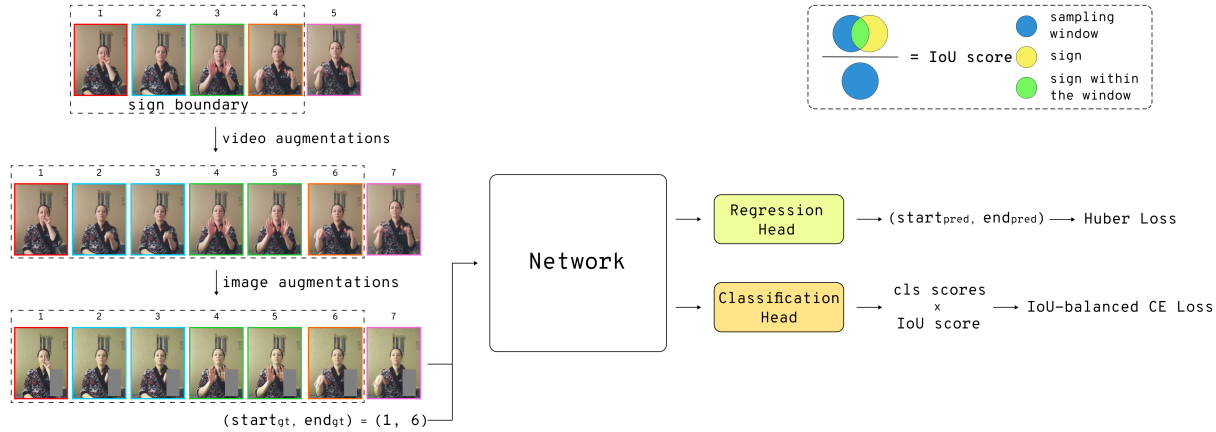

Figure 2: Overall training pipeline. Video-level and image-level augmentations are applied, and the neural network is further trained with augmented data and sign boundary annotations.

enable the model to learn temporal and spatial features. In [13], authors used correlation and identification modules to capture hands and face trajectories more effectively across consecutive frames. Other researchers have employed data-based approaches to boost their model. The authors of [46] utilized datasets of various sign languages for training to address the issue of insufficient data. Some approaches have incorporated not only RGB but also pose information [3], hand and body keypoints [20, 6, 44], and mouthing cues [32], combining data- and model-level modifications. This paper focuses on developing a training pipeline to be applied to any ISLR model that takes RGB videos as input and outputs classification scores.

Image & Video Augmentations. Many studies focusing on SLR have employed random crop [5, 6, 30, 1, 34] and horizontal flip [12, 30, 1, 34] as standard image augmentations. However, these studies often do not address how to handle non-mirrored signs, which should be kept the same since altering them could change their meaning. Zhou et al. [43] utilized strong data augmentation techniques, such as geometric and color space transformations, to enhance the model's robustness to data perturbations.

Video augmentations have proven widely effective in the action recognition task to enhance the model's capacity to capture temporal dependencies. However, actions in this task (e.g., walking or running) are frequently repetitive and can be correctly classified at any time. On the contrary, the order of movements is essential in the SLR task. The specificity is the main reason for the inability to apply some of the transformation proposed in [10], where authors augmented data with video reversing, frame mixing, and temporal extension of CutMix [40]. In [34], the authors proposed a video augmentation method where part of the video frames corresponding to one gesture is removed, added, or replaced with frames depicting a different gesture within the same video. This technique allowed the authors to address the data deficiency and leverage the contextual information within the sign sequence. In this paper, we only remove and add frames ("speed up" and "slow down" in Section 3.1) to speed adjustments because replacing is unsuitable for the ISLR task, given that only isolated signs are present. Ahn et al. [1] used different frame sampling rates to capture spatial and temporal information separately. We adopt a similar approach with randomness ("random add" and "random drop" in Section 3.1) to simulate real-life SLR cases. These video augmentations, shown in Figure 3, enable sign recognition regardless of display speed and address data insufficiency in the SLR task.

