IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3672~3682
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3672-3682

Indian sign language video generation using attention-enhanced generative adversarial network

Prachi Pramod Waghmare, Ashwini Mangesh Deshpande
Department of Electronics and Telecommunication, Maharshi Karve Stree Shikshan Samstha's Cummins College of Engineering for Women affiliated to Savitribai Phule Pune University, Pune, India

Article Info

Article history:

Received Apr 2, 2025
Revised May 20, 2026
Accepted Jul 9, 2026

Keywords:

Deep learning
Generative adversarial network
Indian sign language
Sign language translation
Squeeze and excitation
attention mechanism
Video generation

ABSTRACT

Sign language (SL) is the primary mode of communication for the Deaf signers. Despite advancements in deep learning, SL recognition, translation, and video generation face challenges like blurriness and inconsistencies. This research proposes a novel sign language translation (SLT) approach for Indian sign language (ISL) using an attention-driven generative adversarial network (GAN). The preprocessing pipeline includes video frame extraction, skeletal joint coordinate detection via OpenPose, and dynamic time warping (DTW) for pose data refinement. The squeeze and excitation (SE) attention mechanism enhances 2D convolutional layers, allowing the generator to focus on relevant skeletal pose sequences. A motion discriminator refines motion authenticity. Performance evaluation using two SL datasets demonstrates significant improvements in structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and temporal consistency metric (TCM) scores, achieving 99.60 (%) as SSIM, 31.10 dB as PSNR, and 0.9111 as TCM. The proposed model outperforms standard GAN and dynamic GAN in SL video generation.

Corresponding Author:

Prachi Pramod Waghmare
Department of Electronics and Telecommunication
Maharshi Karve Stree Shikshan Samstha's Cummins College of Engineering for Women
affiliated to Savitribai Phule Pune University
Karvenagar, Pune – 411052, India
Email: prachi.waghmare@cumminscollege.in

1. INTRODUCTION

Communication is fundamental to human interaction, primarily facilitated through speech and gestures. Sign language (SL) is a natural visual language that relies on hand movements, facial expressions, and body language to convey meaning. It is the primary mode of communication for the Deaf community, though some hearing individuals, including interpreters and family members, also use it. Each country has its own SL, such as Indian sign language (ISL), which has a distinct grammar and structure. Additionally, regional SLs like Marathi SL exist within India, each adapted to local linguistic and cultural contexts.

Traditional sign language processing (SLP) techniques struggle to generate natural, coherent sign sequences. Many deep learning-based approaches focus primarily on manual gestures while neglecting non-manual features such as facial expressions and body posture, leading to unnatural-looking sign avatars. Recent advancements in SLP have leveraged deep learning techniques for recognition, generation, and translation. Dhanjal and Singh [1] developed an automatic system to convert Punjabi text into ISL using natural language processing (NLP) to generate ISL glosses and corresponding sign representations. To enhance continuous SL recognition, Pu *et al*. [2] introduced a cross-modality augmentation framework integrating multimodal learning for improved accuracy. Similarly, Hu *et al*. [3] proposed the global-local

enhancement network (GLEN), which employs non-negative matrix factorization (NMF)-aware learning to extract robust features, enhancing recognition performance. Addressing continuous 3D SL production, Saunders *et al*. [4] designed a progressive transformer-based model incorporating mixture density networks for more natural sign synthesis.

Bragg *et al*. [5] provided an interdisciplinary perspective on SL recognition, generation, and translation, highlighting challenges, datasets, and future research directions. Stoll *et al*. [6] introduced 'SignSynth', a data-driven approach for generating SL videos using deep learning from textual and gloss inputs. Beyond SL, generative models have been instrumental in image and video synthesis. Xu *et al*. [7] proposed adversarially approximated autoencoder for realistic image manipulation, while Chen and Guo [8] reviewed autoencoders in deep learning and their applications in representation learning and generative modeling.

Text-to-image and video transformation techniques have also influenced SL research. Jung *et al*. [9] developed a language-agnostic model ensuring semantic consistency in generated images. Wu *et al*. [10] introduced a secure generative adversarial network (SecGAN), a cycle-consistent generative adversarial network (GAN) for securely recoverable video modification. Men *et al*. [11] proposed a GAN-based model for controllable person image synthesis, while Nguyen *et al*. [12] applied GAN-based future frame prediction for anomaly detection in surveillance videos. Kwon and Park [13] refined this approach using retrospective cycle GANs for improved temporal consistency, and Wang *et al*. [14] introduced 'Imaginator', a spatio-temporal GAN for realistic video generation. These advancements collectively demonstrate the impact of deep learning in SLP and video synthesis. The ability of GANs to create high-quality synthetic images and videos makes them a promising solution for SL video generation.

This research proposes an attention driven GAN-based model to generate high-quality, photo-realistic SL videos from skeletal positions. By incorporating linguistic context and expressive features, the proposed model aims to overcome the limitations of traditional GANs. The approach ensures the generation of natural and context-aware sign sequences, enhancing accessibility and communication for the ISL signers and non-signing individuals. The advancements in the GAN-based SL video generation can significantly improve real-time sign language translation (SLT) systems.

