SSLAL: Leveraging Semi-Supervised and Active Learning for Robust Depth-Enhanced Sign Language Recognition

Runzhou Chen
Johns Hopkins University
rchen90@jh.edu

Kunlun Li
Johns Hopkins University
kli90@jh.edu

Yixuan Wang
University of Florida
Johns Hopkins University
wang.yixuan@ufl.edu

Eric Nalisnick*
Johns Hopkins University
nalisnick@jhu.edu

Abstract

There is a scarcity of data on which to train models for sign language processing, which presents an obstacle for developing technologies for people whose native or preferred language is visual. We propose a semi-automated pipeline to help with data scarity for the specific task of isolated sign recognition (ISR). First, we compensate for MediaPipe's limitations in depth extraction by using an off-the-shelf neural depth estimator, thereby enriching the representation of hand movements and facial expressions. Second, SSLAL employs a combination of Semi-Supervised Learning (SSL) and Active Learning (AL), allowing the model to prioritize training on the most informative and uncertain data points. This strategy enables the recognition model to focus on high-value data, improving learning efficiency within resource-constrained environments. We demonstrate improvements in error rate and F1 on public benchmarks such as ASL-CITIZEN and GISLR.

1. Introduction

There are around 200-300 signed languages currently being used around the globe, with most users having some connection to Deaf communities. However, building technologies that can process data in the form of sign language is still a challenging technical problem, with computer vision as a core operation. Most of today's computer vision problems enjoy a plethora of data to train upon, but this is not the case for *sign language processing* (SLP). Large data sets cannot simply be extracted from the internet, and the availability of native or fluent sign language users is a bottleneck to collecting high quality annotations.

*Corresponding author

In this paper, we focus on data collection issue for *isolated sign recognition* (ISR), a sub-task of SLP. In this task, a single sign is captured by raw video or a sequence of pose landmarks, and a classifier much predict the discrete label that corresponds to the sign's meaning. It is analogous to the task of matching the string 'dog' to a one-hot encoding for written language processing. ISR is vitally important for just about any downstream SLP technology. For example, a mobile application that helps users learn a sign language would need to analyze the signs produced by a user, to assess correctness and their overall skills at expression. And when performing SLP for a sequence of signs, one approach is to partition the sequence into isolated signs and apply ISR.

In this work, we propose an automated pipeline for ISR to improve its data quality. The pipeline consists of three keys steps:

1. **depth estimation**: As many ISR datasets are formed of single-view images or videos, they lack depth information that is crucial for distinguishing signs that have outward motion. We apply off-the-shelf neural depth estimation and use it to update the landmark representation.
2. **semi-supervised pre-training**: As labeled data is scarce, we apply a pre-trained model to an unlabeled set and incorporate additional training points vs confidence-based pseudo-labeling.
3. **active learning**: To efficiently collect additional labels, we apply active learning, using uncertainty sampling to select the 'most interesting' points from the pool set for labeling.

We apply this pipeline to common ISR benchmarks—-ASL-CITIZEN, MS-ASL, WLASL, GISLR—showing that it improves the error rate and F1 score of the down-stream classifier.

2. Related Work

ISLR refers to the task of recognizing individual sign gestures from video sequences in which each sign is presented independently, without the influence of surrounding contextual signs. In contrast to continuous sign language recognition (CSLR), which processes sequences of connected signs within sentences, ISLR focuses on the identification of distinct, discrete signs [15, 22]. This task presents significant challenges due to the substantial variability in hand configurations, motion trajectories, and non-manual markers, such as facial expressions, which can differ markedly across signers and signing environments [18, 24]. Traditional methods for ISLR have predominantly relied on fully supervised learning approaches that require large-scale annotated datasets to achieve satisfactory performance [25]. However, the annotation of sign language data is a resource-intensive process, making the acquisition of extensive labeled datasets a persistent challenge [9]. Consequently, recent research has shifted toward more efficient learning paradigms that can leverage both labeled and unlabeled data to address the limitations posed by the scarcity of labeled sign language data [17, 31].

A survey by Sheikhpour *et al*. [35] presents SSL as a learning paradigm designed to leverage both labeled and unlabeled data, significantly reducing the reliance on large annotated datasets. The core theory of SSL is built on the cluster and manifold assumptions, which suggest that data points within the same cluster or on a low-dimensional manifold are likely to share the same label, thereby enabling effective label propagation [7, 39]. Various SSL methods, such as self-training and co-training, exploit these assumptions to refine learning processes through iterative or multi-view approaches [40]. SSL has been widely applied in diverse fields, including image classification, where it enhances object detection by incorporating unlabeled images [34, 37], speech recognition, where it helps models better understand audio data [20], and natural language processing, where it improves text classification and entity recognition tasks by utilizing large unlabeled corpora [9].

AL has emerged as a critical technique in machine learning, aimed at reducing labeling costs by strategically selecting the most informative data points for annotation [4, 30]. The primary idea behind AL is that not all data contribute equally to model performance, and by querying the most uncertain or representative samples, AL allows models to achieve better accuracy with fewer labeled examples [1, 33]. Researchers have developed several AL strategies, including uncertainty sampling, which selects instances with the least confident predictions, and query-by-committee, where multiple models identify disagreements on certain data points to highlight ambiguity [12, 21]. AL has been widely applied in domains such as image classification, where it helps prioritize complex images for labelling, and natural language processing, where it selects uncertain text passages for human annotation [29]. In medical diagnostics, AL is utilized to identify the most informative patient data for annotation, enhancing predictive models while minimizing the need for exhaustive labeling [39]. The adaptability and efficiency of AL make it indispensable in data-scarce environments where manual annotation is expensive and time-consuming [36].

