BDSL-SPOTER: A Transformer-Based Framework for Bengali Sign Language Recognition with Cultural Adaptation.

Sayad Ibna Azad, Md. Atiqur Rahman
Islamic University of Technology (IUT), Gazipur, Bangladesh.

Abstract

BdSL-SPOTER is a pose-based transformer framework for accurate and efficient recognition of Bengali Sign Language (BdSL).

It integrates cultural-specific preprocessing, a compact four-layer transformer encoder with learnable positional encodings, and curriculum learning for rapid convergence on limited data.

BdSL-SPOTER provides a scalable, low-resource framework for other regional sign languages.

Introduction

Sign language serves as the primary mode of communication for over 70 million deaf individuals globally. In Bangladesh, more than 13.7 million people rely on Bengali Sign Language (BdSL).

However, the lack of automatic BdSL recognition systems continues to limit educational and social inclusion.

Recent deep-learning approaches such as CNN–LSTM or transformer architectures have achieved strong results for Western sign languages but fail to generalize to BdSL because of its distinct signing-space structure and cultural variations.

To address these limitations, we present BdSL-SPOTER, a lightweight pose-based transformer that integrates cultural-specific preprocessing, optimized encoder layers, and curriculum learning for faster convergence on limited data.

The model is trained and evaluated on the BdSL-LW60 dataset, summarized below.

BdSLW60 Dataset Characteristics

Attribute	Value
Total Videos	9,307
Classes	60
Signers	18
Avg Duration	2.3s
Split	70%/15%/15%
FPS	30
Environment	Controlled

Related Works

- **Sensor-based systems**
- **Vision Based**
- **Transformer-based**
- **CNN + LSTM**
- **Sign Pose-based (SPOTER)**

Objectives

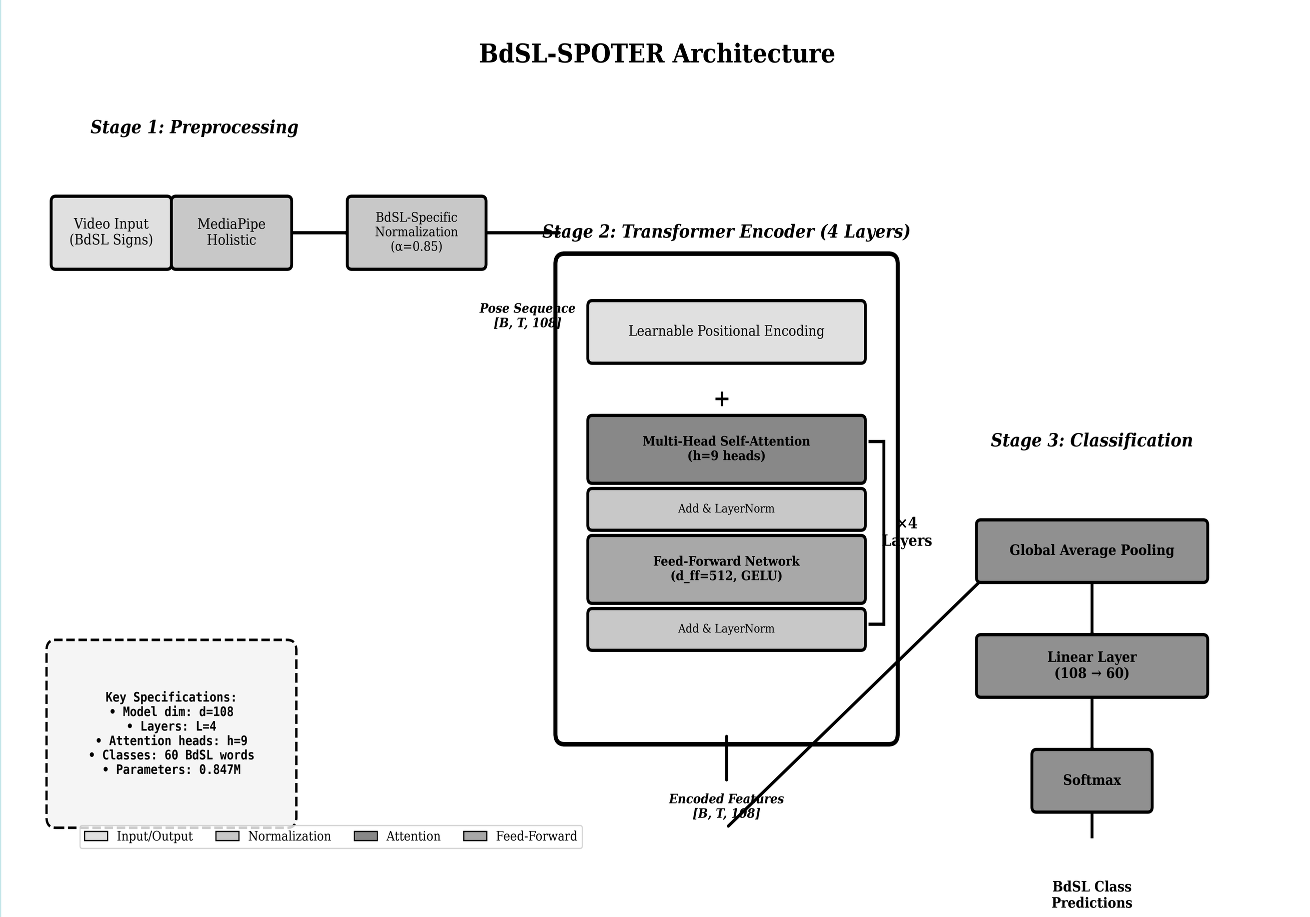
- Develop a culturally aware, pose-based transformer framework for BdSL recognition.
- Optimize model design for efficiency (fewer layers, faster inference).
- Employ curriculum learning for rapid convergence on limited data.
- Deliver a deployable, low-cost solution for accessibility applications