Auxiliary Regression Task. Localizing the action boundaries via regression loss has demonstrated benefits in solving action recognition tasks. This approach enabled Zhang et al. [41] to achieve single-stage anchor-free temporal action localization. Similarly, the authors of [45] divided the temporal action localization task into classification and action boundary regression, applying L1 loss for the latter. As was stated in [25], supplementary tasks can boost the performance of the main task by pushing the backbone toward learning robust and generalized representations. Considering this concept, we incorporate an auxiliary sign boundary regression task by adding a regression head optimized by Huber loss [14] to the training pipeline, as it has demonstrated better model convergence in the experiments provided below (see Table 5). It assists the model in understanding the temporal position of the sign within the video, allowing it to focus more on the frames containing the sign.

IoU-balanced CE Loss. Extracting information about an object's spatial and temporal boundaries proves beneficial in both action recognition [39] and SLR tasks [17, 32]. Techniques like IoU-loss [17, 27, 38] can aid this process. Wu et al. [38] proposed employing

the IoU-balanced CE loss to address the independence between classification and localization predictions. Liu et al. [27] employed a 1D IoU loss in the action detection task to localize relevant segments within the video. In [17], the Generalized IoU loss was employed for SLR by calculating IoU between the bounding boxes of hands in each frame. In the proposed pipeline, we combine the concepts of 1D IoU and the IoU-balanced classification losses to devise the classification IoU loss tailored for the video domain and the particularities of the ISLR task. When computing the classification score, this approach allows us to consider the relative localization of the sign (see Figure 1 for details).

3 TRAINING STRATEGIES

To address the existing gap in SLR methodologies, designing an effective training pipeline requires tailored approaches that account for both data characteristics and model requirements. Figure 2 shows the proposed strategies divided into three components: (1) video-level augmentations, (2) image-level augmentations, and (3) additional losses. Note that these techniques do not depend on specific SL datasets or model architectures, except for certain subcomponents detailed in Section 3.2.

3.1 Video Augmentation

Video augmentations are designed to reduce the gap between real cases and training data regarding subjects' gesturing speedby artificially changing the speed and shifting the sign boundaries.

Speed Up & Slow Down. In real life, signs can appear at various speeds, so we artificially speed up or slow down videos to match the real-world gesturing speed distribution. We sample every N-th frame to speed up by a factor of N (see Figure 3a), pre-sampling additional frames on the right to maintain the same clip length. Conversely, we repeat each frame N times to slow down the video (see Figure 3b), pre-dropping frames so that the final length remains unchanged.

Random Add & Random Drop. In some real-life scenarios, gesturing speed is non-uniform because of the difficulty or simplicity of some gesture parts or due to external distractions during the hand movement. We partially change the speed of the videos to make the model more resistant to such uncommon cases. Figure 3c illustrates the process of randomly dropping frames to speed up the video partially (see Algorithm 1). As in the case of uniform video speed up, we retain the same clip length using the identical process. Similarly, a random add of frames is applied to the video for a partial slow down (see Figure 3d). We are maintaining the video length in the same way as in a uniform slow down.

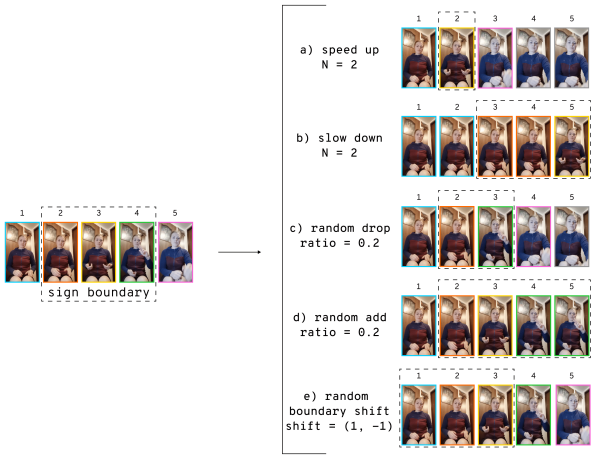

Figure 3: Applying video augmentations to untrimmed videos. Identical frames are highlighted in the same color, and sign boundaries are outlined with a dashed line. Grey boundaries indicate duplicates of the last frame. a) speed up the video 2 times by removing every second frame; b) slow down the video 2 times by duplicating every second frame; c) random frames drop remains 80% of the total video length; d) random frames duplication increases the total video length by 20%; e) random boundary shift is applied, e.g., with shift of (1, -1), which means one frame is added on the left and one is removed on the right. These values were chosen for ease of demonstration.