In image classification and face recognition, SSL with AL frameworks have been applied to improve classification models by iteratively refining both labeled and unlabeled datasets [13]. For instance, one approach leveraged SSL to extract class-relevant features from unlabeled data, while AL selected only the most informative samples for labeling, thereby improving model robustness and reducing annotation costs [15, 16]. Similarly, semi-supervised dictionary active learning (SSDAL) [41] has demonstrated significant benefits in large-scale image classification, overcoming noise in unlabeled data and providing a highly efficient framework for tasks with limited labeled data. These approaches effectively balance SSL's capacity to exploit abundant unlabeled data with AL's capability to enhance classifier quality through strategic sample selection [5, 34]. This study builds upon these advancements, applying the combined strengths of SSL and AL to isolated word sign language recognition, where labeled gloss data is scarce and difficult to obtain [20, 28].

3. Datasets

To enhance spatial data representation in existing ISLR datasets, we propose a novel pipeline that refines depth information and addresses landmark occlusions. We apply this approach to datasets such as ASL-Citizen [11], MS-ASL [19], and WLASL [26], which traditionally consist of raw videos with gloss annotations but lack detailed landmark data.

3.1. Depth Refinement

MediaPipe can extract reliable 2D joint landmarks $J_{f,i}$, representing the x and y coordinates of each landmark i at frame f, along with relative depth values z^{MP} that indicate relative position of the object. However, it does not provide accurate depth measurements. To obtain precise depth information, we incorporate DepthAnythingV2 at the beginning of our pipeline to extract depth values at the 2D joint positions $J_{f,i}$.

While DepthAnythingV2 enhances the depth information, it faces challenges in determining the depth of occluded objects. If occluded landmarks share the same 2D positions as visible ones, identical depth values will be assigned to them, leading to inaccuracies in depth representation. To address this limitation, we design a Depth Overlap Refinement block that combines the depth values z_k^{depth} with

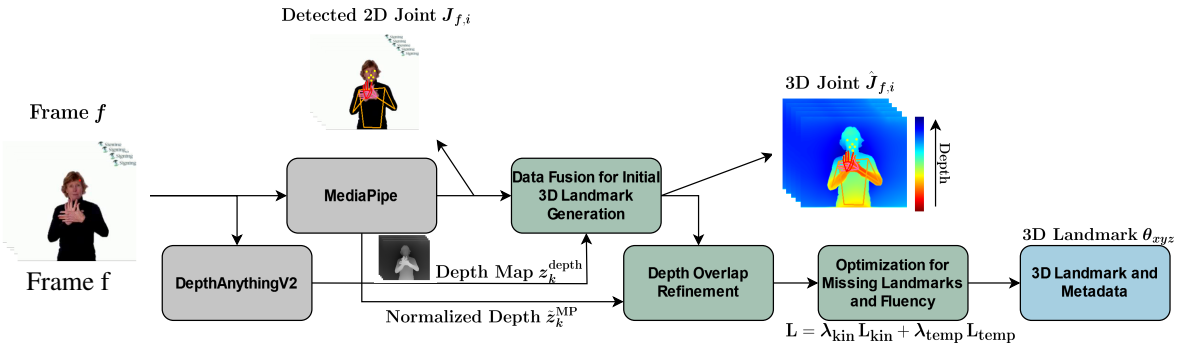

Figure 1. Overview of the 3D landmark optimization pipeline. Each frame f is processed by MediaPipe [27] and DepthAnythingV2 [38] to extract 2D landmarks $J_{f,i}$, relative depth z_k^{MP}, and depth map z_k^{depth}. These are fused to generate initial 3D landmarks, followed by depth overlap refinement. An optimization step then estimates missing landmarks and improves fluency by enforcing kinematic and temporal constraints, resulting in in perform$_r etrieval_a nd_s ave_s klearnacoherent3Drepresentation\boldsymbol{\theta}_{xyz}$.

normalized relative depth cues z^{MP}, ensuring overlapping landmarks are assigned distinct depth values. Specifically, the adjusted depth z_k^{adj} for each landmark k is computed as:

$$z_k^{\text{adj}} = z_k^{\text{depth}} + \alpha\, \tilde{z}_k^{\text{MP}}, \tag{1}$$

where z_k^{depth} is the depth from DepthAnythingV2, $\tilde{z}_k^{\text{MP}}$ is the normalized MediaPipe relative depth, and α is a scaling factor set to 0.3. The normalization of MediaPipe's relative depth is given by:

$$\tilde{z}_k^{\text{MP}} = \frac{z_k^{\text{MP}} - \mu_{\text{MP}}}{\sigma_{\text{MP}}}, \tag{2}$$

with μ_{MP} and σ_{MP} representing the mean and standard deviation of the relative depth values across all landmarks in a frame.

