

SiTeBERT: Pose-Based Transformer with Temporal Downsampling for Isolated Sign Language Recognition

Duy Pham[1,2], Ai Tang[1,2], Nhi Nguyen[1,2], Duc Nguyen[3], and Tho Quan[1,2(✉)]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, Vietnam
qttho@hcmut.edu.vn

[2] Vietnam National University Ho Chi Minh City, Linh Xuan Ward, Ho Chi Minh City, Vietnam

[3] National University of Singapore, Singapore, Singapore

Abstract. Sign Language Recognition (SLR) bridges the communication gap between deaf individuals and the non-signing public, with isolated sign language recognition (ISLR) as a primary focus. Despite recent progress, pose-based ISLR methods still face challenges in scalability, long-range temporal modeling, and label ambiguity when applied to large-scale datasets. In response to these challenges, we propose SiTeBERT, a pose-based framework for ISLR that integrates temporal downsampling, Transformer-based temporal modeling, and confusion-aware learning within a unified architecture. Temporal downsampling is employed to reduce sequence redundancy and computational cost while retaining salient motion patterns, whereas the Transformer encoder is able to model dependencies across long temporal intervals in pose sequences. In addition, a confusion-aware learning strategy is introduced to mitigate label ambiguity and class imbalance without requiring manual relabeling or additional model complexity. Experimental results on two public benchmark datasets indicate the effectiveness of the proposed design choices, achieving a top-1 accuracy of 99.06% on LSA-64 and 36.80% on WLASL-2000. Quantitative evaluations and ablation analyses indicate that each proposed component contributes to improved recognition accuracy, validating the capability of the proposed approach to improve performance in pose-based SLR on large-scale datasets.

Keywords: Isolated Sign Language Recognition · Temporal Downsampling · Pose-based Transformer · Confusion-aware Training

1 Introduction

Social inclusion for the deaf and hard-of-hearing community has been fundamentally bolstered by the development of Sign Language Recognition (SLR) systems,

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2027
H. Fujita et al. (Eds.): IEA/AIE 2026, LNAI 16614, pp. 143–154, 2027.
https://doi.org/10.1007/978-981-92-2885-0_12

which streamline daily interactions. The global prevalence of disabling hearing impairment has reached approximately 430 million individuals, as documented in recent World Health Organization (WHO) estimates. This number is estimated to surpass 700 million by the year 2050 if proactive measures and interventions are not implemented [23]. This motivates the development of precise and adaptable SLR technologies, where Isolated Sign Language Recognition (ISLR) serves as the bedrock by enabling the separate classification of individual gestures. Early research on SLR relied on handcrafted features and classical sequence modeling, which limited the representation of complex spatio-temporal and semantic characteristics [12,13,15,20,26]. Subsequent video-based approaches operating on raw RGB streams enabled end-to-end spatio-temporal modeling; however, they incur high computational costs while remaining sensitive to background and appearance variations [3,10,21,22,25]. To mitigate these limitations, multimodal SLR integrates RGB, pose, and linguistic cues to improve robustness, at the cost of increased system complexity and limited scalability [1,7,9,27,28]. Consequently, pose-based SLR has emerged as a compact and background-invariant alternative; however, existing GCN-based and Transformer-based methods often rely on fixed-length temporal modeling or heuristic sampling, which restrict scalability and may discard informative motion dynamics [2,4,6,8,14,17,18,24,27].

To address these limitations, we propose SiTeBERT (pronounced as 'site-BERT'), a pose-based SLR framework that integrates temporal downsampling, Transformer, and confusion-aware learning. Downsampling reduces redundancy while preserving motion dynamics for more efficient long-sequence modeling. Meanwhile, the Transformer models temporal relationships across extended sequences, and our confusion-aware strategy aims to improve robustness against label ambiguity and class imbalance – common challenges in large-scale datasets like WLASL [11] – without requiring manual relabeling. The primary contributions of this work can be outlined as follows:

- We introduce a temporal downsampling strategy to improve efficiency while preserving informative motion patterns.
- We adopt a confusion-aware learning mechanism to better handle label ambiguity and class confusion.
- We evaluate the proposed design on LSA-64 [19] and WLASL-2000 [11], demonstrating its effectiveness through consistent empirical analysis.