Random Boundary Shift. During inference, an SLR system generally captures frames sequentially using a window of a specific length. Therefore, signs that are too short or too long may not fit into the window properly. To ensure accurate recognition, we augment the data by randomly shifting sign boundaries, removing part of the videos, or adding frames without signs (see Figure 1 and Figure 3e for illustrations).

3.2 Image Augmentation

Due to the real-life limitations, the model must be adaptive to variations in background, subjects, image quality, and overall visual appearance. Image augmentations diversify the data in a frame-independent way to simulate real-life cases.

Image Quality. We imitate artifacts in video recordings caused by low-resolution capturing, due to defects in the video camera, or problems with video transmission over the network by compressing and downscaling images and adding random noise and sharpness.

Horizontal Flip. Most meanings of the signs remain unchanged after horizontal mirroring. Hence, horizontal flip transformation facilitates the equal processing of signs shown by different hands. However, some signs are not invariant to the horizontal flip (e.g., RSL gestures such as "left", "right", "heart", and "liver"

Algorithm 1 Random Drop Realization

1: $y = [video_i]_{i=0}^{31}$
// Initialize the video sequence y with frames indexed from 0 to 31
2: $drop\_ratio \leftarrow d; d \in (0,1)$
// Set the drop ratio d, which determines the rate of frames to drop
3: $y\_len \leftarrow len(y)$
// Calculate the original length of the video sequence y
4: $\hat{y}\_len \leftarrow \frac{y\_len}{1-drop\_ratio}$
// Compute the new length $\hat{y}\_len$ to extend the video sequence such that after dropping frames, the original length y_len is preserved
5: $y.extend([video_k]_{k=32}^{\hat{y}\_len})$
// Extend the video sequence y with additional frames indexed from 32 to $\hat{y}\_len$
6: $\hat{y} \leftarrow sort(random.choice(y, y\_len))$
// Randomly select y_len frames from the extended video sequence y without repetitions. Sort the selected frames to maintain the original order of frames

stop transferring their meanings if shown with the inappropriate hand). In the experiments, this augmentation only affects mirrored signs (more details in Section 5.1).

General. In addition to domain-specific augmentations, we utilize general ones to diversify the dataset: color jittering to provide heterogeneity over the color context, and CutMix [40] and MixUp [42] to induce the model to learn more generalizable features.

3.3 Additional Losses

The following domain-specific training techniques incorporate domain knowledge of the sign boundaries into the model by scaling the classification loss and solving an additional task to enhance the main task's solution accuracy. These modifications are not model-level, as they can be integrated into either architecture.

Sign Boundary Regression Head. The average gesturing speed of various signs may differ, and even the same sign may take a different number of frames, affecting data distribution. Motivated by this, we add the regression head with a fully connected layer parallel to the classification head. This auxiliary head predicts the start and end of a sign in the input video to embed an implicit understanding of the length and gesturing speed into the model's backbone. To train the regression head, we utilize Huber loss [14], which combines the strengths of both MSE and MAE losses: it offers sensitivity and robustness, being less affected by the influence of outliers.

Dataset	Classes	Videos	Signers	Resolution	Language
LSE-Sign [11]	2,400	2,400	2	FullHD	Spanish
LSA64 [35]	64	3,200	10	FullHD	Argentinian
MS-ASL [18]	1,000	25,513	222	varying	American
TheRuSLan [19]	164	13	13	FullHD	Russian
K-RSL [15]	600	28,250	10	FullHD	Kazakh-Russian
FluentSigners-50 [31]	278	43,250	50	varying	Kazakh-Russian
WLASL2000 [23]	2,000	21,083	119	varying	American
AUTSL [36]	226	38,336	43	512 × 512	Turkish
Slovo [21]	1,001	20,400	194	HD / FullHD	Russian
SlovoExt (ours)	1,001	51,000	241	HD / FullHD	Russian

Table 1: The main characteristics of the existing ISLR datasets.

Classification Loss Scaling. Conventional classification losses, like cross-entropy, have the limitation to not consider information about the localization of the sign in the video, which can cause irrelevant gradients to optimize. We use a IoU-balanced CE loss to incorporate information about the location of the sign relative to the window. To calculate IoU scores, the length of the intersection of the sign with the sampled window is divided by the size of the window:

$$IoU\,score = \frac{\min(w_{end}, s_{end}) - \max(w_{start}, s_{start})}{w_{end} - w_{start} + 1}, \quad (1)$$

where s_{start}, w_{start} are the first frames of the sign and the sampled window, and s_{end}, w_{end} are the last ones, respectively; $min(a,b)$ and $max(a,b)$ functions denote the minimum and maximum of two values, respectively (for illustration, see Figure 1 and Figure 2). The overall classification loss is the combination of classification and IoU scores.