3.2. Optimizing for Landmark Occlusions

Occlusions can lead to missing landmarks, especially those connected to detected ones. To predict their positions, we employ an optimization procedure using the refined depth z^{adj} from the previous step and the detected x and y coordinates from MediaPipe. This enhances the precision of hand poses and corrects potential misalignments.

Our optimization aims to minimize a loss function that balances kinematic constraints and temporal coherence:

$$L = \lambda_{\text{kin}} L_{\text{kin}} + \lambda_{\text{temp}} L_{\text{temp}}, \tag{3}$$

where λ_{kin} and λ_{temp} are hyperparameters controlling the influence of each term.

The kinematic constraint term L_{kin} ensures anatomical plausibility by enforcing consistent distances between connected landmark pairs:

$$L_{\text{kin}} = \sum_{(i,j)\in E} \left(\|\hat{\mathbf{J}}_{f,i} - \hat{\mathbf{J}}_{f,j}\|_2 - l_{ij} \right)^2, \tag{4}$$

where $\hat{\mathbf{J}}_{f,i} = (x_i, y_i, z_i)$ denotes the estimated 3D position of landmark i at frame f, l_{ij} is the known distance between landmarks i and j, and E is the set of connected landmark pairs in the hand skeleton.

Temporal coherence is promoted through the term L_{temp}, encouraging smooth transitions between consecutive frames:

$$L_{\text{temp}} = \sum_i \left(\|\hat{\mathbf{J}}_{f,i} - \hat{\mathbf{J}}_{f-1,i}\|_2^2 + \|\hat{\mathbf{J}}_{f,i} - \hat{\mathbf{J}}_{f+1,i}\|_2^2 \right), \tag{5}$$

where $\hat{\mathbf{J}}_{f\pm1,i}$ are the estimated positions in adjacent frames.

In this optimization, we adjust the x, y, and z coordinates of missing landmarks connected to detected ones. For detected landmarks, x and y are fixed from MediaPipe, and z is set to z^{adj}. Minimizing L allows us to infer missing landmark positions while ensuring anatomical plausibility and temporal smoothness.

3.3. Data Representation

The updated 3D representation is stored in Parquet files, with each frame containing:

$$\mathcal{L}_{\text{frame}} = [\theta_f \,\|\, \theta_r \,\|\, \tau \,\|\, \theta_l \,\|\, \theta_{xyz}], \tag{6}$$

where θ_f is the frame number, θ_r is the unique row ID, τ is the landmark type (e.g., face, left hand, pose, right hand), θ_l is the landmark index, and θ_{xyz} are the spatial coordinates (x, y, z); $\|$ denotes concatenation.

A corresponding gloss transcript file includes metadata for each sequence:

$$\mathcal{T} = [\psi \,\|\, \theta_d \,\|\, \theta_s \,\|\, \sigma], \tag{7}$$

where ψ represents the file path, θ_d is the participant ID, θ_s is the sequence ID, and σ is the sequence label.

An overview of the proposed pipeline is depicted in Figure 1, illustrating the flow from depth refinement to occlusion optimization and data representation.

4. Methods

The proposed methodology employs a combination of SSL and AL to effectively leverage both labeled and unlabeled data in the context of ISLR. As depicted in Figure 2, the process begins with an initial set of labeled data that serves as the foundation for model training. This labeled dataset is subsequently expanded through an SSL framework, which incorporates unlabeled data, exploiting its structure to enhance the model's generalization capabilities [40]. AL is then applied to selectively identify the most informative and representative samples from the remaining unlabeled data pool, ensuring that only high-value instances are annotated and integrated into the training set. These refined samples are then fed into a Transformer model, where they undergo further training and prediction.

4.1. Semi-Supervised Learning Processing

Following the initial supervised training, the model utilizes **confidence-based pseudo-labeling** to expand its training set by leveraging unlabeled data. In this step, the model predicts labels for unlabeled samples and selectively accepts those with high prediction confidence, minimizing the risk of incorporating noise from uncertain predictions. This process is formalized as:

$$\hat{y}_{\text{p},j} = \begin{cases} g_\phi(x_j) & \text{if } \max(g_\phi(x_j)) > \tau \\ \text{discard} & \text{otherwise} \end{cases} \tag{8}$$

where $\hat{y}_{\text{p},j}$ represents the pseudo-label assigned to the unlabeled sample x_j if it meets the confidence threshold τ. Here, $\max(g_\phi(x_j))$ denotes the highest probability among the predicted class probabilities for x_j, serving as an indicator of the model's confidence in its prediction. By establishing a threshold τ, this function selectively filters out low-confidence predictions, ensuring that only reliable pseudo-labels contribute to the model's training, which strengthens the learning process while mitigating noise.

To further refine the model's robustness, we incorporate **consistency regularization**, which encourages stable predictions under minor input perturbations[29]. This regularization is defined as:

$$\mathcal{L}_{\text{c}} = \frac{1}{N} \sum_{j=1}^{N} \|g_\phi(x_j) - g_\phi(x_j + \delta)\|^2 \tag{9}$$

where δ represents a small perturbation applied to the input x_j. This term penalizes discrepancies between the model's predictions on original and perturbed inputs, reinforcing the model's resilience to variations in sign language gestures. By combining high-confidence pseudo-labeling with consistency regularization, the model achieves greater generalization, effectively utilizing both labeled and unlabeled data to optimize performance in isolated word sign language recognition.