2 Related Work

Early SLR methods relied on handcrafted visual descriptors and classical temporal or discriminative models. Features such as SIFT [26] and HOG [13] provided fixed representations, while sequence modeling techniques including HMMs [20] and DTW [12] captured temporal order under simplified assumptions. Discriminative models such as SVMs [15] achieved reasonable performance on static gestures but remained limited by fixed feature spaces and static decision boundaries.

These constraints hindered the modeling of complex spatio-temporal dynamics and semantic variability in sign language, motivating the adoption of deep learning–based approaches.

2.1 Video-Based Methods

Video-based SLR methods employ deep convolutional models to learn spatial appearance and temporal dynamics from raw RGB videos. Early works evolved from two-stream CNNs [10] to 3D CNN architectures [25] that explicitly capture spatio-temporal correlations, while recent transformer-based models such as VideoMAE [21] further enhance long-range temporal modeling via large-scale pre-training. Hand-centric designs using ResNet backbones, attention mechanisms, and multi-branch architectures have also been explored to handle challenges such as occlusion and small hand regions [3,22]. Despite their effectiveness, video-based methods incur substantial computational and storage costs due to dense RGB processing, motivating the exploration of more compact pose-based representations.

2.2 Multi-modal Methods

Multi-modal SLR methods leverage complementary cues from pose, RGB, and semantic modalities to improve recognition robustness. Representative approaches encode pose keypoints as image-like inputs for spatio-temporal CNNs [1], or fuse pose and RGB by using pose as structural guidance for feature pooling in 3D CNNs [7]. Transformer-based frameworks further model pose–appearance interactions via joint tokenization and masked pre-training [27], while language-assisted methods incorporate gloss semantics to disambiguate visually similar signs [28]. Large-scale systems also combine skeleton-based graph models with RGB, depth, or optical flow streams to capture both structural motion and global appearance [9]. Despite their effectiveness, such multi-modal designs introduce complex fusion pipelines and high computational overhead, limiting scalability. In contrast, pose-based representations offer a compact and background-invariant alternative with reduced system complexity.

2.3 Pose-Based Methods

GCN-Based. Recent studies have increasingly focused on pose-based SLR, as it offers a compact representation that is inherently invariant to background variations. Early methods mainly adopted GCNs to model skeletal spatial structures and local temporal dynamics, often assuming explicit sign boundaries and fixed-length pose sequences enforced via manual alignment or uniform temporal windowing [2,6]. Such designs limit scalability under label ambiguity and class imbalance [6], and reduce temporal flexibility, potentially discarding subtle motion cues. To enhance representation capacity, multi-stream GCNs with attention and residual connections [14] and ST-GCN models [24] have been proposed

to capture local structural motions and short-range temporal dependencies. However, convolution-based GCNs remain limited in modeling long-range temporal dependencies and typically rely on fixed or heuristic temporal sampling, motivating the use of flexible temporal downsampling and Transformer-based temporal modeling for pose-based SLR.

Transformer-Based. Recent pose-based SLR methods increasingly adopt Transformers for temporal modeling. SPOTER [4] processes full-length pose sequences with strong performance but incurs high computational cost on long videos, while other approaches introduce additional inductive biases to improve robustness at the expense of architectural complexity [18]. To reduce sequence length, methods such as SignBERT [8] and BEST [27] employ fixed-length inputs via random sampling or center-frame selection, which may discard important temporal transitions. SignBART [17] instead uses a compact encoder–decoder Transformer to process full sequences. Existing methods either process full-length sequences with high computational cost or adopt heuristic temporal sampling that may discard discriminative motion cues. SiTeBERT addresses these issues by integrating temporal downsampling and confusion-aware learning to handle class imbalance and visually similar glosses.