4 DATASETS

We assess the training strategies described above on three large-scale ISLR datasets: WLASL [23], AUTSL [36], and Slovo [21]. They are the most diverse open-source data in terms of dataset contributors, performing gestures (signers), and contexts (see Table 1). Evaluating on these datasets alone is sufficient to confirm the pipeline's effectiveness.

- **WLASL** dataset consists of 21,083 RGB-based videos trimmed by sign boundaries. Each video contains only one sign in American Sign Language (ASL), and each sign is performed by at least 3 different signers in various dialects[3]. The dataset was recorded in a studio with solid-colored backgrounds. The WLASL contains non-mirrored signs, but since such signs in its vocabulary are not established, we estimate the impact of flip augmentation for all samples in Section 6.1.
- **AUTSL** was designed to simulate real-life context, i.e., different indoor and outdoor environments

[3] There is more than one way to show one word. Therefore, intraclass diversity occurs.

and diverse lighting conditions. It contains 38,336 trimmed video samples with 512 × 512 frame resolution and 20 different backgrounds. The AUTSL is divided into 226 Turkish Sign Language (TSL) signs performed by 43 signers. It comprises numerous similar signs, so models that can extract complex information from the input data must be used for training. The situation with non-mirrored signs is identical to the WLASL's, so applying the horizontal flip to videos is analyzed below.

- **Slovo** is the largest, most diversified, and the only publicly available RSL dataset. The dataset contains 20,400 HD and FHD untrimmed videos performed by 194 signers. It is divided into 1,001 classes, including the additional "no event" class, which indicates videos where the signer is not performing a sign. Slovo was crowdsourced, featuring real-life conditions and non-invariant signs to horizontal flip.

4.1 SlovoExt Dataset

We expand the Slovo dataset with a self-assembled 30,600 videos. The combination is called SlovoExt and comprises 51,000 samples divided into the same 1,001 classes as Slovo (see Table 1). The process of SlovoExt creation is constructed similarly to maintain the Slovo distribution characteristics such as video length, signer's appearance, and context heterogeneity identical. SlovoExt is also utilized to evaluate the proposed training strategies.

The dataset was created by native speakers of RSL and interpreters proficient in RSL. We involved diverse contributors in the data collection process to address concerns about differences in sign presentation between deaf and hard of hearing people. This approach aims to help neural networks manage gesture variability and mitigate the effects of native signer bias [9].

In collecting the dataset, we carefully considered the specific features of RSL, ensuring that all data was recorded in a single dialect. The dataset was compiled using pre-recorded video templates provided by the All-Russian Society of the Deaf. An exam consisting of 20 questions assessed RSL proficiency for participation in dataset recording. Experts who scored at least 90% were allowed to take on the tasks.

5 EXPERIMENTS

The effectiveness of the training pipeline is assessed by comparing two metrics: utilizing (1) the basic approach[4] and (2) proposed strategies. We fine-tune three architectures: two transformers (MViTv2-S with different pre-training and MViTv2-B [24]) and one CNN (I3D [4] with a ResNet-50) – on four ISLR datasets. The experiment results are provided in Table 2.

Dataset	Model	Pretrain Task	top-1 accuracy	
			basic approach	proposed pipeline
WLASL	MViTv2-S	MaskFeat	49.83	56.37 +6.54
		Classification	51.88	57.17 +5.29
	MViTv2-B		54.31	57.33 +3.02
	I3D		35.55	36.38 +0.83
AUTSL	MViTv2-S	MaskFeat	91.69	95.62 +3.93
		Classification	90.27	95.05 +4.78
	MViTv2-B		93.00	95.75 +2.75
	I3D		85.22	87.81 +2.59
Slovo	MViTv2-S	MaskFeat	71.45	81.57 +10.12
		Classification	77.54	80.97 +3.43
	MViTv2-B		79.31	81.34 +2.03
	I3D		62.79	63.82 +1.03
SlovoExt	MViTv2-S	MaskFeat	81.55	87.31 +5.76
		Classification	83.36	85.90 +2.99
	MViTv2-B		84.14	86.72 +2.58
	I3D		77.30	79.74 +2.44

Table 2: Evaluation results. We present two top-1 accuracy metrics for each setup: the metric on the basic approach and the metric obtained via the proposed training pipeline. Two pre-trained models on the K400 dataset [22] are utilized: MViTv2 trained on the classification task and MaskFeat trained in a self-supervised manner to reconstruct masked pixels. Green values show gains over the basic approach.

