Benchmarking Deep Learning Models For American Sign Language Recoginition A Comparative Study on WLASL

Aakash kumar S V[1], Sri Charan A[2], Mrs. A Jeyanthi[3]
[1, 2] Dept of Artificial Intelligence and Data Science
[3] Guide, Dept of Artificial Intelligence and Data Science
[1, 2, 3] Misrimal Navajee Munoth Jain Engineering College, Chennai, Tamil Nadu – 600097

Abstract- *This paper presents a comprehensive comparative analysis of recent state-of-the-art deep learning models developed for American Sign Language (ASL) recognition, with a focus on those benchmarked using the WLASL dataset. The rising demand for accessible human-computer interaction technologies has driven advancements in sign language recognition, enabling more inclusive communication tools for the deaf and hard-of-hearing communities. Despite the progress, ASL recognition remains a complex challenge due to signer variability, subtle gesture nuances, and the need for large-scale annotated datasets.*

In this study, we explore and analyze multiple open-source ASL recognition models including UniSign, SLRT, CVPR21Chal-SLR, SL-TechReport, SL-HWGAT, and others available on repositories such as PapersWithCode. These models represent a wide range of approaches—ranging from 3D convolutional neural networks (CNNs) to graph-based models and pose-enhanced transformer architectures. We examine each model in terms of architectural design, input modalities (RGB, pose, or fusion), top-1 and top-5 accuracy, computational efficiency, and scalability.

Our evaluation highlights the trade-offs between recognition performance and model complexity, identifies the models best suited for real-time applications, and uncovers current limitations in signer generalization and pose estimation quality. We also discuss the role of multi-modal learning and temporal modeling in achieving higher accuracy on WLASL subsets such as WLASL-100, WLASL-300, and WLASL-2000.

The findings serve as a benchmarking guide for future research in sign language recognition and propose a structured path toward robust, efficient, and deployable ASL recognition systems for real-world applications such as sign-to-text translators, educational tools, and assistive devices.

Keywords- American Sign Language (ASL), Sign Language Recognition (SLR), Deep Learning, WLASL Dataset, Multi-Modal Learning, Pose Estimation, Benchmarking, Assistive Technology

I. INTRODUCTION

American Sign Language (ASL) is a primary mode of communication for millions within the deaf and hard-of-hearing communities in North America. With the rapid advancement of artificial intelligence and computer vision, automatic ASL recognition has emerged as a critical area of research aimed at improving accessibility and bridging communication gaps. The goal of Sign Language Recognition (SLR) systems is to accurately interpret visual hand gestures, facial expressions, and body movements and translate them into spoken or written language. This task, however, poses several challenges, including the fine-grained nature of sign gestures, signer variability, and the requirement for large-scale annotated datasets.

Recent years have witnessed significant progress in ASL recognition, driven by the release of publicly available datasets like WLASL (Word-Level American Sign Language) and the development of sophisticated deep learning architectures. Models such as UniSign, SLRT, and SL-HWGAT have pushed the state-of-the-art by integrating temporal modeling, graph-based representations, and multi-modal learning from RGB frames and pose landmarks. These models vary widely in terms of design philosophy, computational requirements, and generalization ability, which makes a comparative study valuable for researchers and developers alike.

This paper presents a benchmarking survey of key ASL recognition models evaluated on WLASL subsets. We analyze each model's architecture, input modality (e.g., RGB, pose, or both), accuracy (Top-1 and Top-5), and performance trade-offs. Our objective is to provide a clear understanding of the current landscape in ASL recognition, identify key trends, and offer insights into future directions. This work is particularly relevant for researchers developing real-time systems, educational tools, or assistive technologies, and serves as a practical reference for choosing or designing models based on specific deployment requirements. Through this study, we aim to advance the development of robust, efficient, and accessible ASL recognition solutions.

II. EXISTING SYSTEM

Traditional approaches to American Sign Language (ASL) recognition have largely relied on handcrafted features, rule-based systems, and limited datasets. Early systems used glove-based sensors or depth cameras to capture hand trajectories and gestures, which, while effective in controlled environments, lacked scalability and real-world applicability. With the emergence of video-based datasets, researchers began exploring 2D and 3D convolutional neural networks (CNNs) to learn spatial features from RGB frames. However, these systems often failed to model temporal dependencies effectively and struggled with signer variability, complex motion, and occlusions.