2. LITERATURE SURVEY

Recent advancements in SL production and video synthesis have leveraged neural networks and GANs to improve realism and accuracy. Dhanjal and Singh [15] proposed a project that aims to create an automatic system called 'SISLA' that can convert speech to ISL automatically. Three steps comprise the system's operation: translation of the source language into ISL; voice recognition (speech recognition (SR)) for isolated words in English, Hindi, and Punjabi; and a HamNoSys-based 3D avatar depicting ISL motions. The system's four main implementation modules comprise the requirement analysis, data collection, technical development, and assessment modules. The system is more effective because it supports multiple languages. Speech sample files for English, Hindi, and Punjabi were trained and tested. According to empirical findings, the proposed models have minimum accuracy levels for English, Punjabi, and Hindi of 91%, 89%, and 89%, respectively. Usability tests verified the system's usefulness in helping hearing-impaired people communicate and learn.

However, this system lacks the incorporation of non-manual SL characteristics and sentence-level translation. Stoll *et al*. [16] introduced Text2Sign, a neural machine translation and GAN-based framework for SL production, bridging the gap between textual input and sign videos. Mehta *et al*. [17] focused on automated 3D SL caption generation, enhancing accessibility for the hearing-impaired. Similarly, Natarajan and Rajasekar [18] proposed a dynamic GAN model for generating high-quality SL videos from skeletal poses, improving motion realism. Zelinka and Kanis [19] explored neural SL synthesis, emphasizing gloss-based representation learning for sign video generation. In the broader domain of video generation, the motion and content decomposed generative adversarial network (MoCoGAN) by Tulyakov *et al.* [20] introduced a motion-content decomposition framework, effectively separating motion and content in video synthesis. Aigner and Körner [21] developed FutureGAN, leveraging spatio-temporal 3D convolutions to predict future frames in video sequences. Siarohin *et al.* [22] advanced pose-based human image generation through deformable GANs, achieving natural motion synthesis. InfoGAN by Chen *et al.* [23] enhanced interpretability in representation learning by maximizing mutual information, benefiting structured video generation. Zhang *et al*. [24] proposed StackGAN, a two-stage architecture that generates photorealistic images from textual descriptions, influencing text-to-sign synthesis. These studies demonstrate the potential of deep learning techniques in improving SLT, video synthesis, and motion representation. By combining neural machine translation, pose estimation, and GAN-based models, researchers continue to advance SL accessibility and realistic motion synthesis.

The following are the major contributions for this framework. First, the attention-enhanced GAN framework for ISL video generation. Second, incorporating squeeze and excitation (SE) attention blocks to

enhance signer representation and pose-aware feature learning and fine-tuning of sign pose features. Third, SL video generated with enhanced output quality based on SE-GAN framework. The created/generated SL videos' visual realism, motion continuity, and temporal consistency closely matches with computationally demanding transformer/diffusion models for real-time ISL production with lightweight GAN frameworks.

3. METHOD

SL is essential for communication among Deaf signers, yet its interpretation remains challenging. SLT faces difficulties due to alignment learning, non-manual features, and a lack of expertise in ISL. Traditional GANs struggle with understanding linguistic context, leading to ineffective video generation. This work aims to generate realistic ISL videos by incorporating both manual and non-manual expressions. The preprocessing phase involves extracting skeletal joint coordinates using OpenPose and refining pose data via dynamic time warping (DTW). These skeletal features serve as input for an attention-driven GAN. Unlike traditional GANs, this model integrates an SE attention mechanism for 2D convolutional layers, enhancing video realism. The system employs a motion discriminator to differentiate between real and fake motion samples, while a motion generator creates lifelike movements. The attention mechanism guides the network, ensuring accurate sign representation and improved SL video translation, as shown in Figure 1. This section describe the steps of the approach followed to convert ISL videos into frames.

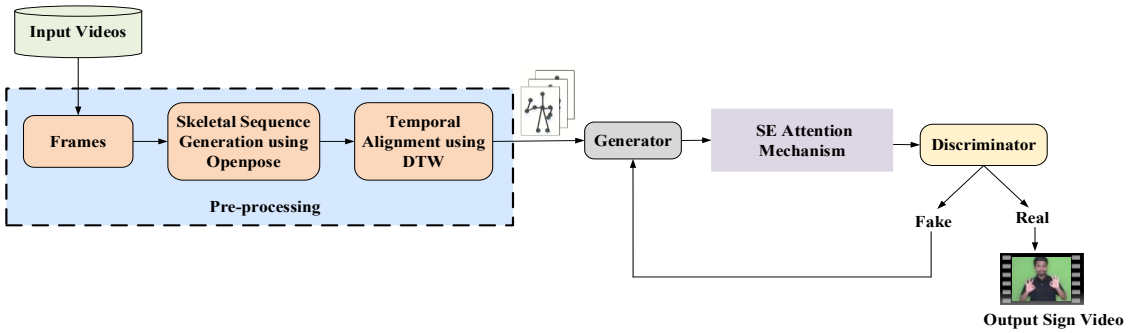


Figure 1. Proposed system for ISL video generation

3.1. Video to frame conversion

RGB frames are created from each ISL input video. The videos were originally recorded at 25 frames per second, so each video yields a different number of frames depending on its duration. After the captured video has been converted into a sequence of frames, an image preparation step (cropping, scaling, and normalizing) to remove noise during data collection once the captured video has been converted into a series of framed pictures for video processing.