The top-1 accuracy is the primary metric for evaluation, measuring the percentage of videos where the predicted class matches the correct class. This metric was chosen due to its widespread use in the SLR task, facilitating comparison with other models and benchmarking. Mean accuracy was used to demonstrate state-of-the-art result on the Slovo dataset and ensure consistency with the metric on the leaderboard[5].

5.1 Preprocessing

Input videos are resized to 300 resolution and converted to HDF5 video format. During the training stage, videos are square-padded and randomly cropped to achieve a resolution of 224 × 224. The low side of the videos is resized to 224 and square-padded during the validation and testing stages. The model input is a sequence of 32 frames sampled with a step of 2. If the video does not have enough frames for sampling, the last frame is repeated to form a complete clip. For the Slovo and SlovoExt datasets, non-mirrored signs are excluded from the horizontal flip augmentation. The WLASL signs are not being flipped, and the AUTSL signs are flipped with no exceptions because Section 6 shows the efficacy of such decisions.

5.2 Training Methodology

As shown in Figure 2, the preprocessed video is initially exposed to video augmentations. One randomly se-

[4] The basic approach is the same as the proposed pipeline, but without video and image augmentations, regression head, and classification loss balancing.

[5] https://paperswithcode.com/sota/sign-language-recognition-on-slovo-russian

Ablation on	Image Augmentations		Video Augmentations					Additional Losses		top-1 acc.
	Basic	CutMix & MixUp	Bound. Shift	Rand. Add	Speed Up	Slow Down	Rand. Drop	Huber Loss	IoU-balanced CE	
None: entire pipeline	✓	✓	✓	✓	✓	✓	✓	✓	✓	**87.31**
Image Augs.	✗	✓	✓	✓	✓	✓	✓	✓	✓	86.08 −1.23
	✗	✗	✓	✓	✓	✓	✓	✓	✓	83.25 −4.06
	✓	✗	✓	✓	✓	✓	✓	✓	✓	85.42 −1.89
Video Augs.	✓	✓	✗	✗	✗	✗	✗	✓	✓	86.77 −0.54
	✓	✓	✗	✓	✓	✓	✓	✓	✓	87.00 −0.31
	✓	✓	✓	✗	✓	✓	✓	✓	✓	86.96 −0.35
	✓	✓	✓	✓	✗	✓	✓	✓	✓	86.99 −0.32
	✓	✓	✓	✓	✓	✗	✓	✓	✓	86.65 −0.66
	✓	✓	✓	✓	✓	✓	✗	✓	✓	87.01 −0.30
Losses	✓	✓	✓	✓	✓	✓	✓	✗	✓	86.79 −0.52
	✓	✓	✓	✓	✓	✓	✓	✗	✗	86.95 −0.36
	✓	✓	✓	✓	✓	✓	✓	✓	✗	87.26 −0.05

Table 3: Ablation study results. We employ the MViTv2-S model to assess the influence of each component by removing it from the pipeline and comparing the results. Accuracy is evaluated on the SlovoExt combination. Red values indicate differences with the entire pipeline.

Method	top-1 accuracy
RandAugment	86.79 −0.52
UniformAugment	86.75 −0.56
Sequential (utilized)	87.31

Table 4: Ablation study for different image augmentation methods.

Regression Loss	top-1 accuracy
MAE	86.68 −0.63
MSE	87.00 −0.31
Huber (utilized)	87.31

Table 5: Ablation study for different regression losses.

lected from four modifications is applied to each batch. In contrast, “random boundary shift" is constantly affected. In experiments, “speed up" increases the video speed by 2 times, and “slow down" decreases it by 2 times. “Random drop" shortens video by 10%, while “random add" lengthen it by 30%. The random values for sign boundaries shifting are from the interval [-5, 5], except the interval [-5, 0] for the WLASL and AUTSL datasets due to their trimming by the sign boundaries.

