IEEE Open Journal of the Computer Society

Received 9 January 2025; revised 15 April 2025; accepted 16 April 2025. Date of publication 28 April 2025; date of current version 2 June 2025. The review of this article was arranged by Associate Editor Agma Juci Machado Traina.

Digital Object Identifier 10.1109/OJCS.2025.3564828

Optimized Multi-Modal Conformer-Based Framework for Continuous Sign Language Recognition

NEENA ALOYSIUS [1], GEETHA M [2], AND PREMA NEDUNGADI [2]

[1] AmritaCREATE, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala 690525, India
[2] Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala 690525, India

CORRESPONDING AUTHOR: GEETHA M (email: geetham@am.amrita.edu).

This work was supported by the Ministry of Electronics and Information Technology (MeitY), Government of India.

ABSTRACT This study introduces Efficient ConSignformer, a novel framework advancing Continuous Sign Language Recognition (CSLR) by optimizing the Conformer-based CSLR model, ConSignformer. Central to this advancement is the Sign Query Attention (SQA) module, a computationally efficient self-attention mechanism that enhances both performance and scalability, resulting in the Efficient Conformer. Efficient ConSignformer integrates video embeddings from dual-modal CNN pipelines that process heatmaps and RGB videos, along with temporal learning layers tailored for each modality. These embeddings are further refined using the Efficient Conformer for the fused data from two modalities. To improve recognition accuracy, we employ an innovative task-adaptive supervised pretraining strategy for Efficient Conformer on a curated dataset of continuous Indian Sign Language (ISL). This strategy enables the model to effectively capture intricate data relationships during end-to-end training. Experimental results highlight the significant contributions of the SQA module and the pretraining strategy, with our model achieving competitive performance on benchmark datasets, PHOENIX-2014 and PHOENIX-2014 T. Notably, Efficient ConSignformer excels in recognizing longer sign sequences, leveraging a computationally lightweight Conformer backbone.

INDEX TERMS Conformer, ConSignformer, continuous sign language recognition, optimization, supervised pretraining, task-adaptive pretraining.

I. INTRODUCTION

Sign language plays a crucial role as the principal mode of communication for individuals who are hearing-impaired, characterized by its unique grammar, vocabulary, and structure. Key visual tasks in this domain revolve around continuous sign language recognition (CSLR) and sign language translation (SLT). Our previous work on continuous sign recognition resulted in the state-of-the-art ConSignformer [1]. Integrating convolutional layers [2] into Transformer architecture [3] enhances the capability of Conformer to capture both local and global dependencies, resulting in a relatively efficient neural network architecture in terms of size. Despite achieving cutting-edge results with ConSignformer via Conformer adaptation, its design currently lacks suitability for real-time recognition systems. Its primary drawback pertains to computational and memory efficiency. The utilization of the attention mechanism in Conformer, crucial for capturing and preserving long-term information in input sequences, is widely recognized as a computational bottleneck. Consequently, the original Conformer architecture demonstrates slower performance during training and inference phases than other alternative models, posing an engineering challenge for its integration into large-scale CSLR systems. The self-attention mechanism of Conformer often demands substantial memory for sequence processing. Managing extended sequences may result in memory constraints, necessitating the application of techniques such as chunking or attention mechanisms with reduced memory requirements.

This work explains the optimizations performed on the Conformer and ConSignformer models. Optimizations are

done to develop a production-ready sign recognition model capable of deployment on an extensive scale, making optimal use of the exceptional modeling capabilities inherent in the original Conformer architecture. This is achieved by introducing Sign Query Attention (SQA) in the MHA of Conformer with reduced space and time complexity of the attention module. The complex multi-head attention (MHA) is replaced with parameter-efficient SQA, to result in the Efficient Conformer. Also, the ensemble architecture of ConSignformer is simplified by replacing two Conformers with simple temporal learning layers. This design enables harnessing the advantages of multi-cue learning while maintaining architectural simplicity. Temporal convolution layers are added to rgb and heatmap CNNs and the fused CNN features are passed to the Efficient Conformer. In addition to these architectural enhancements, a task-adaptive supervised pretraining step is introduced to strengthen the Conformer. The pretraining methodology strengthened the model with valuable priors, enhancing its competence in addressing various sign language-related tasks. Following the pretraining phase, the ensemble network is fine-tuned with the pre-trained Efficient Conformer for continuous sign recognition. The Efficient Conformer and the dual CNNs with temporal learning layers are collectively called the Efficient ConSignformer.

CSLR poses several inherent challenges, including complex background segmentation, large vocabulary scalability, signer-dependent variations in gestures and handling of transitional movements like Movement Epenthesis (ME) [4] and co-articulation [5]. Additional complexities arise from frequent hand occlusions, integration of manual and non-manual cues and the detection of short or fingerspelled signs. These factors demand models that are not only expressive but also robust and efficient. Motivated by these challenges, Efficient ConSignformer is proposed to effectively combine spatio-temporal cues from multiple modalities to advance the state of CSLR.

In summary, our contributions are as follows:

- Efficient ConSignformer, a computationally light model yet superior in performance compared to the state-of-the-art models.
- Parameter-efficient MHA called Sign Query Attention, resulting in a computationally efficient Conformer called Efficient Conformer.
- Task-adaptive supervised pretraining of the Efficient Conformer, on pose data extracted from continuous ISL videos, with Connectionist Temporal Classification (CTC) loss.
- An Efficient Conformer model pretrained on ISL, that can be employed for a wide range of sign language-related tasks.
- Experimental results demonstrate word error rates (WER) comparable to the state-of-the-art on the benchmark German datasets. Our proposed model outperforms the state-of-the-art on longer video sequences in the test set of the benchmark datasets.

II. RELATED WORKS

A. CONTINUOUS SIGN LANGUAGE RECOGNITION

The study commences with an examination of pertinent literature in sign recognition, spotlighting recent advancements in the domain of Continuous Sign Language Recognition (CSLR) through the lens of Deep Learning [6], [7], [8]. CSLR comprises three pivotal modules: a feature extraction module, a temporal learning module and a module for learning the alignment between input-output sequences. Multi-modal networks, despite their added complexity and susceptibility to noise introduced by fusion techniques, yield the most promising results [9], [10], [11], [12]. The recent Two-Stream model proposed by Chen et al. [9], [10], notable for incorporating knowledge distillation, multiple auxiliary losses, and Spatial Pyramid Networks, is expertly employed to address data scarcity. The recent work by Zuo et al. [13] introduced an innovative online CSLR framework that applies an ISLR model in a sliding-window manner over a sign video stream, opening up a new research direction in this area.

