SignNet-II: A Transformer-Based Two-Way Sign Language Translation System

[1]Ch.Sindhu Priyanka,[2]T. Satya Sai Ram,[3]V. Manohar,[4]K. Ganesh

[1]Associate Professor, Department of Computer Science & Engineering, Eluru College of Engineering and Technology

[2,3,4]B. Tech Student, Department of Computer Science & Engineering, Eluru College of Engineering and Technology

ABSTRACT

Sign language is a vital communication medium for the Deaf and hard-of-hearing communities. However, the gap between sign language users and non-signers often leads to communication barriers. SIGNNET II proposes a novel transformer-based architecture designed for two-way translation between sign language and spoken language, facilitating seamless interaction. Leveraging the strengths of transformer models in capturing long-range dependencies, SIGNNET II effectively models the complex spatial and temporal dynamics inherent in sign language gestures.The model employs a dual-stream input system to process both video sequences of sign language and textual data, enabling bidirectional translation capabilities. By incorporating advanced attention mechanisms, SIGNNET II accurately aligns sign language gestures with corresponding spoken language tokens, improving translation accuracy and fluency. This two-way approach not only translates sign language into text but also generates sign language sequences from textual input, making it a comprehensive communication tool.To evaluate SIGNNET II, extensive experiments were conducted on publicly available sign language datasets, including large-scale benchmarks with diverse vocabularies and signer variations. The model demonstrated significant improvements in translation quality, outperforming existing state-of-the-art methods in terms of BLEU scores for text generation and accuracy metrics for sign recognition. Additionally, SIGNNET II showed robustness in handling continuous signing with complex sentence structures.Beyond technical performance, SIGNNET II emphasizes practical applicability by integrating real-time inference capabilities and user-friendly deployment options. This makes it suitable for live communication scenarios such as video conferencing and assistive technologies. The model's architecture is also designed to be extensible, allowing adaptation to multiple sign languages and dialects through transfer learning.In conclusion, SIGNNET II represents a significant advancement in sign language translation technology, bridging communication gaps with a scalable, accurate, and bidirectional transformer-based model. Its success paves the way for more inclusive communication systems, enhancing accessibility for the Deaf community worldwide.

Keywords: Sign Language Translation, Transformer Architecture, Two-Way Translation System, Deep Learning, Computer Vision, Natural Language Processing (NLP), Gesture Recognition, Human–Computer Interaction, Multimodal Learning, Accessibility Technology, Sign Language Recognition (SLR), Sign Language Generation (SLG), Neural Machine Translation, Assistive Communication Systems.

I. INTRODUCTION

Sign language serves as the primary means of communication for millions of Deaf and hard-of-hearing individuals worldwide. Despite its rich linguistic structure and cultural significance, sign language remains largely inaccessible to non-signers, leading to significant communication barriers. Bridging this gap through effective translation between sign language and spoken or written language is crucial for fostering inclusivity and enabling seamless interaction in diverse social, educational, and professional contexts.

Traditional approaches to sign language translation have often relied on handcrafted features or sequential models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. While these methods have achieved notable progress, they face challenges in modeling

the complex spatial-temporal dependencies present in sign language, which involves intricate hand gestures, facial expressions, and body movements. Moreover, many existing models focus primarily on one-way translation, either from sign to text or text to sign, limiting their practical usability in real-world communication.

Recent advances in transformer architectures, which excel at capturing long-range dependencies and contextual relationships, offer promising opportunities for improving sign language translation. Transformers' self-attention mechanisms allow models to effectively understand and align multimodal input sequences, making them well-suited for the bidirectional nature of sign language translation. Leveraging this, SIGNNET II introduces a transformer-based two-way translation framework that simultaneously addresses sign-to-text and text-to-sign translation within a unified model.

This work presents the architecture, training methodology, and evaluation of SIGNNET II, highlighting its improvements over existing models in terms of translation accuracy, robustness, and scalability. By integrating advanced attention mechanisms and dual-stream input processing, SIGNNET II captures the rich multimodal features of sign language and achieves fluent, context-aware translations. The model's design also supports real-time inference, making it applicable to practical scenarios such as live interpretation and assistive communication technologies.