Image augmentations modify the video batch by randomly choosing a combination of them with expertly selected probabilities and magnitudes. Note that CutMix [40] and MixUp [42] influence only transformer models in our experiments because they can degrade CNNs due to a lack of correlation between neighboring pixels [2].

We incorporate IoU scores into the classification head, multiplying classification scores by them. The additional regression head is fed with video features and sign boundaries. The network is trained by optimization IoU-balanced CE and Huber losses for classification and regression heads, respectively. Both are not applied to “no event" videos in the Slovo and the SlovoExt datasets by assigning an IoU score of one and sign boundaries of zeros. Although sign boundaries trim the WLASL and AUTSL datasets, the regression head works appropriately with them.

Training is performed until convergence on four Tesla H100s with 80GB RAM. Early stopping is triggered if the top-1 accuracy metric does not increase by at least 0.003 after 7 epochs. For the first 20 epochs, the learning rate is modified by a linear scheduler, and a cosine scheduler is used from epochs 20 to 100. Cross-entropy loss is minimized using the AdamW [28] optimizer. Other parameters are variable and can be adjusted as needed.

5.3 Results

As Table 2 shows, the proposed pipeline consistently improves top-1 accuracy across different datasets and architectures. Specifically, MViTv2-S achieves additional gains of 6.54%, 3.93%, 10.12%, and 5.76% on WLASL, AUTSL, Slovo, and SlovoExt, respectively, while I3D obtains an average improvement of 1.72% across these datasets. The ablation study (see Table 3) further underscores the effectiveness of each proposed component in the pipeline.

Moreover, this approach attains state-of-the-art performance on two benchmark ISLR datasets. On WLASL, using MViTv2-S pre-trained with MaskFeat and further pretrained on SlovoExt, the model achieves 62.89% top-1 accuracy, surpassing the NLA-SLR [47] model by 1.63%. On the Slovo dataset, this method reaches 78.21% mean accuracy, exceeding the prior best result by 14.12%. A comparable increase of 13.06% is achieved with the MViTv2-S pre-trained on K400, confirming the robustness of the proposed training strategies.

6 ABLATION STUDY

This section estimates the impact of each part of the proposed pipeline individually. We divide all modifications into three blocks: (1) image augmentations, (2) video augmentations, and (3) additional losses. We test the necessity for each block by disconnecting it from the entire pipeline and for its parts by shutting down a particular part while not changing anything else. The

ablation study uses MViTv2-S pre-trained on the MaskFeat to train on the SlovoExt combination.

6.1 Image Augmentations Necessity

The image augmentations block is additionally divided into two parts – basic image transformations and CutMix-MixUp pair – to simplify assessing their influence. Each part and combinations of them are shut down, resulting in three experiments. The first three rows in Table 3 illustrate that the metric significantly decreases by 4.06% in the absence of all image augmentations. The influence of basic image augmentation is two times less than the CutMix-MixUp one.

We also evaluate various augmentation-applying approaches, including RandAugment [7], UniformAugment [26], and random sequential applications. Table 4 indicates that RandAugment and UniformAugment approaches are not effective in the proposed pipeline.

Additionally, since we are uncertain which signs are non-mirrored in WLASL and AUTSL datasets, we search for the optimal option via two experiments for each: one with flip augmentation and one without for all signs. On WLASL, the metric with a flip is lower by 1.3%, indicating that this dataset probably contains numerous non-mirrored signs. On AUTSL, conversely, the metric with flip was higher by 0.11%. Therefore, in the main experiments, we apply flip augmentation to the AUTSL dataset and refrain from using it with the WLASL dataset.

6.2 Video Augmentations Necessity

Similar to the above subsection, we split the estimation of video augmentations necessity by disconnecting each of them separately and the whole block, providing six experiments. The most notable impact achievable with one modification is produced by the “slow down" augmentation, as without it, the top-1 accuracy drops by 0.66%.

6.3 Additional Losses Necessity

The process of evaluation is constructed identically. The last three rows in Table 3 show that the regression loss provides a more substantial improvement than the IoU-balanced CE loss, resulting in a 0.52% increase in metrics compared to 0.05% increase. The result suggests that information about absolute sign boundaries is more crucial than relative sign position within the sampling window.