Single-cue models offer a practical balance between complexity and performance. A notable method involves combining a single cue with cross-modal alignment, as demonstrated by CVT-SLR [14], which leverages a Variational Autoencoder for sequence learning. The SMKD method integrates a single cue with a 2D CNN-BiLSTM-CTC-based recognition network, employing a three-stage optimization process to minimize error rates [15]. The latest SignBERT+ model [16] presents a novel pretraining approach for the BERT framework, incorporating hand-prior information to enhance recognition performance.

Iterative training is an additional approach for enhanced learning and mitigating overfitting concerns, initially introduced by Koller et al. [17]. Subsequent refinements of this approach are documented in [18], as well as in recent works such as [19], [20], [21], [22], [23] and [24]. This methodology significantly contributes to improving the feature extractor. Taking this further, the Visual Alignment Constraint (VAC) [25], proposed by Yuecong et al., enhances the feature extractor via alignment supervision. The Masked Iterative Training (MAIT) [26] constitutes a tailored iteration approach that fortifies the visual and sequence learning modules, particularly when BERT is integrated.

B. TRANSFORMER VARIANTS

The Transformer architecture is a highly adaptable encoder-decoder framework with components tailored to various tasks. Depending on the application, one can utilize the encoder, the decoder, or both. Tasks such as sequence labeling, classification and action or gesture recognition often rely on the encoder, with prominent examples including BERT [27], RoBERTa [28], and BigBird [29]. In language modeling, the decoder plays a central role, as demonstrated by models such as GPT [30], GPT-2 [31], and GPT-3 [32]. Encoder and decoder components are crucial in applications like sign

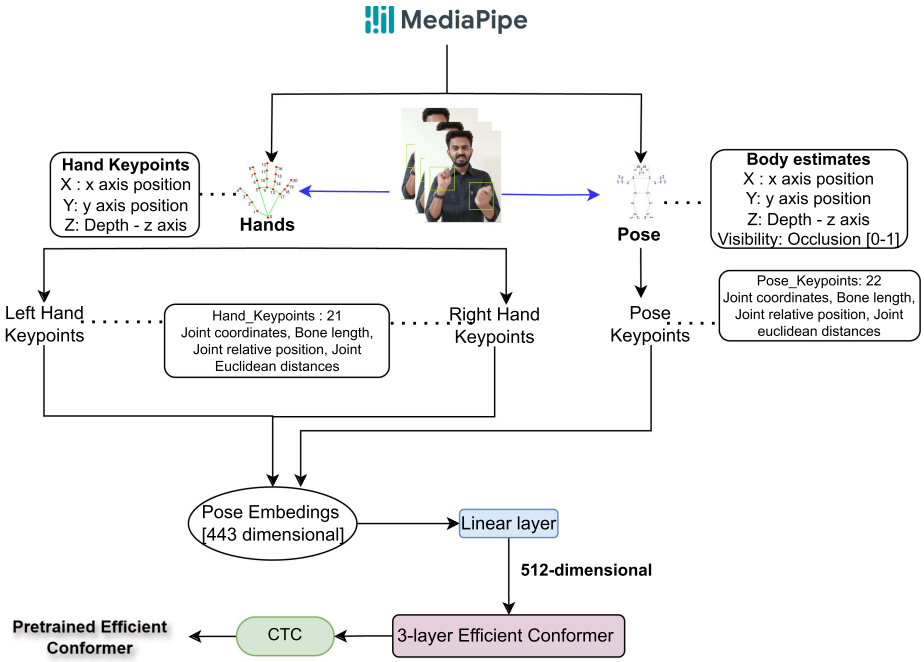

FIGURE 1. Supervised CTC Pretraining of Efficient Conformer. Keypoints are extracted using Mediapipe Hands [37] and Pose [38] from each video frame and are concatenated to form a 443-dimensional feature vector. Since pretraining and downstream tasks are guided by CTC loss, the pretraining technique is also referred to as Task Adaptive Pretraining.

language translation (SLT) or text and speech translations. Commonly used encoder-decoder models include BART [33], T5 [34], and the Switch Transformer [35]. For a detailed examination of Transformer architectures, including their variations, pretraining techniques, and applications, readers can refer to the comprehensive review by Lin et al. [36].

Following the development of ConSignformer, this work focuses on creating a production-ready sign recognition model designed for large-scale deployment while leveraging the exceptional modeling capabilities of the original Conformer architecture. This goal is achieved by implementing an optimized version of the ConSignformer architecture. A novel self-attention module, called Sign Query Attention, is integrated into the Conformer to enhance computational efficiency, resulting in the Efficient Conformer. The optimized variant of the ConSignformer architecture, which incorporates the Efficient Conformer, is referred to as the Efficient ConSignformer.

III. PROPOSED MODEL

A. TASK-ADAPTIVE PRETRAINING

Pretraining helps the model learn generalized representations of temporal and spatial features, making it more adept at capturing dependencies and variations in input gestures. The choice of data and the pretraining task play a vital role in determining the effectiveness of acquiring the valuable priors for the downstream task. In this context, we aimed for a pretrained Efficient Conformer model for CSLR. This initiative was a key component of a major project backed by the Government of India (GI) to promote Indian Sign Language (ISL). We collected a large dataset of around 12,000 videos, all recorded by hard-of hearing signers and annotated by ISL interpreters certified by the GI. The vocabulary size of this pretraining dataset is 164. Utilizing Mediapipe Pose [38], we extracted 22 upper body keypoints for each video frame. The extraction of pose embedding is the same as detailed in our previous work [1].

The keypoints are arranged into a 443-dimensional vector, which is then processed through a linear layer to transform into a 512-dimensional vector, aligning with the hidden dimension of Efficient Conformer. On ISL keypoints data, the model is trained under CTC supervision. The model converged in 47 epochs, with a validation WER of around 20. The pretraining process is outlined in Fig. 1. Since both the pretraining and downstream recognition tasks are trained with CTC loss, this pretraining is called Task-Adaptive Pretraining.