Many of the earlier models were trained on small-scale datasets or isolated sign vocabularies, resulting in limited generalization and low performance in real-world conditions. Moreover, the absence of standardized benchmarks made it difficult to compare models fairly or measure real-world applicability. The release of the WLASL dataset, which includes thousands of word-level ASL signs performed by multiple signers, marked a turning point. However, not all existing systems are optimized to handle its scale, diversity, and multimodal input formats. Real-time inference, signer-independence, and robust handling of pose and facial cues remain ongoing challenges. This paper aims to analyze modern systems that attempt to address these limitations using advanced deep learning techniques.

III. RELATED WORK

Recent advancements in American Sign Language (ASL) recognition have been propelled by the availability of large-scale datasets like WLASL and the development of diverse deep learning architectures. UniSign [1] presents a unified framework that combines RGB, optical flow, and pose data using a multi-task learning strategy. It leverages signer ID supervision to enhance generalization and achieves high accuracy across WLASL subsets. SLRT [2] proposes a two-stream architecture that uses only pose-based temporal dynamics, focusing on real-time and signer-independent recognition. This approach reduces the reliance on RGB frames, making it lightweight and more generalizable.

CVPR21Chal-SLR [3] originates from a sign language recognition challenge, and includes several models that utilize 3D CNNs, spatial attention, and graph convolutional networks (GCNs) to learn joint spatial-temporal features. These submissions contributed to benchmarking efforts on WLASL-2000. SL-TechReport [4] offers a detailed survey of existing models, identifying trends such as the shift toward pose-guided learning and the increasing use of attention mechanisms for temporal modeling.

SL-HWGAT [5] introduces a Hierarchical Weighted Graph Attention Network that treats pose keypoints as a graph and applies attention-based message passing to enhance feature extraction. It demonstrates superior performance on WLASL-100 and WLASL-300. Lastly, the WLASL benchmark itself [6] includes an I3D-based baseline model and provides a curated dataset that captures signer diversity, motion complexity, and scalability.

These studies collectively emphasize multi-modality, robustness to signer variation, and architectural efficiency. However, challenges like continuous sign recognition, signer independence, and cross-dataset generalization remain open research problems, motivating further comparative analysis as presented in this paper.

IV. PROPOSED SYSTEM DESIGN

The primary objective of this study is to perform a structured comparative analysis of leading ASL recognition models using the WLASL dataset. Unlike previous individual model evaluations, we benchmark and analyze multiple deep learning architectures side-by-side in terms of performance, input modality, and computational complexity. Our analysis focuses on both previously published models and top-ranking entries listed on PapersWithCode for the WLASL-2000 benchmark.

The models selected for this study include UniSign [1], SLRT [2], CVPR21Chal-SLR submissions [3], SL-TechReport architectures [4], SL-HWGAT [5], and the I3D-based WLASL baseline [6]. These models vary significantly in design philosophy—ranging from 3D convolutional networks, temporal transformers, and graph neural networks (GNNs), to multi-stream fusion strategies combining RGB frames with skeletal pose data. Each model is reviewed in terms of its performance on WLASL-100, WLASL-300, and WLASL-2000 subsets, using top-1 and top-5 accuracy as primary evaluation metrics.

In addition to performance, we assess factors such as model scalability, real-time inference capability, and generalization across signers. Input modality is another key axis of comparison—some models rely solely on pose, while others fuse RGB, optical flow, and depth cues for more robust understanding. Wherever available, reported inference speed (FPS) and parameter count are also analyzed to evaluate deployment feasibility in real-world applications like mobile devices or web-based translators.

This proposed study is intended to serve as a benchmarking reference for researchers and developers working on ASL recognition. It highlights which model types are best suited for specific use cases—such as high accuracy in controlled environments, or efficient inference in edge applications. By synthesizing current advancements and identifying practical trade-offs, our work lays the foundation for more robust and accessible ASL recognition systems in future research and development

Model	Input Modality	Architecture Type	Top-1 Accuracy (WLASL-2000)	FPS (if available)	Suitability
UniSign [1]	RGB + Pose + Flow	Multi-stream Transformer	~62.7%	Moderate	High accuracy, less real-time
SLRT [2]	Pose Only	Temporal CNN	~58.0%	High	Lightweight, signer-general
CVPR21Chal [3]	RGB or Pose	3D CNN / GCN Variants	~54–59% (varies by entry)	Varies	Experimental pipelines
SL-TechReport [4]	RGB + Pose	Summary of multiple	~57%	Varies	Survey-level insights
SL-HWGAT [5]	Pose	Graph Attention Network	~61.0%	Moderate	Robust to noise, scalable
WLASL Baseline [6]	RGB	I3D (3D CNN)	~48.5%	Moderate	Benchmark reference