3.1.1. Frame to skeletal pose

A popular computer vision library called OpenPose offers skeletal pose estimation from images or videos as well as real-time multi-person keypoint recognition. The model can estimate the 2D positions of important body joints, often called key points, like the head, shoulders, elbows, wrists, hips, knees, and ankles. Furthermore, when given a depth map, OpenPose can generate 3D pose predictions. Convolutional neural networks (CNNs) are used in combination with a deep learning techniques-based model to simultaneously detect body parts and infer their spatial correlations. OpenPose is flexible and has been implemented in different sectors, including human-computer interaction, augmented reality, and human behavior analysis. Its availability and open-source status have aided in its broad acceptance and incorporation into a multitude of keypoint detection and skeletal pose estimation applications.

In this work, OpenPose is used to extract skeletal joint coordinates from sign videos. The lower-body joints are not taken into consideration as they play no role in SL, and use 10 joints describing the upper body (head, neck, shoulders, elbows, wrists, and hips). Each joint is defined as a pixel in the image with coordinates x and y. To utilize multiple datasets from different domains, we perform the following normalization.

Every skeleton is positioned at the neck joint in relation to a selected reference skeleton as in (1).

$$Skel_T = Skel(x,y) + (N_{ref}(x,y) - N_{in}(x,y)) \quad (1)$$

Where the reference and input skeletons' neck joints are denoted by N_{ref} and N_{in}, respectively. After alignment, $Skel_T$ is the translated input skeleton. Next, the shoulder-to-shoulder distances of the input and reference skeletons are used to compute a scaling factor f as in (2).

$$f = \frac{abs\left(Sl_{ref}(x,y) - Sr_{ref}(x,y)\right)}{abs\left(Sl_{in}(x,y) - Sr_{in}(x,y)\right)} \quad (2)$$

Where Sl_{in} and Sr_{in} represent the left and right shoulder joints of the input skeleton, respectively, and Sl_{ref} and Sr_{ref} represent the reference skeletons. Using the translated skeleton, $Skel_T$, that was previously acquired and the scaling factor, f, finally compute the normalized skeleton, $Skel_{norm}$ as in (3).

$$Skel_{norm} = N_{ref}(x,y) + \left(Skel_T(x,y) - N_{ref}(x,y)\right) * f \quad (3)$$

Further, the frame annotations are used to categorize all skeletal sequences by frames to create the frame to pose sequence. Next, use DTW is used to align every skeletal sequence for every frame, and then combine them into a single representative mean skeleton sequence as in (4).

$$Skel_{frame} = \frac{1}{i}\sum_{j=1}^{i} DTW(Skel_{norm_j}, Skel_{norm_0}) \quad (4)$$

Where i is the number of frame sample sequences, and $Skel_{frame}$ is frame's representative mean sequence.

3.2. Dynamic time wrapping

Differentiating the local temporal sequences of different action categories is a challenging process due to several factors, such as frame rate fluctuations and temporal independence in each sublevel, regardless of whether several levels of shape-based alignment/matching (SHs) are used. To solve these problems, the distance between the various multiple/multidimensional DTW (MSH/multiple sequence alignment) levels for every action category is calculated using DTW. The time series alignment algorithm is called DTW. By repeatedly warping the time axis, it calculates a path between two sequences until ideal match is identified.

Assume that we have two time series, A and B, with respective lengths of n and m. The two time series are defined as (5).

$$A = a_1, a_2, \ldots, a_i, \ldots, b_n \; and \; B = b_1, b_2, \ldots, b_j, \ldots, b_m \quad (5)$$

Using DTW, n-by-m matrix was created, and its (i^{th}, j^{th}) element holds distance $d(i,j) = d(a_i, b_j) = \|a_i - b_j\|_2$ between the two points, a_i and b_j. every matrix element (i,j) represents how the two points are aligned. A contiguous set of matrix components that defines a mapping from A to B is called a warping path G. With $max(m,n) \leq K < m + n - 1$, we obtain $G = g_1, g_2, \ldots, g_k, \ldots, g_K$ as the k^{th} element of G, defined as $g_k = (i,j)_k$. The DTW distance is calculated as (6).

$$DTW(A,B) = min\left\{\frac{\sqrt{\sum_{k=1}^{K} g_k}}{K}\right\} \quad (6)$$

Likewise, nominal distances among MSHs of successive levels for every local joint are determined by DTW. Next, each local joint displacement's distance vector is created. For every action category, the temporal model of the skeleton joints is encoded as a concatenation of the distance vector $D^t = [T_1, \ldots, T_J]$. A skeleton representation feature vector is created by combining the skeleton joints' posture and motion feature vectors, which are computed for each action sequence. The skeleton representation feature vector is defined as (7).

$$S = \delta H^s + kD^t \quad (7)$$

Where δ and k are the weighting parameters.

Information regarding the spatiotemporal function over an action's time sequence can be found in the static pose and in the temporal dynamics of the harmonics. This harmonic information can be viewed as a compact representation of the skeletal joints in the body. As a result, such harmonic information enables accurate classification of activities.