B. EFFICIENT CONSIGNFORMER

The proposed model, namely Efficient ConSignformer, comprises an ensemble of three networks: (i) rgb-CNN + temporal learning layers, (ii) heatmap-CNN + temporal learning layers, and (iii) rgb and heatmap features fused + pretrained Efficient Conformer. Separable 3D CNN or S3D is the CNN model used in the network. The architecture is depicted in Fig. 3. The representations from the first four layers of the S3D are used as the rgb and heatmap features. Heatmaps represent the keypoints of the face, hands, and upper body, which are extracted using HRNet [39] trained on the COCO-WholeBody dataset [40]. The temporal learning layers mainly consist of normalization and ReLU activation applied to temporal

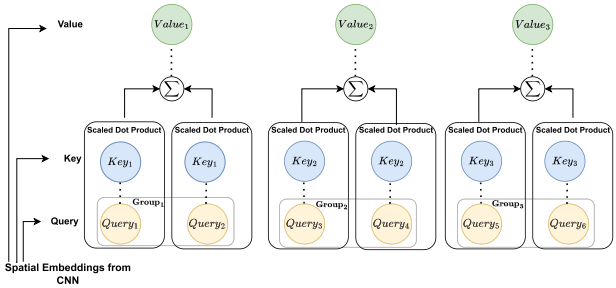

FIGURE 2. Sign Query Attention. An example of 3 clusters is shown in the figure.

convolutions. The number of trainable parameters for the temporal layers is significantly less than that of the Conformer. This combination of temporal layers and Conformer for sequential learning in CSLR makes Efficient ConSignformer parameter-efficient while benefiting from the advantages of the Conformer architecture.

A Conformer block distinguishes itself from a Transformer block through several key features. Importantly, it incorporates a convolution block and adopts a Macron-style architecture [41] with a pair of Feedforward Networks (FFNs) flanking the block. The convolution sub-block stands out as a crucial element, and using a Macron-style FFN pair proves to be more effective than a single FFN with an equivalent parameter count. Furthermore, the Conformer models demonstrate accelerated convergence when swish activations are employed. The original Conformer architecture is complex and computationally heavy. With our goal of an optimized architecture, the multi-head attention (MHA), whose space and time complexity is quadratic to the input length, is replaced by a less complex attention mechanism called Sign Query Attention.

1) SIGN QUERY ATTENTION

Sign Query Attention (SQA) is applied to the self-attention of the Efficient Conformer encoder module.In the SQA module, the query vectors are grouped into clusters based on predefined boundaries over an index that represents spatial or semantic information from the input sequence. This allows queries with similar positional or contextual relevance to be grouped together. Each cluster then attends to a shared key and value representation, significantly reducing redundancy in the attention computation. The rationale behind this grouping is to enforce localized attention, which is particularly effective in sign language sequences where specific hand or body regions convey tightly coupled semantic meaning. This strategy enables the model to balance expressiveness and efficiency, particularly under resource constraints. The sign query heads are clustered inside the self-attention and each cluster shares a single key and value head. An example of 3 clusters (same as the count of key-value heads) is shown in Fig. 2. The complexity reduction comes from performing attention operations on smaller clusters of queries, which is especially useful in scenarios where the sequence length of the gesturing video is large.

If the number of heads in the standard MHA is denoted as H, the dimension of the query, key, and value spaces as d_{model} and the sequence length as L, then in the standard MHA, the time and space complexity is $O(H \cdot d_{model} \cdot L^2)$ because each head independently attends to all positions in the sequence. Now, let us denote the number of clusters as C, query heads as Q, key heads as K and value heads as V where

$$C = K = V << Q \tag{1}$$

in case of SQA and

$$C = K = V = Q \tag{2}$$

for the original MHA.

$$Q \div C = total\ number\ of\ heads\ per\ clustered\ attention \tag{3}$$

In SQA, the time and space complexity decreases to $O(C \cdot d_{model} \cdot L^2)$. This reduction in complexity is particularly beneficial when C is much smaller than H. While attention computations for individual heads are already parallelized in standard MHA, clustering the query vectors provides an additional layer of parallelization, offering computational benefits, especially in scenarios with numerous query heads. This helps in faster decoding in the inference stage. We have also experimented with the complex clustered self-attention reported in Fast Transformers [42] but it did not yield positive results for our task.

2) LOSS FUNCTION

The training process for the Efficient ConSignformer involves supervision through the CTC loss. The RGB encoder produces spatiotemporal features denoted as λ^{rgb}. These features are projected through a linear layer followed by a softmax to estimate frame-wise gloss probabilities $p(f_t|\mathbf{v}^{\text{rgb}})$. To compute the overall probability of a target gloss sequence S, the model marginalizes over all valid alignment paths π belonging to the set $\mathcal{A}$, where each π corresponds to a potential alignment:

$$p\left(S|\lambda^{\text{rgb}}\right) = \sum_{\pi \in \mathcal{A}} p\left(\pi|\lambda^{\text{rgb}}\right) \tag{4}$$

The loss for the RGB branch, denoted as L_{rgb}, is then defined as the negative log-likelihood of the correct gloss sequence S^*:

$$L_{rgb} = -\log p\left(S^*|\lambda^{\text{rgb}}\right) \tag{5}$$

Analogously, the heatmap encoder outputs features λ^{heatmap}, from which the gloss sequence probability is computed. The corresponding loss $L_{heatmap}$ is formulated similarly:

$$L_{heatmap} = -\log p\left(S^*|\lambda^{\text{heatmap}}\right) \tag{6}$$

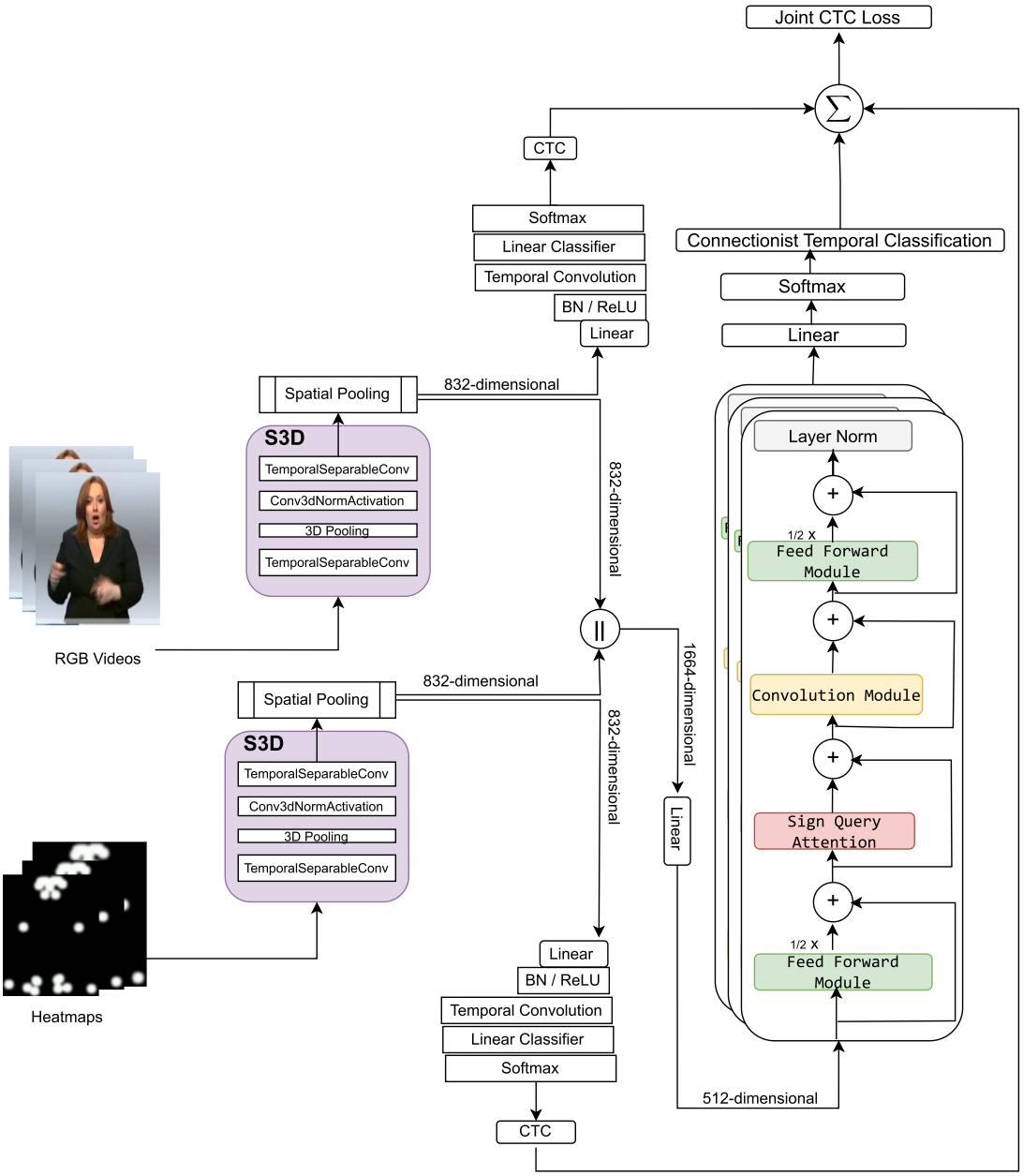

FIGURE 3. The Efficient ConSignformer: RGB Encoder and Heatmap Encoder are composed of S3D-Temporal Convolution layers-CTC. Fusion Network is composed of S3D-Conformer-CTC.

Processing the RGB and Heatmap pipelines separately allows the network to specialize in extracting modality-specific features without interference or dilution. This design ensures that appearance-based features (from RGB) and motion/posture-based features (from heatmaps) are independently optimized and preserved before being effectively fused. Such modular processing enables the system to leverage the strengths of both low-level texture and high-level geometry, improving the robustness and generalization capability of the sign language recognition model.

The combined modality features, denoted by λ^{fusion}, are passed through Conformer layers to estimate frame-wise gloss probabilities $p(f_t|\lambda^{\text{fusion}})$. The probability of generating a target gloss sequence S from these fused representations is calculated by summing over all possible alignment paths $\pi \in \mathcal{A}$, where $\mathcal{A}$ represents the set of valid CTC alignments for S:

$$p\left(S|\lambda^{\text{fusion}}\right) = \sum_{\pi \in \mathcal{A}} p\left(\pi|\lambda^{\text{fusion}}\right) \tag{7}$$

The design of the Fusion network is motivated by the fact that RGB frames provide rich spatial and contextual appearance information, capturing fine-grained visual cues such as hand shape, facial expressions, and scene context. In contrast,

keypoint-based heatmaps abstract the motion dynamics and structural posture of the signer, offering a compact and invariant representation of body movements. By combining these complementary modalities, the network benefits from both detailed appearance features and high-level pose dynamics. The fusion before feeding into the Efficient Conformer enables the model to jointly learn global and local temporal dependencies across modalities, ultimately improving recognition performance, especially for subtle or ambiguous signs that are hard to interpret from a single modality alone.

To train this fusion pathway, a CTC-based loss function is employed. The loss, denoted as $L_{\text{conformer}}$, is the negative log-likelihood of the ground truth sequence S^*:

$$L_{\text{conformer}} = -\log p\left(S^* | \lambda^{\text{fusion}}\right) \tag{8}$$

Similarly, the training objective for the RGB-based APN is defined as the negative log-likelihood of the correct sequence S^*:

$$\mathcal{L}_{\text{aux\_rgb}} = -\log P\left(S^* | \lambda^{\text{aux\_rgb}}\right) \tag{9}$$

where $\lambda^{\text{aux\_rgb}}$ is the spatio-temporal features derived from S3D feature maps of RGB videos, through a series of temporal modules.

Likewise, the loss for the Heatmap-based APN is computed using the features $\lambda^{\text{aux\_heatmap}}$ as:

$$\mathcal{L}_{\text{aux\_heatmap}} = -\log P\left(S^* | \lambda^{\text{aux\_heatmap}}\right) \tag{10}$$

The combined auxiliary loss function is expressed as:

$$\mathcal{L}_{aux} = \omega_1 \mathcal{L}_{\text{aux\_rgb}} + \omega_2 \mathcal{L}_{\text{aux\_heatmap}} \tag{11}$$

where ω_1 and ω_2 denote weights for the auxiliary CTC losses of the RGB and the Heatmap pyramids. Based on empirical analysis, we set $\omega_1 = 0.4$ and $\omega_2 = 0.3$.