3 Methodology

3.1 Pose Sequence Preparation

Before being fed into the Transformer encoder, raw pose sequences are preprocessed to obtain normalized and compact skeleton representations. This stage operates on extracted keypoints and serves as the input preparation module of SiTeBERT.

Keypoint Extraction and Selection. We use publicly available OpenHands datasets [16] that provide 2D skeleton sequences for ISLR. Each frame contains joint coordinates (x, y), forming sequences $X \in \mathbb{R}^{T \times J \times 2}$ with T frames and J joints. Although 75 landmarks are available per frame (33 for the body, 21 per hand), we select 48 keypoints, including 6 upper-body joints and 21 joints for each hand.

Normalization. To standardize inputs across signer distance and camera scale, joint coordinates are normalized to $[0, 1]$ using distinct bounding boxes for different regions corresponding to the upper body and each hand. This scaling is performed independently for each frame and each joint group $\mathcal{J} = \mathcal{J}_{body} \cup \mathcal{J}_{left} \cup \mathcal{J}_{right}$ with a small padding $p = 5\% \times \max(w, h)$.

Augmentation. We apply temporal pose-specific augmentations during training: (i) temporal rescaling to simulate variable signing speeds, (ii) frame dropout, (iii) spatial translation, and (iv) Gaussian noise on joint positions. These transformations improve robustness to temporal and spatial variations.

3.2 Pose Embedding and Temporal Downsampling

After preprocessing, normalized pose sequences are transformed into fixed-dimensional embeddings suitable for Transformer-based modeling. This stage bridges raw skeleton inputs and the encoder backbone by converting spatial joint coordinates into temporal token representations.

Feature Flattening. The selected joint coordinates (x, y) from each frame are flattened into a single vector:

$$x_t^{\text{flat}} = [x_{t,j_1}, y_{t,j_1}, \ldots, x_{t,j_J}, y_{t,j_J}] \in \mathbb{R}^{2J}. \tag{1}$$

Linear Embedding. Flattened vectors are projected into the d_{model}-dimensional Transformer latent space via a learnable linear layer:

$$e_t = W x_t^{\text{flat}} + b, \quad e_t \in \mathbb{R}^{d_{\text{model}}}. \tag{2}$$

The full sequence of embeddings is $E = [e_1, \ldots, e_T] \in \mathbb{R}^{T \times d_{\text{model}}}$.

Temporal Downsampling. To reduce computational cost by shortening sequence length while retaining temporal dynamics, a 1D convolution is applied along the temporal axis where the kernel size, stride, and padding are set to 3, 2, and 1, respectively, producing downsampled embeddings with length $T' \approx T/2$. The attention mask is downsampled via max-pooling to ensure the Transformer attends only to valid frames.

3.3 Proposed SiTeBERT Architecture

Overall Pipeline. SiTeBERT is a Transformer-based encoder–classifier for ISLR from skeleton pose sequences. Given pose embeddings (T, d_{model}), the encoder captures long-range temporal dependencies and joint interactions, while a classification head predicts sign probabilities over C classes, as illustrated in Fig. 1.

Transformer Encoder Backbone. The encoder stacks Transformer layers with multi-head self-attention and feed-forward sublayers. Attention is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{3}$$

where Q, K, V are learned projections. Training stability is improved through residual connections, layer normalization, and dropout. The output is contextualized frame features (T', d_{model}) encoding temporal and inter-joint relations.

Temporal Pooling and Classification Head. Mask-aware mean pooling aggregates encoder outputs $H = \{h_1, \ldots, h_{T'}\}$ into a video-level vector:

$$h_{\text{pooled}} = \frac{\sum_{t=1}^{T'} m_t \odot h_t}{\sum_{t=1}^{T'} m_t}, \tag{4}$$

excluding padded frames. The classifier applies dropout and a linear layer to produce logits $(C,)$.