3.3. Squeeze and excitation-generative adversarial network framework

GANs are extensively utilized in numerous domains for generation tasks. The conventional architecture of GANs, which consists of a generator (G) and a discriminator (D), has received exceptional consideration in recent years. The attention-driven GAN is introduced in our work for SL video generation. The proposed architecture shows significant improvement in the generation of SL video quality by utilizing SE attention module and carefully emphasizing skeletal poses. A graphical depiction of the proposed framework's general structure is shown in Figure 1. Its thorough description is given in this section.

3.3.1. Squeeze and excitation-generative adversarial network architecture

SE-GAN is made up of two key components: a generator and a discriminator. The discriminator is used to characterize the differences between the created detection image and the reference image, whereas the generator is intended to provide a detection image that is as accurate as the reference image. The framework minimizes the difference between the generated detection image and the reference image by use of adversarial training to minimize f-divergence.

The generator and discriminator are comprised of both encoders and decoders. The encoder in the generator is constructed using a rectified linear unit (ReLU) activation layer, a normalization layer, and a convolutional layer. A deconvolution layer, a normalization layer, and a ReLU activation layer make up the decoder. The generator G has a U-shaped structure and is made up of n encoders and decoders in addition to the SE attention module that is covered in the subsection that follows. It is the responsibility of the SE module to enhance the significant features transmitted by the convolution process in the encoder and decoder while disregarding the less significant ones. The next layers receive the reweighted features that were acquired from the SE module.

A convolution layer, a ReLU activation layer, three encoders, and then another convolution layer and a tanh layer make up the discriminator D. The generator generates an ISL image S from a skeletal pose I. The discriminator's inputs are thought to be a skeletal pose I, the generated detection image S, and the reference image R. For each image, the discriminator creates a representation of the variational variable, which is specified as v. To get the generated detection images to resemble the reference images, adversarial training of the discriminator and generator aims to minimize f-divergence.

3.3.2. Squeeze and excitation module

The result of the convolution operation is formed by summarizing all channels, and the relationship among the channels is invisibly intertwined in the spatial data. Therefore, the SE module's introduction aims to increase network sensitivity to key aspects by simulating channel interdependence. Less beneficial traits are therefore suppressed. The SE module is a component of the U-shaped generator construction that comes just after the encoders and decoders. The module consists of two distinct operations: excitation and squeezing. The result of the decoder and encoder U convolution process provides spatial and channel data. To reconstruct features, U is considered the input to the SE module. Here, consider that the combination of channels $u_i \in R^{H\times W}$ serves as the input for the SE module $U = [u_1, u_2, \ldots, u_C]$. A global average pooling operation is then used to compress the feature maps U from the convolution operation into a channel descriptor. Squeezing U with spatial dimensions $H \times W$ yields $sq \in R^{1\times 1\times C}$. The expression for the k^{th} element of $sq \in R^{1\times 1\times C}$ is as in (8).

$$sq_k = F_{sq}(u_k) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_k(i,j) \tag{8}$$

As a result, the vector sq, which encodes all of the information in the image, incorporates the global spatial information through the squeeze process. Figure 2 depicts the SE module's operation.

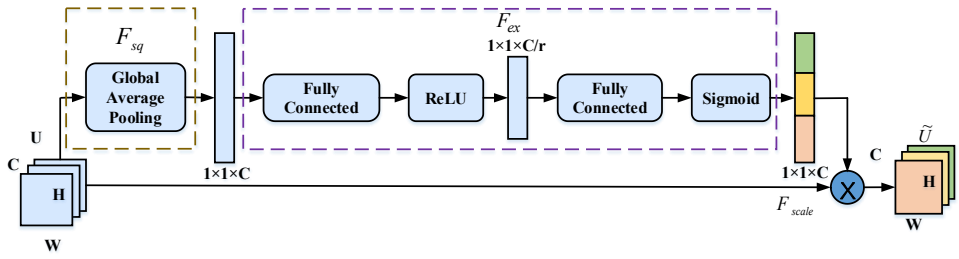


Figure 2. The SE module architecture

The excitation operation provides a comprehensive derived set of channel-wise dependencies, which can be better utilized in conjunction with the integrated global information gained by the squeeze operation. The global average pooling operation's sq output is converted to $W_2\left(\delta\left(W_{1sq}\right)\right)$. Here, the weights of two completely linked layers with r set to 2 are denoted by $W_1 \in R^{C/r\times C}$ and $W_2 \in R^{C\times C/r}$. The non-linearity in the excitation block then forms a bottleneck among the two fully connected layers. As shown in Figure 2, the framework of the excitation operation consists of a dimensionality-increasing layer with W_2, a ReLU layer, and a dimensionality-reduction layer with W_1. The interval [0,1] is the constraint of $W_2\left(\delta\left(W_{1sq}\right)\right)$ by a sigmoid activation function $\sigma(\cdot)$. The initial fully connected layer and the ReLU layer combine to provide a feature vector with a channel dimension of C/r. This feature vector is then transformed by the sigmoid layers and the second fully connected layer into a vector with a channel dimension of C. Thus, the final sigmoid layer's result as in (9).