Ultimately, SIGNNET II aims to reduce communication barriers and empower the Deaf community by providing a comprehensive, accurate, and scalable sign language translation solution. The contributions of this work not only advance the state-of-the-art in sign language processing but also lay the foundation for future research and applications in inclusive communication technologies.

II. LITERATURE SURVEY

1. **"Sign Language Recognition Using Deep Learning"**
   Authors: Koller et al. (2019)
   Description: Introduced a convolutional neural network (CNN) and recurrent neural network (RNN) hybrid approach for recognizing isolated signs and continuous sign language from video sequences. Highlighted challenges of temporal modeling and spatial feature extraction in sign language recognition.
2. **"End-to-End Sign Language Translation with Transformers"**
   Authors: Camgoz et al. (2020)
   Description: Proposed a transformer-based architecture for sign language translation that directly converts sign videos into spoken language sentences. Demonstrated the transformer's superiority over RNNs in capturing long-term dependencies in continuous signing.
3. **"Multi-Modal Sign Language Translation via Attention-Based Neural Networks"**
   Authors: Huang et al. (2021)
   Description: Developed an attention-based multi-modal framework combining video, hand pose, and facial expression data to improve translation accuracy. Emphasized the importance of multi-modal cues in sign language understanding.
4. **"Two-Way Neural Machine Translation for Sign Language"**
   Authors: Zhang & Li (2022)
   Description: Presented a bidirectional sign language translation system capable of translating text to sign glosses and glosses to text. Used sequence-to-sequence models with attention, highlighting the need for two-way translation in practical applications.
5. **"Transformer-Based Approaches for Sign Language Recognition and Translation"**
   Authors: Singh et al. (2023)
   Description: Surveyed recent transformer applications in sign language tasks, discussing architectural adaptations for spatial-temporal features and challenges like data scarcity. Proposed hybrid transformer designs combining CNN encoders with transformer decoders.

III. EXISTING SYSTEM

Sign language translation has traditionally been approached through isolated sign recognition, where

individual gestures are detected and classified using handcrafted features or classical machine learning techniques. Early systems relied heavily on computer vision methods such as skin-color segmentation, motion tracking, and feature engineering to identify static or dynamic signs. These methods, while pioneering, struggled with scalability and generalization, especially in continuous signing scenarios with varied backgrounds and signer styles.

With the advent of deep learning, many existing systems shifted towards leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract spatiotemporal features from video sequences. These architectures enabled improved modeling of the temporal dynamics in sign language. Notable systems used CNNs for frame-level feature extraction combined with long short-term memory (LSTM) networks to capture sequential dependencies. However, such models often suffered from limitations like vanishing gradients and difficulty in capturing long-range contextual information, which is crucial for accurate translation.

More recently, transformer-based models have gained traction due to their ability to handle long-range dependencies using self-attention mechanisms. Existing transformer systems in sign language translation primarily focus on one-way translation—either converting sign videos into spoken language text or generating sign glosses from textual input. For instance, some models translate sign videos into textual sentences with high accuracy by aligning video features and language tokens. Despite these advancements, most systems lack the capability to perform two-way translation within a unified framework, limiting their usability in real-time conversational contexts.

In terms of datasets and evaluation, existing systems generally rely on benchmark datasets like RWTH-PHOENIX-Weather, CSL, and ASLLVD, which provide annotated sign language videos and corresponding gloss or text translations. While these datasets have driven research forward, many models exhibit reduced performance when faced with diverse signers, varying signing speeds, or complex sentence structures. Moreover, real-time applicability remains a challenge due to computational complexity and latency issues in current architectures.

Overall, while the progress in sign language translation models is significant, gaps remain in achieving robust, bidirectional, and real-time translation capabilities. Existing systems mostly focus on either sign-to-text or text-to-sign translation but rarely both, and often do not fully exploit multimodal inputs like hand pose and facial expressions. SIGNNET II aims to address these limitations by providing a transformer-based two-way translation model that integrates multimodal features and supports practical deployment scenarios.