Methodology

Overview:

We extract MediaPipe holistic keypoints, apply signing-space normalization, resample sequences to a fixed length (T = 200), and encode the preprocessed pose sequence P using learnable positional encodings.

A compact 4-layer transformer encoder (9 heads) processes the sequence; the output is temporally pooled and classified by an MLP head. Curriculum learning (sequence-length warm-up) stabilizes training and speeds convergence on limited data.



Pipeline steps:

1. Preprocessing:

Extract KeyPoints:

21/hands &12 upper body
Total=42+12=54

In 2D (x,y)= 108

Data Augmentation:

i. Temporal Dropout.
ii. Coordinate Noise.
iii. Horizontal Flipping

Space Normalization:

$$x'_t = \frac{x_t - x_c}{\alpha \cdot w}, \quad y'_t = \frac{y_t - y_c}{\alpha \cdot h}$$

2. Transformer Encoder:

Positional Encodings:

This design enables the model to adapt dynamically

$$X = P + L_{\text{pos}}$$

Optimized Transformer Architecture:

We use 4 transformer encoder layers.

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

Curriculum Learning:

i. Smoother Optimization
ii. Faster Convergence

3. Classifications:

Signle Head Attention:

$$h_i = \text{softmax}\left(\frac{QW_i^Q (KW_i^K)^T}{\sqrt{d_k}}\right) VW_i^V$$

Multi Head Attention:

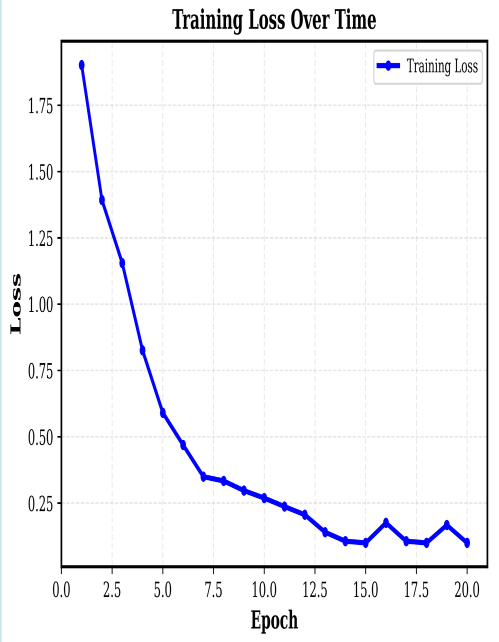
$$\text{MHSA}(X) = \text{concat}(h_1, \ldots, h_9)W^O$$

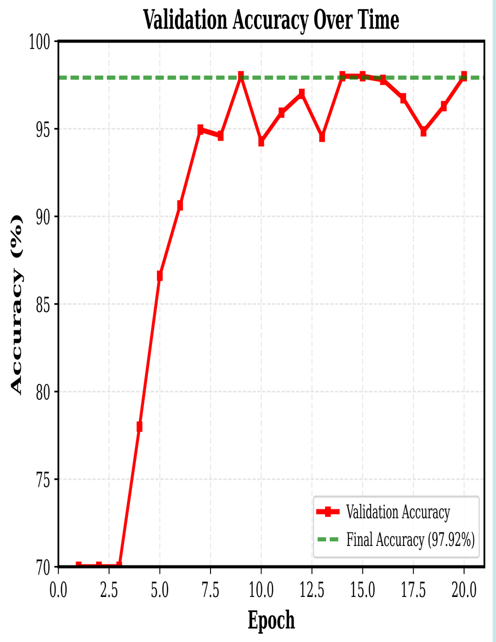
Experiments:

All experiments were conducted using **PyTorch 1.12**.

The model was trained for **20 epochs** with early stopping **(patience = 5)** and OneCycleLR scheduler.

Cross-entropy loss with **label smoothing (ϵ = 0.1)**, **weight decay = 1e−4**, and **dropout = 0.15** were used.