Also, we conduct additional research on the impact of Huber regression loss by replacing it with MAE and MSE losses (see Table 5). We observe lower metrics in both cases, with MAE yielding the lowest at 86.68%. These results could be attributed to the behavior of MAE, which, while descending rapidly, may get stuck on values close to the ground truth, overshooting the desired target value.

7 DISCUSSION

Ethical Considerations. All participants signed consent forms before data mining. The forms authorized the processing and publication of personal data for research purposes. We do not restrict videos with signers under 18 since parental permission was obtained during the registration, which complies with the Civil Code of the Russian Federation[6]. To preserve contributors' privacy, we employ anonymized user hash IDs in the dataset annotations. Furthermore, we have ensured that the Slovo dataset meets these ethical criteria. We provide the dataset for research purposes only, but we understand that it could be misused for malicious purposes, such as identifying people or large-scale surveillance.

Positional Statement. Through the research of the ISLR task, we involved the All-Russian Community of Deaf experts and professional sign language interpreters. The expertise of the All-Russian Society of the Deaf was utilized at every stage of the SlovoExt dataset creation, including data collection, validation, and verification processes, as well as video tagging. We also involved deaf consultants in developing training strategies to apply Considerations to particular solutions. Additionally, some of our researchers took formal courses on RSL to enhance their knowledge of this domain.

Limitations. The regression head and IoU-balanced CE loss require sign boundary annotations that are frequently difficult to achieve. Thus, when the dataset is pre-trimmed, the “random boundary shift" augmentation can adjust the sign boundaries only inward toward the center rather than allowing shifts in both directions. We processed RGB frames and did not analyze articulation, keypoints, or depth information. Moreover, a lack of awareness regarding non-mirrored signs in sign language may lead to issues when applying flip augmentation. There are additional challenges in each sign language, and each has its specifics that must be considered when adapting the pipeline.

8 CONCLUSION

In this paper, we introduce a training pipeline for the ISLR models, considering the specifics of the SLR domain and the constraints of real-world usage. We demonstrate the effectiveness of applying image and video augmentations to address the issues of low data quality and varying gesturing speed. The importance of integrating temporal information into the model by the regression head combined with IoU-balanced CE loss is also presented. The developed pipeline delivers state-of-the-art results on the WLASL and Slovo datasets.

[6] https://ihl-databases.icrc.org/en/national-practice/federal-law-no-152-fz-personal-data-2006

Future work will extend the training strategies to CSLR and SLT tasks.

ACKNOWLEDGEMENTS

We are grateful to Alena Fenogenova, Albina Akhmetgareeva and Anastasia Vasyatkina for the discussions and comments on this work.

REFERENCES

[1] Junseok Ahn, Youngjoon Jang, and Joon Son Chung. Slowfast Network for Continuous Sign Language Recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3920–3924. IEEE, 2024.

[2] Sihun Baek, Jihong Park, Praneeth Vepakomma, Ramesh Raskar, Mehdi Bennis, and Seong-Lyun Kim. Visual transformer meets cutmix for improved accuracy, communication efficiency, and data privacy in split learning. In *arXiv preprint arXiv:2207.00234*, 2022.

[3] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 301–319. Springer, 2020.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] 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.

[6] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. In *Advances in Neural Information Processing Systems*, volume 35, pages 17043–17056, 2022.

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[8] Aakash Deep, Aashutosh Litoriya, Akshay Ingole, Vaibhav Asare, Shubham M Bhole, and Shantanu Pathak. Realtime sign language detection and recognition. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–4. IEEE, 2022.

[9] Aashaka Desai, Maartje De Meulder, Julie A Hochgesang, Annemarie Kocab, and Alex X Lu. Systemic Biases in Sign Language AI Research: A Deaf-Led Call to Reevaluate Research Agendas. In *arXiv preprint arXiv:2403.02563*, 2024.

[10] Artjoms Gorpincenko and Michal Mackiewicz. Extending Temporal Data Augmentation for Video Action Recognition. In *Computer Vision – ECCV 2022 Workshops*, pages 116–133. Springer, 2022.

[11] Eva Gutierrez-Sigut, Brendan Costello, Cristina Baus, and Manuel Carreiras. LSE-sign: A lexical database for spanish sign language. In *Behavior Research Methods*, volume 48, pages 123–137. Springer, 2016.

[12] Aiming Hao, Yuecong Min, and Xilin Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11303–11312, 2021.