To enhance the stability of pseudo-labels across training epochs, we employ **Temporal Ensembling**, a technique that maintains a running ensemble prediction for each unlabeled sample by combining predictions over multiple epochs. This approach minimizes fluctuations in pseudo-labels, creating more consistent targets for subsequent training iterations. The ensemble update mechanism is formalized as:

$$\hat{y}_{\text{e},j}^{(t)} = \alpha \cdot \hat{y}_{\text{e},j}^{(t-1)} + (1 - \alpha) \cdot g_\phi^{(t)}(x_j) \tag{10}$$

where $\hat{y}_{\text{e},j}^{(t)}$ represents the ensemble prediction for the j-th sample at the current epoch t. The parameter α is a momentum factor that controls the influence of the previous ensemble prediction, $\hat{y}_{\text{ensemble},j}^{(t-1)}$, in the current update, while $g_\phi^{(t)}(x_j)$ denotes the model's current prediction for the sample x_j at epoch t with parameters ϕ. By adjusting α, the ensemble method balances past and present predictions, ensuring that the pseudo-labels evolve gradually, thereby reducing sensitivity to transient fluctuations. This technique is particularly effective in isolated word sign language recognition, where high variability in gesture patterns can lead to inconsistent pseudo-labels. Temporal Ensembling thus enables more robust and reliable learning from unlabeled data by providing stabilized pseudo-labels for further training.

4.2. Active Learning Processing

Building upon our semi-supervised learning approach, we incorporate **entropy-based selection** within the Active Learning framework to further enhance the model's performance on isolated word sign language recognition. This selection method is particularly beneficial in this domain, as it highlights ambiguous gestures with overlapping hand shapes or similar motions—areas where the model often encounters uncertainty. By calculating the entropy $H(f(x_i))$ of each sample x_i in the unlabeled pool, we can prioritize samples where labeling would yield the greatest improvement to the model's decision boundaries [4]. Entropy, defined as:

$$H(f(x_i)) = -\sum_{j=1}^{C} p(y_j|x_i) \log p(y_j|x_i) \tag{11}$$

where $p(y_j|x_i)$ is the probability that sample x_i belongs to class y_j, and C denotes the total number of classes, serves as a quantitative measure of uncertainty. Higher entropy values indicate a lack of confident predictions across multiple classes, suggesting that the model is unsure of the correct label. By focusing labeling efforts on high-entropy samples, entropy-based selection identifies cases most likely to refine the model's decision boundaries.

Following entropy-based selection, we apply **K-Means Clustering with Core-Set Selection** to ensure diversity

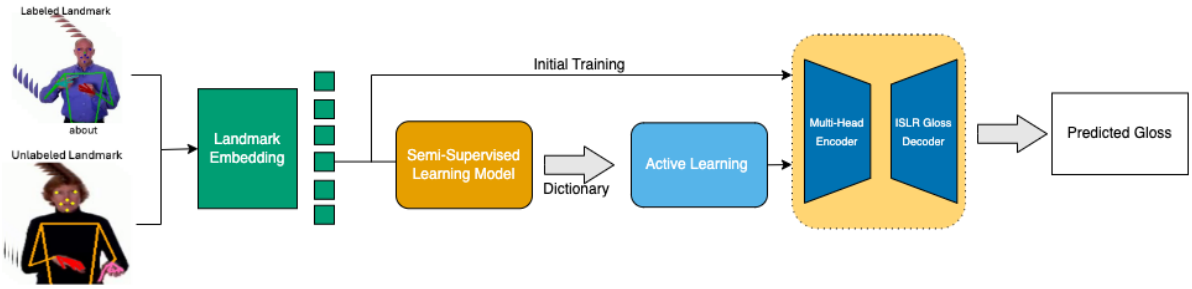

Figure 2. Architecture of the proposed SSLAL model for isolated word sign language recognition[41]. The model begins with labeled and unlabeled landmark data, which are processed through a Landmark Embedding module to capture spatial features. These embeddings are then utilized by a SSL module for initial training, incorporating both labeled and unlabeled data to enhance generalization. An AL component iteratively selects informative samples, constructing a dictionary that further refines the model.

among the selected samples. This combined approach maximizes the informativeness of labeled data by selecting high-uncertainty samples that are also diverse, thereby covering a broad representation of the feature space. To achieve this, we first perform **Principal Component Analysis (PCA)** on the high-uncertainty samples, reducing the dimensionality and capturing the main variance within the data. We then perform **K-Means clustering** on the PCA-reduced samples, grouping them into k clusters, each represented by a centroid. This clustering objective is expressed as follows:

$$\min_{\mathcal{C}} \sum_{i=1}^{n} \min_{\mu_j \in \mathcal{C}} \|x_i - \mu_j\|^2 \tag{12}$$

x_i represents a sample in the PCA-reduced feature space, μ_j is the centroid of cluster j, and $\mathcal{C}$ is the set of all centroids. This clustering step minimizes the Euclidean distance between each sample and its assigned centroid, ensuring that each cluster accurately represents a distinct region of the data. After clustering, we apply Core-Set Selection, which iteratively selects points that are farthest from those already chosen, maximizing the minimum distance between selected samples [29].