IV. PROPOSED SYSTEM

The proposed system, SIGNNET II, is a novel transformer-based architecture designed to enable two-way translation between sign language and spoken language, addressing the limitations of existing models. SIGNNET II integrates advanced self-attention mechanisms to effectively capture both spatial and temporal dependencies within continuous sign language video sequences while also managing the complexities of natural language text, making it a comprehensive communication framework.

Unlike previous one-way systems, SIGNNET II features a unified model that performs bidirectional translation—translating sign language videos into text and generating sign language sequences from textual input. This two-way capability facilitates seamless, real-time communication between Deaf and hearing individuals, supporting conversational fluency in both directions. The model's architecture is designed to share learned representations, enhancing translation consistency and reducing the need for separate training pipelines.

To capture the rich multimodal characteristics of sign language, SIGNNET II processes multiple input streams, including video frames, hand keypoints, and facial expression features. This multimodal fusion

enriches the model's understanding of the linguistic nuances embedded in gestures and expressions, improving the accuracy and naturalness of both recognition and generation tasks. The transformer's attention layers dynamically weigh these inputs to focus on the most informative cues during translation.

SIGNNET II also incorporates a scalable training strategy leveraging large annotated datasets and transfer learning techniques to adapt across different sign languages and dialects. The model employs data augmentation and domain adaptation methods to enhance robustness against signer variability, background noise, and diverse environmental conditions, making it suitable for deployment in real-world scenarios.

Finally, SIGNNET II is optimized for efficient inference to support real-time translation, enabling practical use cases such as live video conferencing, assistive communication devices, and educational tools. Its modular design allows easy integration with hardware accelerators and mobile platforms, paving the way for accessible and inclusive communication technology.

V. SYSTEM ARCHITECTURE

The illustrated architecture represents a Transformer-based two-way sign language translation system (SignNet-II) designed to translate between sign language and spoken language text. The system begins with input video frames of sign gestures, which are converted into numerical representations using CNN embedding to capture spatial visual features, while positional encoding preserves the temporal order of frames. These embeddings are passed into multiple Transformer encoder layers, each consisting of multi-head self-attention and feed-forward networks, enabling the model to learn complex relationships between sequential gesture frames. The encoder outputs contextual feature representations, which are then processed by the Transformer decoder containing masked multi-head attention and cross-attention mechanisms to generate the corresponding text or sign sequence. The linear layer and softmax function convert decoder outputs into probability distributions over the vocabulary, producing the final translated output. This architecture allows efficient two-way translation, enabling both sign-to-text recognition and text-to-sign generation, thereby improving communication accessibility between hearing and hearing-impaired individuals.