The overall loss is derived from both the rgb (L_{rgb}) and heatmap pipelines ($L_{heatmap}$), which are then combined with the Conformer pipeline ($L_{conformer}$). This approach ensures that the model learns effectively from visual and temporal information, leveraging the strengths of each pipeline to enhance its overall performance in sequential learning tasks. Furthermore, we incorporate an auxiliary pyramid network with CTC head as in [1] throughout training, to enhance feature learning across multiple scales and improve the overall training process. It is important to note that this module is not included in the proposed model during the inference stage and, as a result, is not depicted in the architecture diagram. However, the auxiliary CTC loss (L_{aux}) is also added while computing the main loss.

$$L_{final} = L_{rgb} + L_{heatmap} + L_{conformer} + L_{aux}$$

IV. EXPERIMENTS

A. DATASETS AND EVALUATION METRIC

The proposed Efficient ConSignformer underwent evaluation on two demanding datasets of German Sign Language: RWTH-Phoenix-Weather-2014 T (PHOENIX14 T) [21] and RWTH-Phoenix-Weather-2014 (PHOENIX14) [43]. PHOENIX14 T is an enhanced version of the PHOENIX14 corpus consisting of German translations for the video content. It features a slightly smaller vocabulary than PHOENIX14. As a result, although the performance on PHOENIX14 and PHOENIX14 T may share similarities, direct comparisons between them might be nuanced.

The evaluation of automatic sign recognition systems utilizes the Word Error Rate (WER) as a key metric, as outlined in [43]. WER measures the difference between the recognized gloss sequence (hypothesis) and the ground truth sequence (reference). The calculation of WER involves counting substitutions, deletions, and insertions and is expressed by the formula:

$$WER = \frac{\#deletions + \#substitutions + \#insertions}{\#words\ in\ reference}. \tag{12}$$

The resultant WER value is typically presented as a percentage, with lower values indicating higher accuracy.

TABLE 1. Configurations

Hyper-parameter	Value
#Encoders	3
#Query heads	32
#Value heads	2
#Key heads	2
Dropout	0.1
Learning rate	0.001
#Epochs	40
Optimizer	Adam
Weight decay	0.001
Batch size	8
Scheduler	cosineannealing
Kernel size	31
Expansion factor	4

B. IMPLEMENTATION DETAILS

We begin by pretraining the Efficient Conformer in a supervised learning approach on the ISL pose dataset. This involves using 3 Conformer encoder layers, each with a dimensionality of 512, 32 query heads and 2 key-value heads. The pretraining is carried out on a single A100 GPU. Following this, we leverage S3D pretrained on Kinetics-400 and pretrained Efficient Conformer for the final joint training of the proposed model, supervised under CTC. Table 1 provides the hyperparameter settings for the final training. The training is conducted in a distributed manner using 8 A100 GPUs on a DGX server.

C. RESULTS

The outcomes of our models' recognition performance on the two benchmark datasets are outlined in Table 2. Initially, we employed the Efficient Conformer pretrained with supervised task-adaptive CTC training. It is worth highlighting that the model processing raw videos exhibits significantly superior

TABLE 2. Performance Results

	Phoenix2014T		Phoenix2014	
Model	Dev WER	Test WER	Dev WER	Test WER
RGB Pipeline	21.78	21.86	21.99	21.98
Heatmap Pipeline	21.64	21.87	22.15	22.29
Conformer Pipeline	21.08	21.70	21.88	21.90
Efficient ConSignformer	19.56	19.70	20.09	19.94

The first three rows illustrate the results from each pipeline, with the final row showcasing the performance of the proposed network.

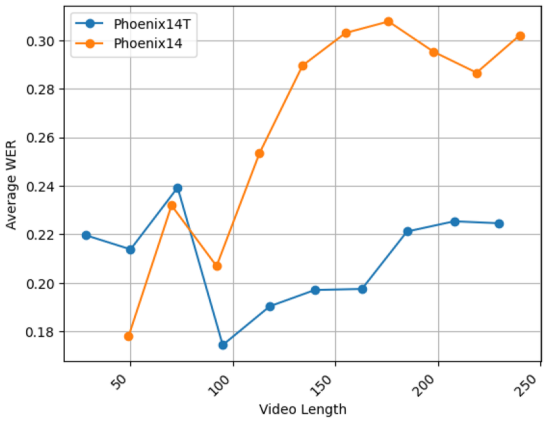

FIGURE 4. Average WER of the proposed model for videos with similar lengths on the benchmark datasets.

performance compared to the one utilizing heatmap inputs. On the other hand, the Efficient Conformer network that processes fused data (RGB + heatmap) outperforms the models relying solely on RGB or heatmap inputs. Moreover, the ensemble configuration of the three networks, known as Efficient ConSignformer, demonstrates superior recognition capabilities compared to the individual networks. Importantly, the proposed model consistently achieves the best results across all experiments.

Fig. 4 illustrates the correlation between the length of sign videos and WER. A notable observation is the increase in WER with longer video lengths. This trend can be attributed to various factors; for instance, longer videos may introduce extended temporal dependencies between words and phrases. This underscores the necessity for an enhancement in the self-attention mechanism of the Conformer to accommodate longer sequences. Additionally, in longer videos, the frequency of certain words may be lower, resulting in sparser data for those instances. This can impact the model's capacity to learn efficiently and generalize, particularly for less common words.

D. ABLATION STUDIES

1) *EFFICIENT CONFORMER VS. TEMPORAL LEARNING LAYER*

We conducted an extensive ablation study on the components of our proposed model, and the resulting WERs for each experiment are documented in Table 3. The experiments involved testing different combinations of the sequence learning module. The Efficient Conformer was excluded in the initial trial (Exp_1) and temporal learning layers were utilized in all three networks. In the subsequent trial (Exp_2), this configuration was reversed.

In the third experiment (Exp_3), the proposed architecture was implemented, except that the Efficient Conformer was not pretrained. Following this, we conducted another trial with a pretrained Efficient Conformer in the same setup (Exp_4).

Using three Efficient Conformers led to an overfitted model while eliminating the Conformer did not yield satisfactory results. This is evident from the third experimental setup, where using a single Efficient Conformer alone resulted in a reduced WER compared to the first two setups.

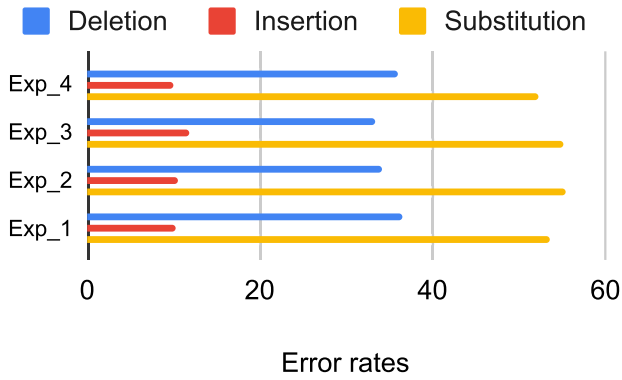

FIGURE 5. Study on the insertion/substitution/deletion error rates, on Phoenix14.