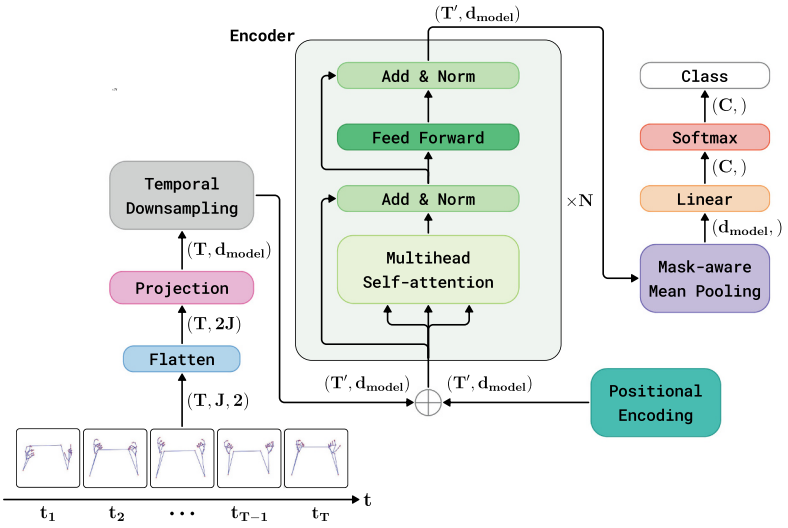


Fig. 1. Overall architecture of SiTeBERT for ISLR.

3.4 Confusion-Aware Training Objective

To improve robustness under large-vocabulary and long-tailed settings, SiTeBERT adopts a two-stage training objective. The model is first trained with label smoothing to learn stable class boundaries, then refined with confusion-aware loss to handle systematic confusions between visually similar signs. This strategy stabilizes early optimization while enabling class-specific supervision in later training. The loss function is defined as:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{LS}, & \text{early training stage,} \\ \mathcal{L}_{conf}, & \text{confusion-aware stage.} \end{cases} \tag{5}$$

Label Smoothing Cross-Entropy. Let C be the number of sign classes, ϵ the smoothing factor, and y the ground-truth class label. The smoothed target distribution $y^{(LS)}$ is defined as:

$$y_c^{(LS)} = \begin{cases} 1-\epsilon, & \text{if } c = y, \\ \frac{\epsilon}{C-1}, & \text{otherwise.} \end{cases} \tag{6}$$

The corresponding label-smoothed cross-entropy loss is given by:

$$\mathcal{L}_{LS} = -\sum_{c=1}^{C} y_c^{(LS)} \log p_c, \tag{7}$$

where $p_c = \text{softmax}(z)_c$ denotes the probability of class c predicted from the model logits z.

Confusion-Aware Loss Weighting. A confusion matrix $\hat{C} \in \mathbb{R}^{C \times C}$ is estimated on the validation set, where $\hat{C}_{i,j}$ represents the probability of samples from class i being predicted as class j.

Based on $\hat{C}$, we construct a weight matrix:

$$W_{i,j} = \max\left(1 - \alpha \cdot \hat{C}_{i,j},\ \lambda\right), \tag{8}$$

where α controls the strength of confusion-based adjustment and λ controls the minimum contribution of each sample to the loss. The confusion-aware loss for a training sample with ground-truth label y and predicted class $\hat{y}$ is defined as:

$$\mathcal{L}_{conf} = W_{y,\hat{y}} \cdot \mathcal{L}_{LS}. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocol. We evaluate SiTeBERT on two benchmark datasets of different scales: LSA-64 and WLASL. Minor preprocessing adjustments are applied to ensure compatibility with the OpenHands releases [16]. LSA-64 [19] contains 64 isolated signs performed by 10 signers under controlled conditions and is used to assess performance in low-data settings. WLASL [11] is a large-scale American Sign Language dataset with 2,000 classes and 119 signers, exhibiting visual variability and long-tailed class distributions. Dataset statistics are summarized in Table 1. Top-1 accuracy is reported for all datasets, while Top-5 and Top-10 accuracies are additionally reported for WLASL.