5. Experiments

5.1. Setting

5.1.1. Datasets Evaluation

As summarized in Table 1, the large-scale isolated sign language datasets in this study show notable differences in size, signer diversity, vocabulary, and temporal resolution. GISLR [14] has the highest number of samples and the largest average samples per class, 377.91, indicating a well-balanced dataset that captures variations in how each sign is performed. ASL-CITIZEN[11] provides the largest vocabulary, covering 2,731 distinct signs, which supports a broad range of sign coverage for model training. MS-ASL [19] is distinguished by its high signer diversity, incorporating 222 unique signers, and by its extended average frame count per sample of 93.08 frames, which facilitates robust generalization across individual variations and continuity within the temporal domain.

The accompanying Figure 3 illustrates the 3D data representation utilized in this study to enhance SSLAL's capacity for capturing critical spatial and depth-related features. The first column displays raw video frames, highlighting examples where visually similar signs, such as "DAD" and "GRANDPA," exhibit nearly identical 2D hand shapes and positions (shown with landmarks in the second column). This similarity in the planar view poses a challenge for distinguishing signs based solely on x and y coordinates. The third column presents depth maps, where variations in hand distance relative to the body differentiate these classes. This visual disparity in depth emphasizes the importance of incorporating depth information for accurate sign recognition, capturing spatial nuances that 2D landmarks alone may overlook.

5.1.2. Evaluation Metrics

The model's performance is evaluated through **Training Accuracy**, **Test F1 Score**, and **Word Error Rate (WER)**, metrics essential for ISLR [2, 3]. Training Accuracy assesses the proportion of correctly classified samples within the training set, providing insights into the model's capacity to generalize learned representations. The Test F1 Score, the primary metric for evaluating unseen data, combines **Precision** and **Recall** to account for both accurate classifications and the balance of false positives and negatives, thus offering a holistic view of performance [17]. The **Word Error Rate (WER)**, calculated as:

$$\text{WER} = (1 - \text{ACC}) \times 100\% \tag{13}$$

quantifies the misclassification rate, where lower values reflect higher accuracy and robustness across diverse ISLR scenarios. Together, these metrics provide a comprehensive evaluation of the model's applicability and reliability in real-world sign language recognition tasks.

Dataset	Avg Samples/Class	Classes	Signers	Total Samples	Avg Frames/Sample
ASL-CITIZEN [11]	30.54	2,731	52	83,399	85.24
MS-ASL [19]	25.51	1,000	222	25,513	93.08
WLASL [26]	10.54	2,000	119	21,083	64.49
GISLR [14]	377.91	250	21	94,478	67.36

Table 1. Comparative Analysis of ASL Datasets: ASL-CITIZEN[11], MS-ASL[19], WLASL [26], and GISLR [14]. The table details average samples per sign, number of distinct signs, number of signers, total samples, and average frames per sample.

5.1.3. Implementation Details

Our isolated word sign language recognition approach utilizes a **Transformer-based model** in conjunction with a **SSL component** to optimize performance under limited labeled data conditions [6, 8]. The Transformer model integrates AL to prioritize informative data points, while input features, including landmarks for lips, left hand, and pose, are normalized by mean and standard deviation, then embedded through dense layers with GELU activations. Positional embeddings based on frame indices are applied to capture temporal dependencies, and a masking mechanism during random frame selection further enhances robustness [29]. Both the Transformer and SSL models are trained with the AdamW optimizer, using a learning rate of 1e-3 and a weight decay of 1e-5. The SSL model employs a batch size of 16, while the Transformer model uses a batch size of 32. For loss computation, sparse categorical cross-entropy with label smoothing is applied [1]. All training is conducted on an NVIDIA A100 80GB GPU.

5.2. Main Result

Figure 4 shows a comparison of the training WER of the AL model and the Baseline model over the training steps, shown as a moving average. Both models share initial weights from the pre-trained model, which explains why the accuracy does not start from zero. In this case, sign language datasets often lack consistent quality, which affects the ability of tools such as Mediapipe to reliably detect hand landmarks - especially in frames with low clarity. To address this issue, we extracted three key frames (the start, middle, and end of the sign gesture) after removing frames that lack visible hand landmarks. This frame selection not only mitigates the impact of the continuity problem inherent in ISLR, but also helps reduce noise in low-quality frames. In this case, we found that the error rate of the SSLAL model is lower than that of the baseline model. This improvement suggests that the AL model benefits from selectively sampling informative data, ultimately building a more powerful model.

The experimental results, presented in Table 2, illustrate the comparative performance of the Baseline and SSL-AL models, both with and without the incorporation of depth values, across several sign language datasets. Notably, the

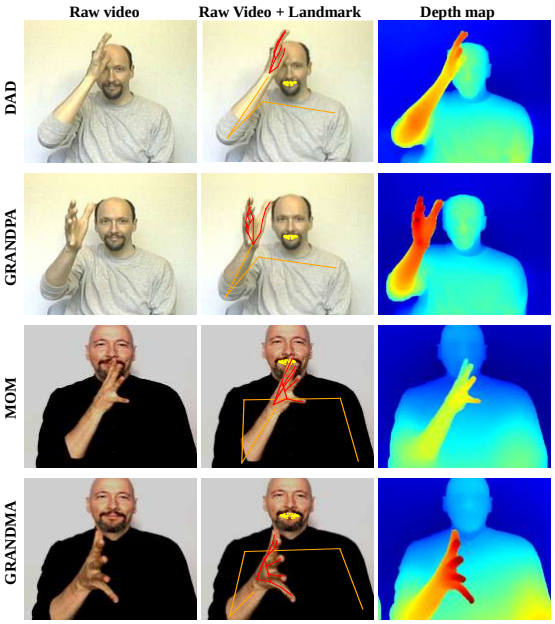

Figure 3. Visualization of multi-modal data for isolated sign language recognition, with fewer lip landmarks shown for clarity.