$$ex = F_{ex}(sq, W) = \sigma\left(W_2\delta\left(W_{1sq}\right)\right) \tag{9}$$

Where ex denotes i th channel's significance. The final result is produced by scaling U using ex as in (10).

$$\bar{u}_k = F_{scale}(u_k, ex_k) = ex_k.u_k \tag{10}$$

In the excitation operation, the first fully connected layer's dimensionality reduction is intended to lower the network's computational complexity, while the second fully connected layer's dimensionality increase guarantees that the dimensions remain the same for determining the weight of each layer feature. The shape of U is recalculated to be $U = [ex_1u_1, ex_2u_2, \ldots, ex_Cu_C]$. The SE attention module adaptively highlights the key channels while ignoring the less significant ones. The detailed layer-wise configuration is shown in Table 1. Figure 3 shows stepwise process of the overall SE-GAN architecture, where GAN-based feature extraction and reconstruction along with SE attention mechanism and discriminator blocks are used. In this figure, T is number of frames, H is height, W is width, Cin is input channels, Conv is convolutional layer, Conv transpose is transposed convolution, s is stride, p is padding, and BN is batch normalization.

Table 1. Layer details for the proposed attention-driven GAN

Layer no.	Network configuration and hyperparameters
	Generator (encoder)
Layer 1	Conv 64(4×4), s =2, p=2, s =2, ReLU
Layer 2	Conv 128(4×4), s =2, p=2, s =2, BN, ReLU
Layer 3	Conv 128(4×4), s =2, p=2, s =2, BN, ReLU
	SE
Layer 1	Global average pooling (64,1,1,3)
Layer 2	Dense, ReLU
Layer 3	Dense, sigmoid
	Generator (decoder)
Layer 1	Conv 128(4×4), s =2, p =2, s =2, BN, ReLU
Layer 2	Conv 128(4×4), s =2, p =2, s =2, BN, ReLU
Layer 3	Conv 64 (4×4), s =2, p =2, s =2, ReLU
	Discriminator
Layer 1	Conv 32 (4×4), s =2, p =2, s =2, ReLU
Layer 2	Conv 32 (4×4), s =2, p =2, s =2, BN, ReLU
Layer 3	Conv 32 (4×4), s =2, p =2, s =2, BN, ReLU

The proposed architecture has generation and discriminator blocks, where the generator block has encoder and decoder units. Each encoder has a three-layer network. The first layer extract low features such as edges and contours. The second layer extracts mid-level features such as upper body parts and joints, while the third layer extracts high-level features such as pose structure. All the learned features are extended to the SE block, where the squeeze block computes global average pooling to capture global information of each channel. Here, channels can be key joints for hand and face. The excitation block learns channel importance and generates appropriate weights and fine details. The generator decoder does upsampling and reconstructs high-level features, recovers mid-level and fine details, and enhances the spatial resolution. The discriminator extracts features and make the decision whether the generated sample is real and fake. With the inclusion of SE-GAN attention mechanism, the discriminator receives information of important channel gestures, which are enhanced with the attention mechanism. There is improved hand motion and pose details and smooth temporal consistency, which is evident with the improvement in the structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and temporal consistency metric (TCM) score calculated for basic GAN, dynamic GAN on two different datasets.