Table 1. Statistical summary of the evaluation datasets.

Dataset	Classes	Signers	Train	Validation	Test
LSA-64	64	10	2,560	320	320
WLASL	2,000	119	14,289	3,916	2,878

Implementation Details. All experiments are implemented in PyTorch 2.6.0 with Python 3.12 and conducted on a single NVIDIA GTX 1080 Ti GPU. Training is performed for 100 epochs with a fixed random seed of 143, using AdamW with a batch size of 32. The learning rate is adjusted using ReduceLROnPlateau with a decay factor of 0.1 after five epochs without validation improvement. We use a two-stage training strategy with label smoothing for the first 35 epochs, after which confusion-aware loss is activated and updated every 10 epochs. Hyperparameter details are reported in Table 2.

Table 2. Hyperparameter configurations of SiTeBERT.

Hyperparameter	LSA-64	WLASL
d_{model}	128	512
d_{ff}	512	2048
Encoder Layers	2	2
Attention Heads	16	16
Dropout	0.2	0.2
Classifier Dropout	0.4	0.4
Max Sequence Length	256	256
Learning Rate	5×10^{-4}	2×10^{-4}
Epochs	100	100

Table 3. Comparison of model performance on WLASL using Top-1, Top-5, and Top-10 accuracy (%).

Method	Top-1	Top-5	Top-10
RGB-based			
I3D [5]	32.48	57.31	66.31
Pose-based			
GCAR [14]	24.10	52.62	63.32
Bidirectional GCN [6]	**77.43**	**94.54**	–
SignBart [17]	68.00	–	–
SignBERT [8]	39.40	73.35	–
BEST [27]	46.25	79.33	–
Multi-modal			
NLA-SLR [28]	61.26	91.77	–
Fusion-3 [7]	38.84	67.58	**75.71**
SiTeBERT (Ours)	36.80	66.92	75.54

4.2 Quantitative Results

We evaluate SiTeBERT on WLASL (Table 3) and LSA-64 (Table 4), as follows.

SiTeBERT achieves 36.80% Top-1, 66.92% Top-5, and 75.54% Top-10 accuracy on WLASL. While the Top-1 performance is lower than several pose-based baselines, the model maintains competitive Top-5 and Top-10 results. Notably, the Top-10 accuracy is more than twice the Top-1 score, indicating that the model consistently ranks the correct sign within the top candidates. This trend suggests that the Transformer effectively narrows the search space in large-vocabulary settings despite long-tail challenges.

On the LSA-64 dataset, SiTeBERT attains a Top-1 accuracy of 99.06%, placing it competitively among recent approaches in Table 4. The proposed

Table 4. Comparison of model performance on LSA-64 using Top-1 accuracy (%).

Method	Top-1
Pose-based	
3DGCN [2]	94.84
SPOTER [4]	**100.00**
Siformer [18]	99.84
SignBart [17]	96.04
Multi-modal	
R3(2+1)D-SLR [1]	98.90
SiTeBERT (Ours)	99.06

method achieves performance comparable to Transformer-based pose models, while surpassing several existing GCN-based and hybrid baselines. Although slightly behind highly specialized architectures such as SPOTER and Siformer, SiTeBERT demonstrates an effective trade-off between recognition accuracy and model complexity, supporting the practicality of the proposed design.