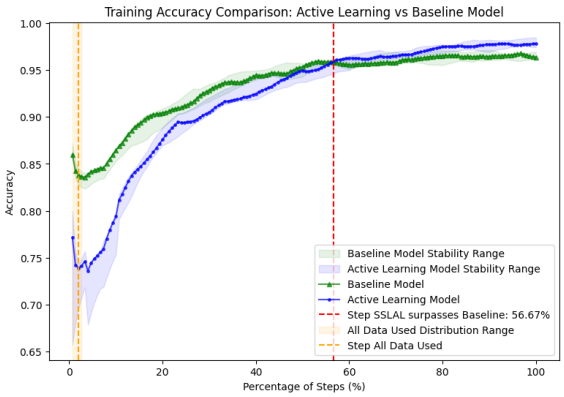

Figure 4. Active Learning model surpasses Baseline model accuracy and step of first all data used.

GISLR dataset lacks original video data, preventing the extraction of depth information; thus, only results without

	Baseline			SSLAL			Baseline 3D			SSLAL 3D		
Method	Dev(WER)	Test(WER)	F1	Dev(WER)	Test(WER)	F1	Dev(WER)	Test(WER)	F1	Dev(WER)	Test(WER)	F1
GISLR [14]	2.960 ± 0.82	2.926 ± 1.19	0.967	1.500 ± 0.46	0.975 ± 0.96	0.971	NA	NA	NA	NA	NA	NA
WLASL [26]	3.020 ± 0.16	3.135 ± 0.22	0.968	1.560 ± 0.02	1.254 ± 0.57	0.986	1.720 ± 0.04	2.146 ± 0.53	0.971	0.850 ± 0.14	1.238 ± 0.44	0.992
MS-ASL [19]	3.960 ± 0.88	2.378 ± 0.52	0.973	2.580 ± 0.74	1.458 ± 0.79	0.984	2.840 ± 1.72	2.192 ± 0.74	0.974	1.680 ± 1.12	0.658 ± 0.30	0.988
ASL-CITIZEN [11]	2.820 ± 0.14	7.823 ± 2.72	0.900	1.980 ± 1.84	4.824 ± 0.66	0.939	0.986 ± 0.08	7.396 ± 0.95	0.914	0.982 ± 0.20	4.007 ± 1.34	0.955

Table 2. Performance WER comparison across multiple datasets in terms of accuracy and F1 score. Results are presented for both the baseline and SSL-AL models, "3D" along with their corresponding versions incorporating depth information.

Method	10K	20K	30K
With SSL & AL	0.976	0.985	0.992
With AL	0.956	0.984	0.985
Baseline	0.952	0.981	0.983

Table 3. F1 scores for models with different configurations (With SSLAL, With AL, and Baseline) across varying training data sizes (10K, 20K, and 30K samples).

Method	GISLR[14]	WLASL[26]	MSASL[19]	ASL-CITIZEN[11]
Semi-SL	2.49 ± 1.07	0.50 ± 0.33	0.78 ± 0.20	1.28 ± 0.41
Self-SL	3.36 ± 1.19	1.85 ± 0.78	1.40 ± 0.22	2.25 ± 0.89
Baseline	4.23 ± 0.86	2.15 ± 0.73	2.38 ± 1.28	4.83 ± 1.53

Table 4. Comparison of WER and variability boundary for Semi-Supervised Learning, Self-Supervised Learning, and Baseline models across various datasets.

depth are reported for GISLR. Across the other datasets, the inclusion of depth values in both the Baseline 3D and SSLAL 3D configurations yields superior performance compared to their non-depth counterparts. This improvement can be attributed to the depth-enhanced data, which simulates the perspective of human vision by capturing the three-dimensional spatial relationships essential in sign language recognition. The variability in results not only indicates model stability but also serves as feedback on data quality. Depth information allows the model to discern the distance of hand and body landmarks from the camera, providing a more comprehensive spatial understanding of the gestures.

Furthermore, each model's variability values were determined by calculating results over five different random seeds, with the median value selected as the primary result and the largest observed difference across seeds used as the variability boundary. This approach ensures a robust estimation of performance variability, underscoring the model's stability and reliability across various initialization scenarios. By incorporating depth, the model can better interpret subtle variations in hand shape, trajectory, and positioning relative to the signer's body, all of which are crucial for accurately distinguishing between similar signs. This added dimensionality helps the model replicate the depth cues that human observers rely on, thereby enhancing recognition accuracy and robustness. Consequently, depth-enhanced models demonstrate notable improvements in both WER reducing and F1 scores, highlighting the importance of depth optimization in capturing the complexity of sign language movements.