[13] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2529–2539, 2023.

[14] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[15] Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The k-rsl. In *Proceedings of the 24th conference on computational natural language learning*, pages 631–640, 2020.

[16] Vaishnavi Jadhav, Priyal Agarwal, Dhruvisha Mondhe, Rutuja Patil, and CS Lifna. A Survey of Sign Language Recognition Systems. In *booktitle of Innovative Image Processing*, volume 4, pages 237–246, 2022.

[17] Sanyam Jain. ADDSL: hand gesture detection and sign language recognition on annotated danish sign language. In *arXiv preprint arXiv:2305.09736*, 2023.

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

[19] Ildar Kagirov, Denis Ivanko, Dmitry Ryumin, Alexander Axyonov, and Alexey Karpov. TheRuSLan: Database of Russian sign language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6079–6085, 2020.

[20] Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. Sign language translation with hierarchical spatio-temporal graph neural network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3367–3376, 2022.

[21] 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.

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

[23] 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.

[24] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya 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.

[25] Lukas Liebel and Marco Körner. Auxiliary tasks in multi-task learning. In *arXiv preprint arXiv:1805.06334*, 2018.

[26] Tom Ching LingChen, Ava Khonsari, Amirreza Lashkari, Mina Rafi Nazari, Jaspreet Singh Sambee, and Mario A Nascimento. Uniformaugment: A search-free probabilistic data augmentation approach. In *arXiv preprint arXiv:2003.14348*, 2020.

[27] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. In *IEEE Transactions on Image Processing*, volume 31, pages 5427–5441. IEEE, 2022.

[28] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[29] Lu Meng and Ronghui Li. An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network. In *Sensors*, volume 21, page 1120, 2021.

[30] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11542–11551, 2021.

[31] Medet Mukushev, Aidyn Ubingazhibov, Aigerim Kydyrbekova, Alfarabi Imashev, Vadim Kimmelman, and Anara Sandygulova. FluentSigners-50: A signer independent benchmark dataset for sign language processing. In *Plos one*, volume 17, page e0273649. Public Library of Science San Francisco, CA USA, 2022.

[32] KR Prajwal, Hannah Bull, Liliane Momeni, Samuel Albanie, Gül Varol, and Andrew Zisserman. Weakly-supervised fingerspelling recognition in british sign language videos. In *arXiv preprint arXiv:2211.08954*, 2022.

[33] Alexey Prikhodko, Mikhail Grif, and Maxim Bakaev. Sign language recognition based on notations and neural networks. In *Digital Transformation and Global Society: 5th International Conference, DTGS 2020, St. Petersburg, Russia, June 17–19, 2020, Revised Selected Papers 5*, pages 463–478. Springer, 2020.

[34] Junfu Pu, Wengang Zhou, Hezhen Hu, and Houqiang Li. Boosting continuous sign language recognition via cross modality augmentation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1497–1505, 2020.

[35] Franco Ronchetti, Facundo Manuel Quiroga, César Estrebou, Laura Lanzarini, and Alejandro Rosete. LSA64: an Argentinian sign language dataset. In *arXiv preprint arXiv:2310.17429*, 2023.

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

[37] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[38] Shengkai Wu, Jinrong Yang, Xinggang Wang, and Xiaoping Li. Iou-balanced loss functions for single-stage object detection. In *Pattern Recognition Letters*, volume 156, pages 96–103. Elsevier, 2022.

[39] Wanru Xu, Zhenjiang Miao, Jian Yu, and Qiang Ji. Action recognition and localization with spatial and temporal contexts. In *Neurocomputing*, volume 333, pages 351–363. Elsevier, 2019.

[40] 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 (ICCV)*, pages 6022–6031, 2019.

[41] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.

[42] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *arXiv preprint arXiv:1710.09412*, 2017.

[43] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881, 2023.

[44] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13009–13016, 2020.

[45] Zixin Zhu, Le Wang, Wei Tang, Ziyi Liu, Nanning Zheng, and Gang Hua. Learning disentangled classification and localization representations for temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3644–3652, 2022.

[46] Ronglai Zuo and Brian Mak. Improving continuous sign language recognition with consistency constraints and signer removal. In *ACM Transactions on Multimedia Computing, Communications and Applications*, volume 20, pages 1–25. ACM New York, NY, 2024.

[47] 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.