4.3 Ablation Study

Table 5 summarizes the impact of each component in SiTeBERT. Overall, all modules enhance performance, though their effectiveness is scale-dependent. On WLASL, data augmentation has the largest impact, with a 7.27% drop in Top-1 accuracy when removed, highlighting its importance for generalization under large-vocabulary and in-the-wild conditions. Temporal downsampling is the second most critical factor, causing a 4.21% degradation on WLASL when disabled, as it enables efficient modeling of long and highly variable pose sequences. Joint selection provides consistent but smaller gains by reducing redundant and noisy joints, while confusion-aware weighting yields stable improvements by mitigating systematic confusions. On the smaller LSA-64 dataset, the relative impact of individual components is reduced, although temporal downsampling and confusion-aware weighting remain beneficial. These results indicate that data augmentation and temporal downsampling are the most influential factors for large-scale ISLR, while joint selection and confusion-aware supervision provide complementary refinements that improve robustness across datasets.

4.4 Discussion

SiTeBERT demonstrates the ability to capture spatio-temporal patterns from pose sequences. Results on LSA-64 indicate that joint selection, temporal downsampling, and data augmentation are suitable for short and relatively clean sign sequences, while the more moderate Top-1 accuracy on WLASL, together with improved Top-5 and Top-10 results, reflects the challenges of large vocabularies and long-tailed class distributions. Although SiTeBERT captures long-range

Table 5. Ablation study results (%) on WLASL (Top-1, Top-5, Top-10) and LSA-64 (Top-1).

Variant	WLASL			LSA-64
	Top-1	Top-5	Top-10	Top-1
Without Joint Selection	35.86	**67.44**	**76.30**	98.44
Without Temporal Downsampling	32.59	62.72	71.82	97.19
Without Augmentation	29.53	59.97	70.12	98.44
Without Confusion-Aware	36.48	67.10	75.09	97.81
SiTeBERT (Ours)	**36.80**	66.92	75.54	**99.06**

dependencies, Top-1 accuracy on large-scale benchmarks remains lower than several specialized methods, highlighting current limitations of pose quality and temporal modeling in real-world SLR scenarios.

5 Conclusion

In this paper, we introduce SiTeBERT, a pose-based framework for ISLR. Our architecture integrates temporal downsampling with Transformer-based modeling to capture the intricate dependencies within sign sequences. Joint selection reduces redundant spatial information, temporal downsampling shortens sequence length while preserving salient motion patterns, data augmentation enhances generalization, and confusion-aware learning alleviates label ambiguity and class imbalance. Experiments on two public benchmarks, together with ablation analyses, demonstrate the effectiveness of our architectural components. This work is expected to facilitate future research on pose-based ISLR and large-scale SLR.

Acknowledgements. This research is funded by Ho Chi Minh City University of Technology (HCMUT), VNU-HCM under grant number SVKSTN-2025-KH&KTMT-07. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

1. Akdag, A., Baykan, O.K.: Enhancing signer-independent recognition of isolated sign language through advanced deep learning techniques and feature fusion. Electronics **13**(7) (2024). https://doi.org/10.3390/electronics13071188. https://www.mdpi.com/2079-9292/13/7/1188
2. Al-Hammadi, M., et al.: Spatial attention-based 3D graph convolutional neural network for sign language recognition. Sensors **22**(12) (2022). https://doi.org/10.3390/s22124558. https://www.mdpi.com/1424-8220/22/12/4558