Table 3 presents the WER comparisons across varying training data sizes for the SSLAL, AL-only, and Baseline models. As expected, WER reduce across all models with an increase in training data, underscoring the impact of data volume on model performance. The SSLAL model consistently outperforms both the AL-only and Baseline models at each data size, demonstrating the advantage of leveraging both unlabeled data through SSL and informative sample selection via active learning. This combination allows SSLAL to effectively utilize limited labeled data while benefiting from additional unlabeled samples, resulting in a more generalized model. The AL-only model also shows notable improvement over the Baseline, highlighting the value of actively selected samples; however, without the supplementary boost from SSL, its performance remains lower than SSLAL. This trend suggests that SSLAL's hybrid approach not only achieves lower WER in data-limited scenarios but also scales effectively as more training data becomes available, offering robust performance advantages for sign language recognition tasks.

5.3. Ablation Study

In this ablation study, we examine the effects of self-supervised learning (Self-SL) versus semi-supervised learning (SSL) in overcoming the challenges posed by limited annotated data in ISLR tasks. Both methodologies aim to leverage unlabeled data to enhance model performance but employ distinct mechanisms suited to minimal labeled data constraints [35]. Self-SL utilizes unlabeled data through contrastive alignment, allowing the model to derive informative representations independently of labeled supervision [30]. This approach enables the model to uncover intrinsic patterns within the unlabeled data, resulting in diverse feature representations, albeit with less alignment to specific ISLR classes. Conversely, SSL employs labeled data as a stabilizing anchor, facilitating unlabeled data learning through consistency regularization and pseudo-labeling, which significantly expand the utility of sparse labeled samples [31]. By anchoring the model to labeled classes, SSL fosters stable learning with improved class alignment and ensures consistency across class distinctions in the dataset.

Such a comparative analysis is critical for ISLR, where annotated data is resource-intensive to obtain, and robust model performance depends heavily on high-quality labeled data.

Table 4 presents a detailed WER comparison across datasets between models trained with SSL, Self-SL, and the Baseline. The results underscore the effectiveness of SSL, which achieves the highest accuracy across all datasets, demonstrating its superior ability to leverage both labeled and unlabeled data. For example, on the GISLR dataset, Semi-SL achieves a WER of 2.49, outperforming both Self-SL (3.36) and the Baseline model (4.23). This trend holds across other datasets as well, with Semi-SL obtaining a WER of 0.50 on WLASL and 1.28 on ASL-CITIZEN, showcasing its improved generalization capabilities in diverse ISLR settings. While Self-SL also surpasses the Baseline, its performance remains below that of SSL, indicating that access to labeled data provides crucial guidance in semi-supervised settings. The Baseline model's higher WER across all datasets highlights the limitations of fully supervised learning in data-scarce environments, where unlabeled data can be effectively exploited [10, 13]. The variability results, represented as ± 1.07 for GISLR and calculated from the largest observed WER variation across trials, indicate the stability of the training results and provide insights into data quality. Notably, the GISLR dataset, characterized by relatively lower data quality, displays a larger boundary, reflecting increased variability and the challenges of training with noisy ISLR data. These findings emphasize the advantages of SSL in complex SLR tasks, where robust generalization across diverse datasets is essential.

The comparative analysis underscores key insights into the strengths of SSL and Self-SL in SLR. SSL's reliance on labeled data enables it to achieve superior accuracy when annotated samples are available, whereas Self-SL proves valuable in scenarios with limited labeled data, leveraging unlabeled data through self-guided learning. Practically, this suggests that SSL may be preferable in cases where labeled data can be obtained with some effort, while Self-SL is beneficial in low-resource contexts.

6. Conclusion

Despite the promising results of the SSLAL model in ISLR, several challenges suggest directions for further refinement, particularly in managing noise from blurred frames, which can compromise the recognition error rate amid inconsistent video quality. While effective in most cases, the model's depth estimation also struggles with occluded body parts, leading to partial imprecision that impacts gesture accuracy [23]. Moreover, although SSL improves performance by utilizing unlabeled data, it imposes significant computational demands, potentially limiting deployment in resource-constrained environments. Enhanced robustness and efficiency could facilitate ISLR model deployment on mobile and low-resource devices, broadening sign language recognition access. Future work might explore advanced filtering to mitigate noise from blurred frames, thereby improving robustness across varied input quality [32]. Additionally, refining depth prediction with methods like 3D reconstruction or occlusion handling could enrich spatial understanding and gesture precision. Optimizing the SSL framework to reduce computational overhead and adding joint detection for detailed finger movements may further enhance precision in complex hand configurations, essential for SSLAL's adaptability in diverse ISLR applications.