Results:

Our method achieved **97.92% top-1** accuracy on the BdSLW60 dataset, repre senting a **15.52% improvement** over the standard SPOTER and required **63.1% less training time** and **29.4% fewer parameters.**

Detailed comparison with other methods are shown in the table

Performance Comparison on BdSLW60

Method	Top-1 Acc (%)	Top-5 Acc (%)	Macro F1
Bi-LSTM	75.1	89.2	0.742
SPOTER	82.4	94.1	0.801
CNN-LSTM	79.8	91.5	0.785
3D-CNN	77.3	88.9	0.761
BdSL-SPOTER (Ours)	97.92	99.8	0.979

Discussion

Key Findings:

- 97.92% Top-1 and 99.80% Top-5 accuracy on BdSLW60.
- Curriculum learning and label smoothing together add +8.7% accuracy.
- Normalization adds +4.4%, confirming cultural adaptation impact.

Ablation Study: Architecture Optimization

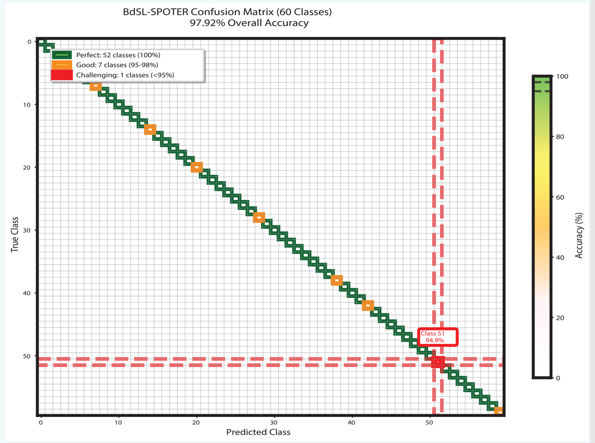
Configuration	Top-1 Acc (%)	Δ Acc (%)
Base Configuration	89.20	–
Layer Analysis		
4 layers (ours)	97.92	+8.72
6 layers	96.80	+7.60
8 layers	95.40	+6.20
Attention Head Analysis		
3 heads	94.30	+5.10
6 heads	96.10	+6.90
9 heads (ours)	97.92	+8.72
12 heads	97.20	+8.00

Error Analysis:

Misclassifications mostly occur among visually similar gesture

Findings:

1. Perfect(100%): 52/60 class.
2. Good (>95%): 7/60 class.
3. Challenging(<=95%): 1/60 class(51th)



Efficiency Analysis:

outstanding computational efficiency across all metrics while maintaining the highest accuracy, as shown in

Statistical Analysis:

- Consistent performance and clear improvement.
- Average accuracy of 97.84%, with a tiny variation(±0.08%).
- A p-value less than 0.001 and Cohen's d of 2.84

Method	Training Time (min)	Parameters (M)	Memory (GB)
Bi-LSTM [14]	45.0	2.1	8.2
SPOTER [2]	13.0	1.2	6.1
CNN-LSTM Hybrid [10]	39.0	3.4	10.5
3D-CNN [4]	52.0	5.2	12.8
BdSL-SPOTER (ours)	4.8	0.847	4.8

Conclusion

We introduced BdSL-SPOTER, a transformer-based model specifically designed for Bengali Sign Language (BdSL) recognition. Our architecture incorporates BdSL-specific preprocessing and cultural adaptations, achieving 97.92% Top-1 accuracy — a 22.82% improvement over baseline models. The model also demonstrates remarkable efficiency, with a 63.1% reduction in training time and 29.4% fewer parameters.

BdSL-SPOTER significantly contributes to improving accessibility for Bangladesh's 13.7 million hearing-impaired people and sets a foundation for inclusive AI solutions that extend beyond BdSL.

Future work:

1. Continuous and multilingual sign recognition.
2. Expanding dataset diversity.

References

1. *Boháček, M., Hrúz, M. Sign Pose-based Transformer for Word-level Sign Language Recognition. WACV, 2022.*
2. *Rubaiyeat, H.A. et al. BdSLW60: A Word-level Bangla Sign Language Dataset. Multimedia Tools and Applications, 2025.*
3. *Li, D., Rodriguez, C., Yu, X., Li, H. Word-level Deep Sign Language Recognition from Video. WACV, 2020.*
4. *Dosovitskiy, A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929, 2020.*
5. *Vaswani, A. et al. Attention Is All You Need. NeurIPS, 2017.*
6. *Hasib, A. et al. BdSL49: A Comprehensive Dataset of Bangla Sign Language. Data in Brief, 2023.*
7. *Hadiuzzaman, M. et al. Baust Lipi: A BdSL Dataset with Deep Learning-based Recognition. ICCA, 2024.*
8. *Ito, T. et al. Learning Interpretable Positional Encodings in Transformers. ICML, 2025.*