2) IMPACT OF TASK-ADAPTIVE PRETRAINING USING ISL POSE DATA

To assess the contribution of our supervised task-adaptive pretraining strategy, we conducted controlled experiments comparing models initialized with and without pretraining. In our setup, pretraining was performed using ISL pose data, enabling the model to learn gesture-relevant spatiotemporal representations before fine-tuning on the target recognition task.

The effect of pretraining is evident in the comparison between Exp_3 (non-pretrained) and Exp_4 (pretrained), as documented in Table 3. The pretrained Efficient Conformer consistently achieved a lower WER, indicating improved recognition accuracy. This gain underscores the benefit of initializing with task-aligned prior knowledge.

Moreover, pretraining contributed to better generalization, as observed in reduced overfitting and improved validation performance. This suggests that the pose-driven pretraining serves as a strong inductive bias, helping the model adapt more effectively to variations in signer styles and gesture execution across the dataset.

3) *INSERTION, SUBSTITUTION, DELETION RATES*

A study on the insertion/substitution/deletion error rates was also conducted on the four experimental setups. The comparison of error rates on the four experimental setups is graphically shown in Fig. 5. The deletion rate is minimal

TABLE 3. Ablation Study on Phoenix14

Model	Sequence Learning Module			
	Exp_1	**Exp_2**	**Exp_3**	**Exp_4**
RGB Pipeline	Temporal Layers	Conformer	Temporal Layers	Temporal Layers
Heatmap Pipeline	Temporal Layers	Conformer	Temporal Layers	Temporal Layers
Conformer Pipeline	Temporal Layers	Conformer	Conformer without pretraining	Pretrained Conformer
WER	20.53	21.62	20.48	19.94

Various combinations of sequence learning modules are experimented.

TABLE 4. Study on the Effect of the Number of Efficient Conformer Encoders on Phoenix14 T

#Encoders	Dev WER	Test WER
3	19.56	19.70
5	20.22	19.22
7	20.71	19.54

for Exp_3 whereas insertion and substitution rates are minimal for the proposed model or Exp_4. The consistent and controlled nature of low insertion errors of Efficient ConSignformer is a promising sign. There has been a slight increase in deletion errors across models owed to the presence of OOVs and singletons in the test set of the datasets. Substitution errors are higher than insertion and deletion errors across models due to low inter-class variability in some vocabulary words and limited training samples, which hinders the learning of distinctive features. Thus, designing datasets carefully is essential to avoid these flaws in real-world applications.

4) *NUMBER OF EFFICIENT CONFORMER ENCODERS*

Table 4 outlines the fluctuations in recognition performance based on the number of Efficient Conformer encoders utilized in the proposed architecture. Recognition rates exhibit a decline as the number of encoders increases. The augmentation of transformer encoders can heighten model complexity, potentially leading to overfitting on the training data and consequent poor performance on new, unseen data.

5) *NUMBER OF HEADS IN SIGN QUERY ATTENTION*

The primary optimization in the Conformer architecture involves organizing query vectors into clusters. The impact of the number of query heads, key-value heads (equivalent to the number of query clusters) and the number of heads per cluster on recognition performance, as well as the variation in trainable parameters with these factors, is elaborated in the results presented in Table 5. This investigation is done on the Phoenix14 T dataset. An important observation from this table is that having fewer heads per cluster results in more trainable parameters, regardless of the total number of query heads. The minimum trainable parameters are achieved when the number of heads per cluster is 32. However, the WERs are minimized when the number of heads per cluster is 16. Considering the trade-off between the efficiency and complexity of the model, the proposed model opts for 32 query heads and 2 query clusters (or key-value heads).

TABLE 5. Study on the Effect of Varying Query Heads and Query Clusters on Phoenix14 T

Query Heads	Key-Value Heads (Clusters)	No. of heads per cluster	#Parameters	WER
16	2	8	120942716	20.59
32	4	8	120942716	20.87
32	2	16	120794876	19.70
64	8	8	120942716	20.83
64	4	16	120794876	19.98
64	2	32	120720956	20.43

E. COMPARISON WITH STATE-OF-THE-ART RESULTS

We have successfully devised a highly optimized architecture that maintains impressive recognition rates compared to state-of-the-art models, as detailed in Table 6. To enhance the comparative study, focusing on the sequence learning module, the works are classified as single-modal and multimodal based on the modalities used for feature extraction. Overall, multimodal approaches tend to yield better performance across both datasets, as the fusion of complementary modalities helps capture richer spatiotemporal features. Among them, the current state-of-the-art model, ConSignformer [1], achieves notable results with a Transformer-based sequence learning backbone.

In contrast, single-cue models show a relatively higher WER, suggesting that relying on a single feature stream limits the model's ability to generalize, particularly for complex sign gestures. Among these models, Transformer-based methods such as SignBERT+ show promising results but still fall short of matching multi-modal performance due to their restricted input representation. We attribute the effectiveness of our proposed model to the synergistic combination of multiple strategies: the use of a Transformer-based backbone, pretraining with large-scale sign language videos, the integration of auxiliary networks, and the exploitation of multiple input cues (e.g., RGB, pose). To the best of our knowledge, no existing work has simultaneously adopted this combination of techniques, which we believe has been instrumental in achieving robust performance across both PHOENIX14 and PHOENIX14 T benchmarks.