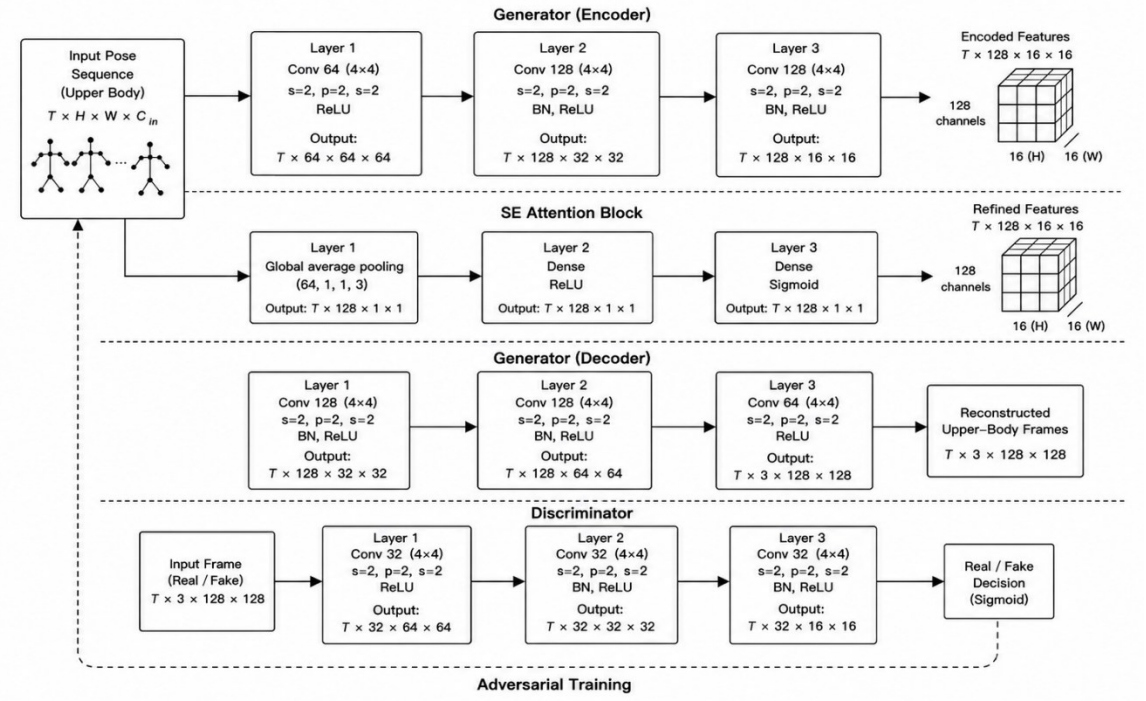


Figure 3. Process and layer wise results of the GAN-based feature extraction and reconstruction with SE attention mechanism

4. RESULTS AND DISCUSSION

This section presents the dataset description. It also results and compares the performance of baseline models and the proposed model. Optimal hyperparameters selected through experimentation are also included.

4.1. Dataset description

An SL video dataset, Indian sign language video and text dataset for sentences (ISLVT) [25] for ISL is created for this research. The purpose of this corpus is to aid the Indian Deaf community. The 700 videos in this unique corpus were gathered from a single signer and contain only a single background with well-set brightness conditions while recording. The other dataset used is the dataset for ISL-continuous sign language translation and recognition (ISL-CSLTR) [26].

4.2. Experimental results

The performance comparison of the baseline GAN, dynamic GAN, and proposed SE-GAN model is presented in Table 2. The comparison is based on three key evaluation metrics: SSIM, PSNR, and TCM. These metrics assess the visual quality, clarity, and temporal consistency of the generated frames.

Table 2. Comparison of generative models

Models	ISL-CSLTR dataset [26]			ISLVT (our) sign language dataset [25]		
	SSIM (%)	PSNR (dB)	TCM	SSIM (%)	PSNR (dB)	TCM
GAN [27]	81	7.6	0.48	80.15	16.84	0.5
Dynamic GAN [18]	93.7	8.5	0.712	85.81	18.79	0.65
SE-GAN (attention-driven) (proposed)	99	9.6	0.911	99.60	31.10	0.911

From Table 2, it is evident that the attention-driven GAN outperforms both the standard GAN and dynamic GAN across all three metrics. The proposed model achieves 99.60% SSIM, demonstrating superior structural similarity to the original frames compared to dynamic GAN (98.81%) and GAN (98.15%). A higher SSIM suggests that the generated frames retain more structural details, making them visually closer to real SL videos. Similarly, the PSNR value of the proposed model is significantly higher at 31.10 dB, compared to 18.79 dB for dynamic GAN and 16.84 dB for the standard GAN. A higher PSNR value signifies reduced noise and improved image clarity in the generated frames, confirming the model's ability to

reconstruct high-fidelity SL videos. Temporal consistency, measured by TCM, is crucial for smooth motion representation in video generation. The proposed model achieves a TCM score of 0.911, substantially outperforming the dynamic GAN (0.48) and GAN (0.50). This indicates that the attention-driven GAN generates temporally stable frames with consistent motion transitions, reducing flickering and unnatural movements. This performance shows consistent results for both datasets.

The quantitative comparison in terms of PSNR, SSIM, and TCM for the implemented GAN, dynamic GAN models, and the proposed SE-GAN model versus the training epochs is shown in Figure 4. It shows that SE-GAN consistently outperforms baseline models in terms of picture quality, temporal consistency, and structural similarity. It proves the efficacy of this method for producing closer to realistic SL videos. Figure 5 shows qualitative results comparison for GAN, dynamic GAN, and the proposed SE-GAN methods for the generated SL frames across various training epochs (20–100). These results also indicate that in comparison to the basic GAN and dynamic GAN models, the SE-GAN model produces more visually consistent, and enhanced-quality SL frames as the number of epochs increases.

The significant improvements in evaluation metrics highlight the model's potential for real-world SLT applications. Further, comparative results for the video frames generated by the baseline models and the proposed model for the original sample video, followed by the pose estimation step by the OpenPose are shown in Figure 6. Figure 6(a) shows the video frames generation results for the signs made for the text "Mother Food Cook" from ISLVT [25] dataset and Figure 6(b) depicts the results for the signs made for the text "R u free today" from the ISL-CSLTR dataset [26]. Overall, the results validate the effectiveness of the Attention-Driven GAN in enhancing the quality, clarity, and smoothness of SL video generation.