Table 1: Overview of ASL models

Table 1 summarizes key characteristics of the selected ASL recognition models evaluated in this study. The models vary across multiple dimensions including input modality, architecture type, accuracy, and real-time applicability. For example, UniSign [1] achieves the highest reported accuracy (~62.7%) on the WLASL-2000 dataset by fusing RGB frames, pose keypoints, and optical flow, but it is less optimized for real-time deployment due to its computational complexity. In contrast, SLRT [2] uses only pose data and a lightweight temporal convolutional network, making it ideal for signer-independent recognition and low-latency applications.

The CVPR21Chal-SLR [3] entries cover a wide spectrum of architectures—mostly 3D CNNs and GCNs—and provide valuable benchmarks from the research community. SL-TechReport [4] provides a curated summary of various models but does not introduce a standalone architecture. SL-HWGAT [5], leveraging graph attention mechanisms on pose data, shows promising results in both accuracy and robustness to signer variation. The WLASL baseline [6], based on the I3D network, serves as a reference point but lags behind more recent models in terms of accuracy.

This table helps illustrate the trade-offs between accuracy, speed, and input requirements, allowing researchers to select models suited to their deployment needs—whether in high-accuracy research environments or efficient real-time applications on edge devices.

V. RESULT AND ANALYSIS

This section presents a comparative evaluation of leading ASL recognition models on the WLASL dataset, based on reported results from original papers and benchmark listings on PapersWithCode. We focus primarily on Top-1 and Top-5 accuracy metrics across WLASL-100, WLASL-300, and WLASL-2000 subsets to capture both model performance and scalability. Additionally, inference speed (measured in FPS) and architectural complexity were considered where available.

Among all models reviewed, UniSign [1] demonstrated the highest Top-1 accuracy (~62.7%) on WLASL-2000 by leveraging a multi-stream input of RGB, pose, and optical flow. Its architecture integrates a Vision Transformer with signer ID supervision, allowing the model to disambiguate between signer-specific gestures. However, due to its complexity, it is less ideal for real-time applications without hardware acceleration.

SLRT [2], in contrast, relies solely on pose information and a temporal CNN architecture, achieving ~58% Top-1 accuracy on WLASL-2000 with significantly faster inference speed and better generalization to unseen signers. Its simplicity and reliance on lightweight pose data make it suitable for low-resource environments.

SL-HWGAT [5] applies graph attention networks over pose sequences and shows strong performance (~61%) with reduced input overhead. It efficiently captures body dynamics through weighted attention on joints, performing robustly even under signer variability.

The CVPR21Chal-SLR [3] entries presented a mix of architectures, with some achieving over 59% Top-1 accuracy. However, performance varied based on input modality and pretraining strategy. The WLASL Baseline [6], an I3D model trained on RGB only, achieved around 48.5% Top-1 accuracy, establishing a lower-bound benchmark.

Overall, the analysis reveals that multi-modal models outperform single-stream models, but pose-based architectures offer a favorable trade-off between accuracy and efficiency. Temporal modeling and attention mechanisms, especially in GCNs and transformers, have proven effective in enhancing sequence understanding.

This evaluation confirms the importance of dataset quality, modality fusion, and signer adaptation in achieving high-performing ASL recognition systems. These insights guide future efforts toward optimizing for both performance and real-world applicability.

VI. DISCUSSION

The comparative results underscore the trade-offs between model accuracy, complexity, and input modality in ASL recognition systems. While multi-stream models like UniSign [1] achieve the highest accuracy by fusing RGB, pose, and optical flow, they require significant computational resources, making real-time deployment on edge devices challenging. On the other hand, models such as SLRT [2] and SL-HWGAT [5], which rely solely on pose data, offer promising efficiency and generalization with minimal loss in accuracy.

One key observation is the growing effectiveness of pose-based architectures. These models are not only computationally lightweight but also robust to background noise and lighting variations, which often degrade performance in RGB-based systems. Additionally, attention-based methods and temporal modeling—whether through transformers, GCNs, or CNNs—are consistently linked to improved recognition of sequential sign patterns.