3. Bajpai, S., Dharamshi, N.: Sign language recognition using transfer learning: performance comparison of GoogleNet, ResNet, and VGG. J. Multi Disciplinary Eng. Technol. **18**(01)
4. Boháček, M., Hrúz, M.: 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, pp. 182–191 (2022)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
6. Dafnis, K.M.: Bidirectional skeleton-based isolated sign recognition using graph convolution networks. LREC proceedings (2022). https://par.nsf.gov/biblio/10381424
7. Hosain, A.A., Santhalingam, P.S., Pathak, P., Rangwala, H., Kosecka, J.: Hand pose guided 3D pooling for word-level sign language recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3429–3439 (2021)
8. Hu, H., Zhao, W., Zhou, W., Wang, Y., Li, H.: SignBERT: pre-training of handmodel-aware representation for sign language recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11087–11096 (2021)
9. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3413–3423 (2021)
10. Kim, J.H., Won, C.S.: Action recognition in videos using pre-trained 2D convolutional neural networks. IEEE Access **8**, 60179–60188 (2020). https://doi.org/10.1109/ACCESS.2020.2983427
11. LI, D., Rodriguez, C., Yu, X., LI, H.: 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) (2020)
12. Lichtenauer, J.F., Hendriks, E.A., Reinders, M.J.: Sign language recognition by combining statistical DTW and independent classification. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 2040–2046 (2008). https://doi.org/10.1109/TPAMI.2008.123
13. Liwicki, S., Everingham, M.: Automatic recognition of fingerspelled words in British Sign Language. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 50–57 (2009). https://doi.org/10.1109/CVPRW.2009.5204291
14. Miah, A.S.M., Hasan, M.A.M., Nishimura, S., Shin, J.: Sign language recognition using graph and general deep neural network based on large scale dataset. IEEE Access **12**, 34553–34569 (2024). https://doi.org/10.1109/ACCESS.2024.3372425
15. Nagarajan, S., Subashini, T.: Static hand gesture recognition for sign language alphabets using edge oriented histogram and multi class SVM. Int. J. Comput. Appl., **82**(4) (2013)
16. NC, G., Ladi, M., Negi, S., Selvaraj, P., Kumar, P., Khapra, M.M.: Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022). https://openreview.net/forum?id=zBBmV-i84Go

17. Nguyen, T., Phan Tran, M.K.: SignBart - new approach with the skeleton sequence for isolated sign language recognition. In: Nguyen, N.T., et al.(eds.) Advances in Computational Collective Intelligence, pp. 92–105. Springer Nature Switzerland, Cham (2026)
18. Pu, M., Lim, M.K., Chong, C.Y.: Siformer: feature-isolated transformer for efficient skeleton-based sign language recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 9387–9396. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3681578. https://doi.org/10.1145/3664647.3681578
19. Ronchetti, F., Quiroga, F.M., Estrebou, C., Lanzarini, L., Rosete, A.: LSA64: an argentinian sign language dataset. In: XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016), pp. 794–803 (2016), xIII Workshop Bases de Datos y Minería de Datos (WBDMD)
20. Starner, T., Weaver, J., Pentland, A.: Real-time american sign language recognition using desk and wearable computer based video. IEEE Trans. Pattern Anal. Mach. Intell. **20**(12), 1371–1375 (1998). https://doi.org/10.1109/34.735811
21. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 10078–10093. Curran Associates, Inc. (2022). https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf
22. Wang, S., Wang, K., Yang, T., Li, Y., Fan, D.: Improved 3d-resnet sign language recognition algorithm with enhanced hand features. Sci. Rep. **12**(1), 17812 (2022)
23. World health organization: hearing loss (2024). https://www.who.int/health-topics/hearing-loss. Accessed Jan 28 2026
24. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (2018). https://doi.org/10.1609/aaai.v32i1.12328. https://ojs.aaai.org/index.php/AAAI/article/view/12328
25. Yao, G., Lei, T., Zhong, J.: A review of convolutional-neural-network-based action recognition. Pattern Recog. Lett., **118**, 14–22 (2019). https://doi.org/10.1016/j.patrec.2018.05.018. https://www.sciencedirect.com/science/article/pii/S0167865518302058, cooperative and Social Robots: Understanding Human Activities and Intentions
26. Yasir, F., Prasad, P., Alsadoon, A., Elchouemi, A.: SIFT based approach on bangla sign language recognition. In: 2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA), pp. 35–39 (2015). https://doi.org/10.1109/IWCIA.2015.7449458
27. Zhao, W., Hu, H., Zhou, W., Shi, J., Li, H.: BEST: BERT pre-training for sign language recognition with coupling tokenization. In: Proceedings of the AAAI Conference on Artificial Intelligence vol. 37, no. 3, 3597–3605 (2023)
28. Zuo, R., Wei, F., Mak, B.: Natural language-assisted sign language recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14890–14900 (2023)