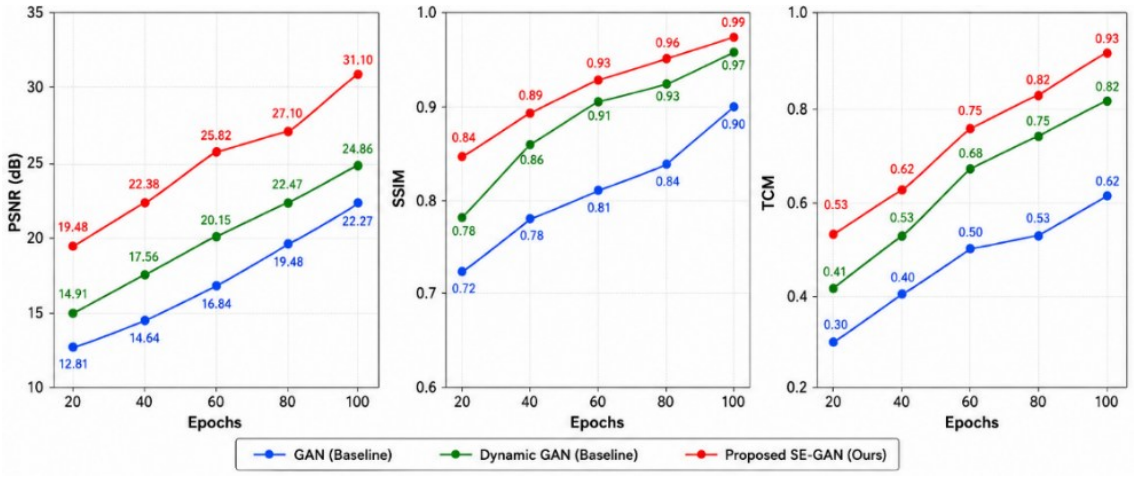


Figure 4. Quantitative comparison of GAN, dynamic GAN, and proposed SE-GAN models

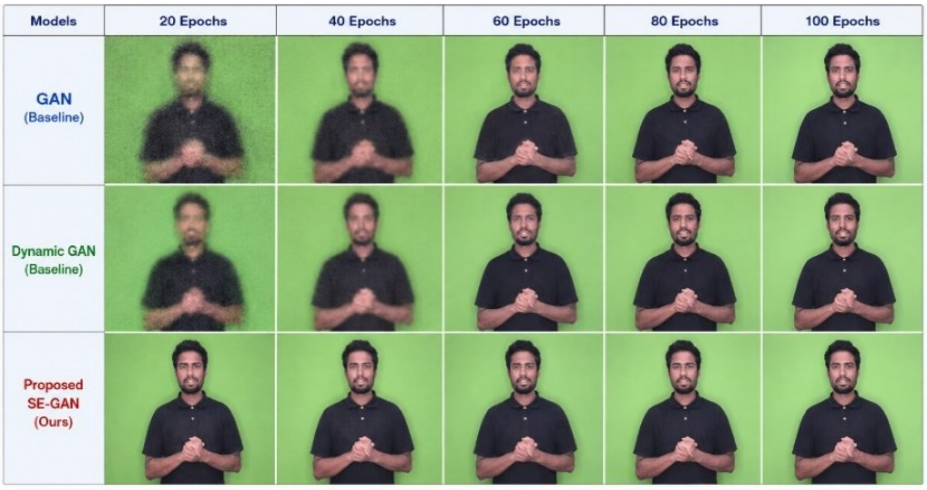
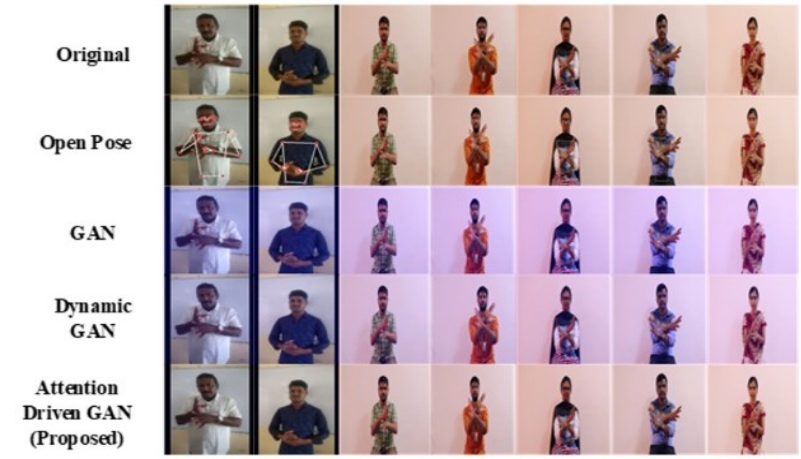


Figure 5. Qualitative comparison for generated SL frames and progression

(a)



(b)

Figure 6. Comparative video frame generation results for (a) sample output for dataset [25] and (b) sample output for dataset [26]

5. CONCLUSION

This research addresses the challenges in SLT for ISL video synthesis with improved video quality by introducing SE-GAN framework. It connects computer vision for gesture and pose understanding, NLP-driven text-to-sign translation, and multimodal learning frameworks that integrate linguistic and visual cues for more effective SL synthesis. Through a comprehensive approach involving video frame extraction, skeletal joint coordinate extraction, and DTW for data interpolation/extrapolation in the preprocessing step, the attention-driven GAN with an SE attention mechanism demonstrates its effectiveness in generating close to realistic SL videos. The attention mechanism dynamically modulates the generation process, allowing the model to attend to relevant skeletal pose sequences and frames, resulting in improved recognition metrics. Performance evaluation using ISL-CSLTR benchmark sign corpora and ISLVT sign corpora created for this research indicates substantial enhancements in video frame quality and temporal coherence. Comparative analysis with implemented models of GAN and dynamic GAN showcases the superior performance of the proposed model in terms of SSIM, PSNR, and TCM. These results validate the efficacy of the proposed system in advancing SLT, paving the way for improved communication for ISL signers and non-signing individuals. Future studies will look at diffusion-based generative models and reinforcement learning as a long-term AI roadmap for continuous SL production. Diffusion models may offer smoother and more accurate video synthesis for continuous sentence-level signing, while reinforcement learning agents can use gesture consistency feedback to dynamically improve temporal motion quality. These developments might

lead to multilingual, adaptive, signer-independent SL communication systems that can support practical accessibility applications in a variety of linguistic contexts.