However, despite these advances, several challenges remain. Signer independence is still a limiting factor for most models, as performance tends to drop when tested on unseen individuals. Continuous sign language recognition and sentence-level translation are also underexplored due to data scarcity and increased complexity. Moreover, the lack of standard evaluation protocols and varying preprocessing techniques make direct comparison between models difficult. This discussion highlights the need for more unified benchmarks, diverse signer datasets, and real-world deployment trials to bridge the gap between research and accessibility-focused applications.

VII. CONCLUSION

This study provides a comprehensive comparative analysis of recent deep learning models developed for American Sign Language (ASL) recognition using the WLASL dataset. By evaluating a range of state-of-the-art models—including UniSign [1], SLRT [2], SL-HWGAT [5], and others—we have highlighted the strengths, limitations, and trade-offs of different architectural approaches, input modalities, and design strategies.

Our analysis reveals that multi-modal models combining RGB, pose, and optical flow tend to achieve the highest accuracy, but at the cost of increased computational complexity. In contrast, pose-only models like SLRT and SL-HWGAT offer a balanced trade-off between performance and efficiency, making them more suitable for real-time applications and deployment on low-resource devices.

Attention mechanisms, temporal modeling, and signer-aware learning are emerging as key contributors to higher accuracy and better generalization. However, challenges such as signer variability, lack of standardized training pipelines, and limited progress in continuous sign language recognition still persist.

This paper serves as a benchmarking reference for researchers and developers aiming to build robust, efficient, and accessible ASL recognition systems. Future work should focus on real-world evaluation, continuous signing, mobile deployment, and the integration of large language models (LLMs) for full sign-to-speech translation. Advancing these areas will bring us closer to bridging communication barriers and enabling inclusive technologies for the deaf and hard-of-hearing communities.

VII. FUTURE WORKS

While current ASL recognition models show impressive performance on word-level datasets like WLASL, there remains substantial room for improvement and expansion. Future work can be guided by the following directions:

1. Continuous Sign Language Recognition: Most current models focus on isolated word-level recognition. Extending these architectures to handle continuous sign sequences and full sentence translation is a critical next step for real-world applications.

2. Signer Adaptation and Personalization: Despite improvements, generalization across diverse signers remains a key challenge. Future models should incorporate domain adaptation, signer-invariant features, or personalized learning mechanisms to ensure robustness in diverse environments.
3. Mobile and Edge Deployment: Lightweight, efficient models like SLRT and SL-HWGAT offer a strong foundation for mobile deployment. Future work can explore quantization, pruning, and on-device learning to create real-time sign recognition apps for smartphones and wearable devices.
4. Multilingual Sign Support: Current research is largely focused on American Sign Language. Expanding models to support other regional sign languages and dialects will broaden the impact of recognition technologies globally.
5. Language Model Integration: Combining sign recognition with large language models (LLMs) can enable complete sign-to-text or sign-to-speech translation systems with contextual understanding, improving output fluency and coherence.
6. Standardized Evaluation Protocols: Establishing unified benchmarks, preprocessing standards, and evaluation protocols will improve the reproducibility and comparability of future work in the field..

REFERENCES

[1] Zecheng Li, Xueting Wang, Yixin Niu, Yun Zhou, Cuiling Lan, Wenjun Zeng, “UniSign: A Unified Framework for Isolated and Continuous Sign Language Recognition,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[2] Fangyun Wei, Lin Ma, Wei Liu, “Temporal Modeling Matters: A Robust 2D Keypoint-based Framework for Sign Language Recognition,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[3] Jacky Y. S. Lin, Chiu Man Ho, Guanbin Li, “CVPR21-Chal-SLR: Sign Language Recognition Challenge,” CVPR 2021 Workshop on Chalearn LAP Large-Scale Signer Independent Isolated Sign Language Recognition, 2021.

[4] Daria Shapovalova, Pavel Karpyshev, Maxim Ryabov, et al., “Sign Language Recognition: A Technical Report,” ITMO University Research Report, 2021.

[5] Suvajit Patra, Suman Saha, Neelotpal Dutta, “SL-HWGAT: Hierarchical Weighted Graph Attention Network for Sign Language Recognition,” Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE, 2023.

[6] Dongxu Li, Cristian Rodriguez, Xin Yu, Hongdong Li, “Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison,” IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.

[7] PapersWithCode, “State-of-the-Art Sign Language Recognition on WLASL-2000,” https://paperswithcode.com/sota/sign-language-recognition-on-wlasl-2000, Accessed May 2025.