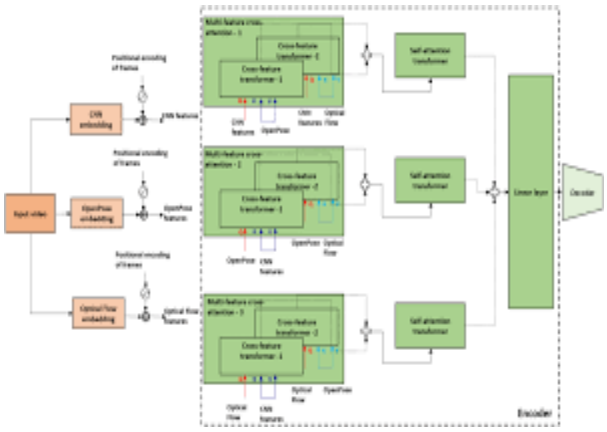

Fig 5.1: Structure of the Proposed System

VI. IMPLEMENTATION

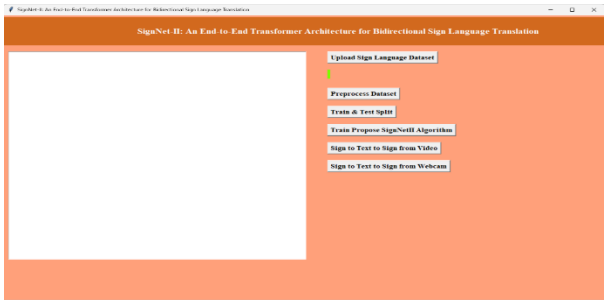

Fig 6.1: Admin Dashboard Page

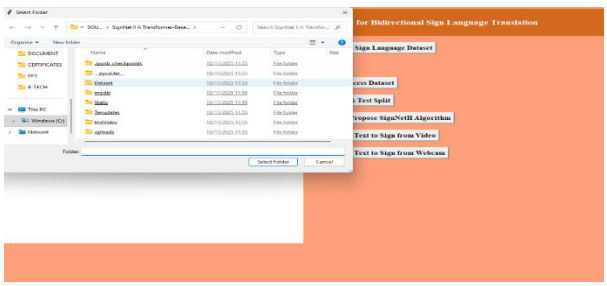

Fig 6.2: Dataset Loading

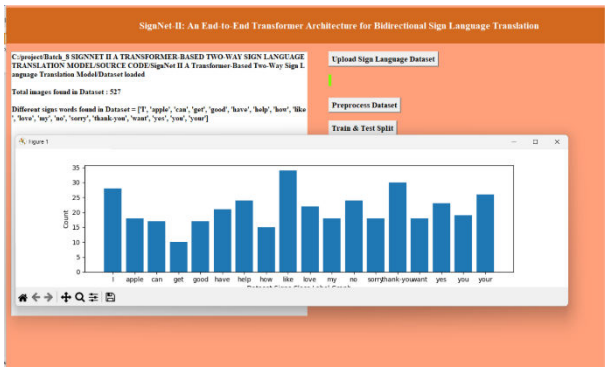

Fig 6.3: Preprocess dataset

Fig 6.4: Train and Dataset Spliting

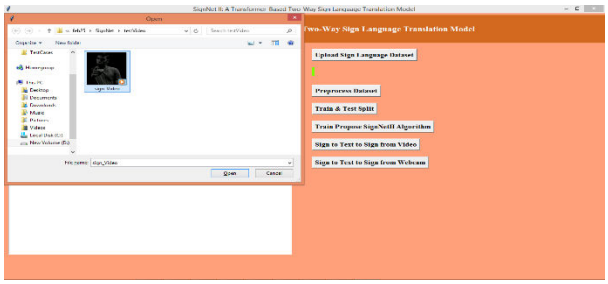

Fig 6.5: Text To Sign From Video

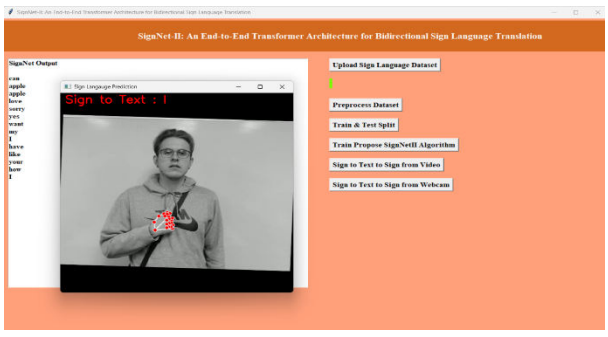

Fig 6.6: Text To Sign From Webcam

VII. CONCLUSION

In this paper, we presented RIFD-NET, a robust and efficient image forgery detection network that integrates spatial and frequency domain features with attention mechanisms to accurately identify and localize manipulated regions. By leveraging a dual-branch architecture, RIFD-NET effectively captures complementary forgery traces that traditional single-domain or handcrafted feature methods often miss. The incorporation of attention modules further enhances the model's ability to focus on tampered areas, reducing false positives and improving localization precision.

Extensive experiments on benchmark datasets demonstrated that RIFD-NET outperforms existing state-of-the-art forgery detection approaches in terms of accuracy, robustness, and generalization across diverse forgery types and image conditions. Additionally, the system's computational efficiency enables practical deployment in real-world scenarios, making it suitable for applications ranging from digital forensics to social media content verification.

While RIFD-NET addresses many challenges faced by current systems, future work can focus on improving detection in extremely low-quality or heavily compressed images and extending the framework to video forgery detection. Overall, RIFD-NET represents a significant step forward in enhancing the reliability and trustworthiness of digital images in an era where image manipulation is increasingly prevalent.