References

[1] S. Agarwal, H. Arora, S. Anand, and C. Arora. Contextual diversity for active learning. In *Computer Vision – ECCV 2020. Lecture Notes in Computer Science*. Springer, Cham, 2020. 2, 6

[2] Muneer Al-Hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsulaiman, Mohammed A. Bencherif, Tareq S. Alrayes, Hassan Mathkour, and Mohamed Amine Mekhtiche. Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation. *IEEE Access*, 8:192527–192542, 2020. 5

[3] Saleh Aly and Walaa Aly. Deeparslr: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition. *IEEE Access*, 8:83199–83212, 2020. 5

[4] George M. Awad. *A framework for sign language recognition using support vector machines and active learning for skin segmentation and boosted temporal sub-units*. PhD thesis, Dublin City University, 2007. 2, 4

[5] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[6] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 182–191, 2022. 6

[7] Danielle Bragg, Oscar Koller, Mary Bellard, and Larwan Berke. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*, 2019. 2

[8] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[9] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey. Asl-lex: A lexical database of american sign language. *Behavior Research Methods*, 49(2):784–801, 2017. 2

[10] H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2568–2574, 2009. 8

[11] Aashaka Desai, Lauren Berger, Fyodor O Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and Danielle Bragg. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *arXiv preprint arXiv:2304.05934*, 2023. 2, 5, 6, 7

[12] R. Elakkiya and K. Selvamani. An active learning framework for human hand sign gestures and handling movement epenthesis using enhanced level building approach. *Procedia Computer Science*, 48:606–611, 2015. 2

[13] M. Gao, Z. Zhang, G. Yu, S. Ö. Arık, L.S. Davis, and T. Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*. Springer, Cham, 2020. 2, 8

[14] Google Research. Google isolated sign language recognition (gislr) dataset. https://www.kaggle.com/competitions/asl-signs, 2021. Accessed: 2024/08/01. 5, 6, 7

[15] Kirsti Grobel and Marcell Assan. Isolated sign language recognition using hidden markov models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, pages 162–167. IEEE, 1997. 2

[16] Saad Hassan, Larwan Berke, Elahe Vahdani, Longlong Jing, Yingli Tian, and Matt Huenerfauth. An isolated-signing rgbd dataset of 100 american sign language signs produced by fluent asl signers. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (SignLang)*. ACL, 2020. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 5

[18] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018. 2

[19] Hamid Reza Vaezi Joze and Yingbo Zhou. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–12, 2019. 2, 5, 6, 7

[20] T. Kapuscinski, M. Oszust, M. Wysocki, and D. Warchol. Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems*, 12(4): 36, 2015. 2

[21] D. Kelly, J. McDonald, and C. Markham. Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(2): 526–541, 2011. 2

[22] O. Koller, N. C. Camgoz, H. Ney, and R. 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, 2020. 2

[23] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005. 8

[24] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[25] 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 Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[26] D. Li, C. Rodriguez, X. Yu, and H. 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 (WACV)*, pages 1459–1469, 2020. 2, 6, 7

[27] F. Lugaresi, L. Tang, J. Brown, M. Hale, C. Camacho, M. Pont-Tuset, A. Blum, K. Suo, R. Castillo, J. Fan, C. Malioutov, S. Steiner, and M. Koss. Mediapipe: A framework for building perception pipelines. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1231–1234, 2019. 3

[28] Hamzah Luqman. An efficient two-stream network for isolated sign language recognition using accumulative video motion. *IEEE Access*, 2022. 2

[29] Maria Papatsimouli, Lazaros Lazaridis, Konstantinos-Filippos Kollias, Ioannis Skordas, and George F. Fragulis. Speak with signs: Active learning platform for greek sign language, english sign language, and their translation. *arXiv preprint arXiv:2012.11981*, 2020. 2, 4, 5, 6

[30] H. Ranganathan, H. Venkateswara, S. Chakraborty, and S. Panchanathan. Deep active learning for image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3934–3938, Beijing, China, 2017. 2, 7

[31] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021. 2, 7

[32] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Word separation in continuous sign language using isolated signs and post-processing. *Expert Systems with Applications*, 249, Part B:123695, 2024. 8

[33] Jeanne Reis, Erin T. Solovey, Jon Henner, Kathleen Johnson, and Robert Hoffmeister. Asl clear: Stem education tools for deaf students. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 441–442. ACM, 2015. 2

[34] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 357–360. ACM, 2007. 2

[35] Razieh Sheikhpour, Mehdi Agha Sarram, Sajjad Gharaghani, and Mohammad Ali Zare Chahooki. A survey on semi-supervised feature selection methods. *Pattern Recognition*, 64:141–158, 2017. 2, 7

[36] Yiqi Tong, Jiangbin Zheng, Hongkang Zhu, Yidong Chen, and Xiaodong Shi. A document-level neural machine translation model with dynamic caching guided by theme-rheme information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4385–4395, 2020. 2

[37] Chong Wu, Jiangbin Zheng, Zhenan Feng, Houwang Zhang, Le Zhang, Jiawang Cao, and Hong Yan. Fuzzy slic: Fuzzy simple linear iterative clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2114–2124, 2020. 2

[38] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 3

[39] Y. Yang, I. Saleemi, and M. Shah. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1635–1648, 2013. 2

[40] Yihao Zhang, Junhao Wen, Xibin Wang, and Zhuo Jiang. Semi-supervised learning combining co-training with active learning. *Expert Systems with Applications*, 41(5):2372–2378, 2014. 2, 4

[41] Q. Zhong, M. Yang, and T. Zhang. Semi-supervised dictionary active learning for pattern classification. In *Pattern Recognition and Computer Vision. PRCV 2018. Lecture Notes in Computer Science*. Springer, Cham, 2018. 2, 5