TABLE 6. Comparison With Recent Works on CSLR

Model	Year	DL Framework	Additional Information				Phoenix14T		Phoenix14	
			Transformer-backbone	Pretraining of Context Learner	Auxiliary Task	Cross-Modality Learner	Dev WER	Test WER	Dev WER	Test WER
Multi-Modal										
DNF (RGB+Flow) [19]	2019	CNN+BiLSTM+CTC	✗	✗	✓	✓	—	—	23.10	22.90
STMC-R (RGB+Pose) [44]	2021	STMC+BiLSTM+CTC	✗	✗	✓	✓	19.60	21.00	21.10	20.70
TwoStream-SLR [9]	2022	CNN+Temporal Layers+CTC	✗	✗	✓	✓	17.70	19.30	18.40	18.80
C^2SLR (RGB+Pose) [45]	2024	CNN+Transformer+CTC	✓	✗	✓	✗	20.2	20.4	20.5	20.4
ConSignformer [1]	2024	**CNN+Conformer+CTC**	✓	✓	✓	✓	17.90	**18.55**	18.59	**18.59**
Optimized ConSignformer (Proposed Model)	2025	**CNN+Conformer/Temporal layers+CTC**	✓	✓	✓	✓	19.56	**19.70**	20.09	**19.94**
Single-Modal										
Joint-SLRT [22]	2020	CNN+Vanilla Transformer+CTC	✓	✗	✗	✗	24.60	24.50	—	—
CMA [46]	2020	CNN+BiLSTM+CTC	✗	✗	✗	✗	—	—	21.30	21.90
VAC [25]	2021	CNN+BiLSTM+CTC	✗	✗	✗	✗	—	—	21.20	22.30
SMKD [15]	2021	CNN+BiLSTM+CTC	✗	✗	✓	✗	20.80	22.40	20.80	21.00
MMTLB [10]	2022	CNN+Temporal Layers+CTC	✗	✗	✗	✗	21.90	22.50	—	—
LCSA [47]	2022	CNN+Context-Aware Transformer+CTC	✓	✗	✗	✗	—	—	21.40	21.90
SignBERT+ [16]	2023	CNN+Vanilla Transformer+CTC	✓	✓	✗	✗	18.8	19.9	19.9	20.0
CorrNet [48]	2023	CNN+BiLSTM+CTC	✗	✗	✗	✗	18.9	19.4	18.8	19.4
SEN [49]	2023	CNN+BiLSTM+CTC	✗	✗	✗	✗	19.3	20.7	19.5	21.0
MSTNet [50]	2024	CNN+(1DCNN-2DCNN-Transformer)+CTC	✓	✗	✗	✗	—	—	20.3	21.4

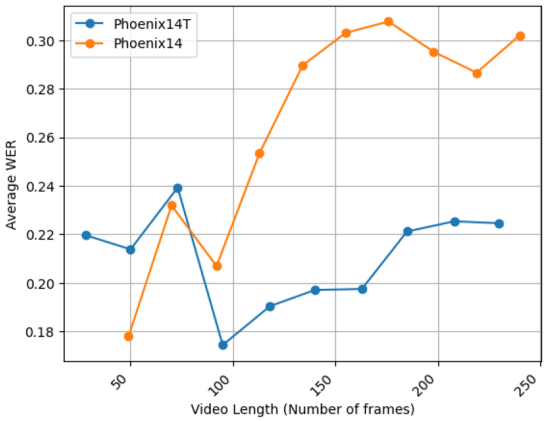

FIGURE 6. Average WER for videos in Phoenix14 with similar lengths with Exp_1 (temporal layers only) and Exp_4 (proposed model) set-ups.

Owing to the clustered query mechanism and simplified ensemble architecture, the model can be easily adapted for a product-level recognition system. This architecture enables fast gesture video decoding, achieving an average inference time of approximately 0.16 seconds for a 12-second video at 60 frames per second on an A100 GPU. Although numerous Transformer variants are available, limited experimentation has been conducted in sign recognition. We have successfully adapted the complex Conformer architecture and transformed it into an efficient version. Ablation studies show that for longer sequences, our Conformer-based recognition network is remarkably robust to varying video lengths.

Fig. 6 depicts the impact of video duration on the recognition rate using models defined by Exp_1 (no Conformer) and Exp_4 (proposed). It demonstrates that a transformer-based architecture outperforms the one without a transformer-based backbone. The previous state-of-the-art (SOTA) model [9] exhibits better overall WER than our optimized version on both benchmarks (current SOTA is ConSignformer). However, their architecture lacks a transformer-based sequential backbone, limiting its applicability in real-time recognition scenarios where long sequences are common. In such cases, the Efficient ConSignformer demonstrates better capability in handling longer sequences than the SOTA models without a transformer backbone. Compared to the limited works utilizing Transformers, our approach employs Conformer, which is more adept at simultaneously learning global and local features. Overall, our findings underscore the significance of incorporating a Conformer-based architecture for robust handling of long video sequences, highlighting its potential to bridge the gap between real-time applicability and recognition accuracy in continuous sign language tasks.

TABLE 7. Comparison of ConSignformer and Efficient ConSignformer

Metric	ConSignformer	Efficient ConSignformer
Inference Time (seconds)	0.24	0.16
Time for MHA (microseconds)	671 (Normal MHA)	560 (Sign Query Attention)
Parameters	≈ 237 million	≈ 120 million

F. CONSIGNFORMER VS. OPTIMIZED CONSIGNFORMER

This work focused on optimizing the ConSignformer architecture while retaining the advantages of the Conformer's innovative design. Parameter reduction was achieved by 49.2%. The optimized CSLR model demonstrates state-of-the-art performance on the renowned German datasets Phoenix14 and Phoenix14 T, particularly excelling with longer sign sequences. Inference time was evaluated across the entire test set of Phoenix14 T (642 videos), resulting in an average runtime of 0.16 seconds per video on an NVIDIA A100 GPU—a 1.5× improvement over the previous version's 0.24 seconds. The 95% confidence interval for inference time was calculated as

[0.158, 0.168] seconds, indicating stable and efficient performance across diverse test samples. Processing time for the MHA alone is reduced by approximately 100 microseconds for a 10-second video @30fps. The details are presented in Table 7.

V. CONCLUSION

In this study, the Conformer architecture is adapted for CSLR by replacing the standard MHA with SQA, resulting in the Efficient Conformer. The Efficient Conformer undergoes pretraining using a novel supervised task-adaptive CTC-based training approach with ISL pose data. Thus, we present the Efficient ConSignformer, a model comprising ensembles of dual-S3D pipelines with temporal convolution layers and the pretrained Efficient Conformer, specifically designed for sign recognition. This optimized CSLR model achieves state-of-the-art performance on Phoenix14 and Phoenix14 T benchmarks, excelling particularly on longer sign sequences. The clustered query mechanism speeds up the training process and video inference is completed in 0.16 seconds. Our future work will concentrate on additional architectural enhancements to achieve faster decoding without augmenting model complexity. We aim to address the challenges of longer video sequences, ultimately striving to develop an efficient, production-ready model.

ACKNOWLEDGMENT

This project draws inspiration and guidance from Amma, the Chancellor of Amrita University. The authors wish to extend their heartfelt appreciation to the Ministry of Electronics and Information Technology (MeitY), Government of India, for their generous financial support towards this research. The authors also extend special thanks to their collaborative partner, the Centre for Development of Advanced Computing (CDAC), for providing access to the PARAM Siddhi supercomputing facility, which was instrumental in carrying out this work.

REFERENCES

[1] N. Aloysius, G. M., and P. Nedungadi, "Continuous sign language recognition with adapted conformer via unsupervised pretraining," 2024, *arXiv:2405.12018*.