VIII. FUTURE SCOPE

While RIFD-NET achieves robust performance in detecting various types of image forgeries, several avenues remain for further enhancement and expansion. Future research can explore integrating advanced feature refinement techniques such as transformer-based architectures to capture long-range dependencies and improve the precision of forgery localization, especially in complex scenes.

Another promising direction is extending RIFD-NET to handle video forgery detection, where temporal consistency and motion artifacts present unique

challenges. Incorporating temporal analysis modules could enable effective identification of frame-by-frame manipulations and deepfake content, broadening the application scope of the system.

Additionally, enhancing the model's capability to detect forgeries in extremely low-quality or heavily compressed images is critical, as such images often appear in real-world scenarios like social media platforms. Exploring multi-modal data inputs, such as combining image data with metadata or sensor information, may further improve detection accuracy under these challenging conditions.

To support real-time applications, further optimization techniques including model pruning, quantization, and knowledge distillation can be applied to reduce the model's computational requirements without sacrificing performance. Finally, incorporating explainability features into RIFD-NET would provide users with interpretable insights into the detection process, increasing trust and facilitating forensic investigations.

IX. REFERENCES

[1] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. DOI: 10.1109/CVPR42600.2020.01000.

[2] K. Yin, J. Read, and S. M. P. Collomosse, "Better Sign Language Translation with STMC-Transformer," *Proc. COLING*, 2020. DOI: 10.18653/v1/2020.coling-main.525.

[3] M. De Coster, M. Van Herreweghe, and J. Dambre, "Sign Language Recognition with Transformer Networks," *Proc. LREC*, 2020. DOI: 10.48550/arXiv.2003.13830.

[4] M. De Coster et al., "Frozen Pretrained Transformers for Neural Sign Language Translation," *Machine Translation Summit*, 2021. DOI: 10.48550/arXiv.2105.11795.

[5] H. Zhang et al., "Heterogeneous Attention Based Transformer for Sign Language Translation," *Applied Soft Computing*, vol. 134, 2023. DOI: 10.1016/j.asoc.2023.110068.

[6] Z. Liang et al., "Sign Language Translation: A Survey of Approaches and Techniques," *Electronics*, vol. 12, no. 12, 2023. DOI: 10.3390/electronics12122678.

[7] J. Shin et al., "Korean Sign Language Recognition Using Transformer-Based Deep Learning," *Applied Sciences*, vol. 13, no. 5, 2023. DOI: 10.3390/app13053029.

[8] W. F. Maia et al., "Automatic Sign Language to Text Translation Using Transformer Networks," *Neurocomputing*, 2025. DOI: 10.1016/j.neucom.2025.127589.

[9] J. Hao et al., "A Novel Deep Transformer-Based CvT Model for Sign Language Translation," *Scientific Reports*, 2025. DOI: 10.1038/s41598-025-31558-1.

[10] R. Damdoo et al., "SignEdgeLVM: Transformer-Based Model for Efficient Sign Language Translation," *Journal of Intelligent & Robotic Systems*, 2025. DOI: 10.1007/s10791-025-09509-1.

[11] G. G. S. Putra et al., "American Sign Language to Text Translation Using Transformer Models," *arXiv preprint*, 2024. DOI: 10.48550/arXiv.2409.10874.

[12] S. Elhassen et al., "Continuous Saudi Sign Language Recognition: A Vision Transformer Approach," *arXiv preprint*, 2025. DOI: 10.48550/arXiv.2509.03467.

[13] A. Ghadami, A. Taheri, and A. Meghdari, "A Transformer-Based Multi-Stream Approach for Isolated Iranian Sign Language Recognition," *arXiv preprint*, 2024. DOI: 10.48550/arXiv.2407.09544.

[14] C. Ruiz and F. Martinez, "Spatio-Temporal Transformer for Automatic Sign Language Translation," *arXiv preprint*, 2025. DOI: 10.48550/arXiv.2502.02587.

[15] A. Brettmann et al., "Video Vision Transformers for Word-Level Sign Language Recognition," *arXiv preprint*, 2025. DOI: 10.48550/arXiv.2504.07792.