[2] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *Proc. IEEE Int. Conf. Commun. Signal Process.*, 2017, pp. 588–592.

[3] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[4] A. Choudhury, A. K. Talukdar, M. K. Bhuyan, and K. K. Sarma, "Movement epenthesis detection for continuous sign language recognition," *J. Intell. Syst.*, vol. 26, no. 3, pp. 471–481, 2017.

[5] M. Bhuyan, D. Ghosh, and P. Bora, "Continuous hand gesture segmentation and co-articulation detection," in *Computer Vision, Graphics and Image Processing*. Berlin, Germany: Springer, 2006, pp. 564–575.

[6] N. Aloysius and M. Geetha, "Understanding vision-based continuous sign language recognition," *Multimedia Tools Appl.*, vol. 79, no. 31–32, pp. 22177–22209, 2020.

[7] S. Renjith and R. Manazhy, "Sign language: A systematic review on classification and recognition," *Multimedia Tools Appl.*, vol. 83, no. 31, pp. 77077–77127, 2024.

[8] S. Alyami, H. Luqman, and M. Hammoudeh, "Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects," *Inf. Process. Manage.*, vol. 61, no. 5, 2024, Art. no. 103774.

[9] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 17043–17056.

[10] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, "A simple multi-modality transfer learning baseline for sign language translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5120–5130.

[11] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13009–13016.

[12] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3413–3423.

[13] R. Zuo, F. Wei, and B. Mak, "Towards online sign language recognition and translation," 2024, *arXiv:2401.05336*.

[14] J. Zheng et al., "CVT-SLR: Contrastive visual-textual transformation for sign language recognition with variational alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23141–23150.

[15] A. Hao, Y. Min, and X. Chen, "Self-mutual distillation learning for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11303–11312.

[16] H. Hu, W. Zhao, W. Zhou, and H. Li, "SignBERT+ : Hand-model-aware self-supervised pre-training for sign language understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11221–11239, Sep. 2023.

[17] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3793–3802.

[18] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4297–4305.

[19] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1880–1891, Jul. 2019.

[20] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4165–4174.

[21] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7784–7793.

[22] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10023–10033.

[23] M. Geetha, N. Aloysius, D. A. Somasundaran, A. Raghunath, and P. Nedungadi, "Towards real-time recognition of continuous indiansign language: A multi-modal approach using RGB and pose," *IEEE Access*, vol. 13, pp. 60270–60283, 2025.

[24] N. Aloysius, M. Geetha, and P. Nedungadi, "Incorporating relative position information in transformer-based sign language recognition and translation," *IEEE Access*, vol. 9, pp. 145929–145942, 2021.

[25] Y. Min, A. Hao, X. Chai, and X. Chen, "Visual alignment constraint for continuous sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11542–11551.

[26] Z. Zhou, V. W. Tam, and E. Y. Lam, "A cross-attention BERT-based framework for continuous sign language recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 1818–1822, 2022.

[27] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2019, Art. no. 2.

[28] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[29] M. Zaheer et al., "Big bird: Transformers for longer sequences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17283–17297.

[30] A. Radford et al., "Improving language understanding by generative pre-training," 2018.

[31] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.
[32] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[33] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
[34] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.
[35] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 5232–5270, 2022.
[36] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.
[37] F. Zhang et al., "MediaPipe hands: On-device real-time hand tracking," 2020, *arXiv:2006.10214*.
[38] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," 2020, *arXiv:2006.10204*.
[39] S. Jin et al., "Whole-body human pose estimation in the wild," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 23–28, 2020, Springer, 2020, pp. 196–214.
[40] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[41] Y. Lu et al., "Understanding and improving transformer from a multi-particle dynamic system point of view," 2019, *arXiv:1906.02762*.
[42] A. Vyas, A. Katharopoulos, and F. Fleuret, "Fast transformers with clustered attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21665–21674.
[43] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Comput. Vis. Image Understanding*, vol. 141, pp. 108–125, 2015.
[44] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for sign language recognition and translation," *IEEE Trans. Multimedia*, vol. 24, pp. 768–779, 2021.
[45] R. Zuo and B. Mak, "Improving continuous sign language recognition with consistency constraints and signer removal," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 20, no. 6, pp. 1–25, 2024.
[46] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in *Proc. 28th ACM Int. Conf. Multimedia*, pp. 1497–1505, 2020.
[47] R. Zuo and B. Mak, "Local context-aware self-attention for continuous sign language recognition," in *Proc. Interspeech*, 2022, pp. 4810–4814.
[48] L. Hu, L. Gao, Z. Liu, and W. Feng, "Continuous sign language recognition with correlation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 2529–2539, 2023.
[49] L. Hu, L. Gao, Z. Liu, and W. Feng, "Self-emphasizing network for continuous sign language recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 854–862.
[50] Q. Zhu, J. Li, F. Yuan, and Q. Gan, "Multiscale temporal network for continuous sign language recognition," *J. Electron. Imag.*, vol. 33, no. 2, pp. 23059–23059, 2024.

NEENA ALOYSIUS received the Ph.D. degree in computer science and engineering from Amrita Vishwa Vidyapeetham, Coimbatore, India. She was a Deep Learning Engineer with Amrita Center for Research in Analytics, Technologies & Education for four years. She also has more than five years of industry experience as a Systems Engineer with Infosys Technologies Limited. She is currently an Assistant Professor with the TKM College of Engineering, Kollam, India. Her research interests include deep learning, computer vision, and pattern recognition.

GEETHA M received the Ph.D. degree from Amrita Vishwa Vidyapeetham, Coimbatore, India. She has been involved in teaching and research since 2003. She is currently the Chairperson and Associate Professor with the Department of Computer Science Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amritapuri. Her research interests include computer vision, video analytics, machine learning, deep learning, and pattern recognition. She has funded projects and patents in the area of video analytics.

PREMA NEDUNGADI received the Ph.D. degree in computer science and engineering from Amrita Vishwa Vidyapeetham, Coimbatore, India. She is currently the Director of the Amrita Center for Research in Analytics, Technologies & Education, Amrita University, and a Professor with the Amrita School of Computing, Amrita Vishwa Vidyapeetham. She was the recipient of the Digital India Award from the Ministry of Electronics and Information Technology, India, in the category of digital empowerment. She was a finalist in the U.S. $7 million Barbara Bush Foundation Adult Literacy XPRIZE Competition.