ACKNOWLEDGMENTS

The authors would like to thank for Mrs. Janhavi and her team from BleeTech Innovation, Pune, who helped us in curating the dataset.

FUNDING INFORMATION

This research was partially supported by MKSSS's Cummins College of Engineering for Women, Pune, under the research initiation fund of the institute.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Prachi Pramod Waghmare	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Ashwini Mangesh Deshpande		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

The partial data that support the experiments and findings of this study is openly available in Mendeley Data at https://data.mendeley.com/datasets/98mzk82wbb/1. Other data used is properly cited.

REFERENCES

[1] A. Dhanjal and W. Singh, "An automatic conversion of Punjabi text to Indian sign language," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 28, Jul. 2018, doi: 10.4108/eai.13-7-2018.165279.
[2] J. Pu, W. Zhou, H. Hu, and H. Li, "Boosting continuous sign language recognition via cross modality augmentation," in *28th ACM International Conference on Multimedia*, Oct. 2020, pp. 1497–1505, doi: 10.1145/3394171.3413931.
[3] H. Hu, W. Zhou, J. Pu, and H. Li, "Global-local enhancement network for NMF-aware sign language recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 3, pp. 1–19, Aug. 2021, doi: 10.1145/3436754.
[4] B. Saunders, N. C. Camgoz, and R. Bowden, "Continuous 3D multi-channel sign language production via progressive transformers and mixture density networks," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2113–2135, Jul. 2021, doi: 10.1007/s11263-021-01457-9.
[5] D. Bragg *et al.*, "Sign language recognition, generation, and translation," in *21st International ACM SIGACCESS Conference on Computers and Accessibility*, Oct. 2019, pp. 16–31, doi: 10.1145/3308561.3353774.
[6] S. Stoll, S. Hadfield, and R. Bowden, "SignSynth: data-driven sign language video generation," in *Computer Vision – ECCV 2020 Workshops*, 2020, pp. 353–370, doi: 10.1007/978-3-030-66823-5_21.
[7] W. Xu, S. Keshmiri, and G. Wang, "Adversarially approximated autoencoder for image generation and manipulation," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2387–2396, Sep. 2019, doi: 10.1109/TMM.2019.2898777.
[8] S. Chen and W. Guo, "Auto-encoders in deep learning—a review with new perspectives," *Mathematics*, vol. 11, no. 8, Apr. 2023, doi: 10.3390/math11081777.
[9] S. Jung, W. S. Choi, S. Choi, and B.-T. Zhang, "Language-agnostic semantic consistent text-to-image generation," in *Proceedings of the Workshop on Multilingual Multimodal Learning*, 2022, pp. 1–5, doi: 10.18653/v1/2022.mml-1.1.
[10] H. Wu *et al.*, "secGAN: a cycle-consistent GAN for securely-recoverable video transformation," in *2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges*, Oct. 2019, pp. 33–38, doi: 10.1145/3349614.3356024.

[11] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed GAN," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp. 5083–5092, doi: 10.1109/CVPR42600.2020.00513.
[12] K.-T. Nguyen, D.-T. Dinh, M. N. Do, and M.-T. Tran, "Anomaly detection in traffic surveillance videos with GAN-based future frame prediction," in *2020 International Conference on Multimedia Retrieval*, Jun. 2020, pp. 457–463, doi: 10.1145/3372278.3390701.
[13] Y.-H. Kwon and M.-G. Park, "Predicting future frames using retrospective cycle GAN," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp. 1811–1820, doi: 10.1109/CVPR.2019.00191.
[14] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, "ImaGINator: conditional spatio-temporal GAN for video generation," in *2020 IEEE Winter Conference on Applications of Computer Vision*, Mar. 2020, pp. 1149–1158, doi: 10.1109/WACV45572.2020.9093492.
[15] A. S. Dhanjal and W. Singh, "An automatic machine translation system for multi-lingual speech to Indian sign language," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 4283–4321, Jan. 2022, doi: 10.1007/s11042-021-11706-1.
[16] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: towards sign language production using neural machine translation and generative adversarial networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 891–908, Apr. 2020, doi: 10.1007/s11263-019-01281-2.
[17] N. Mehta, S. Pai, and S. Singh, "Automated 3D sign language caption generation for video," *Universal Access in the Information Society*, vol. 19, no. 4, pp. 725–738, Nov. 2020, doi: 10.1007/s10209-019-00668-9.
[18] B. Natarajan and E. Rajasekar, "Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks," *Research Square*, Aug. 2021, doi: 10.21203/rs.3.rs-766083/v1.
[19] J. Zelinka and J. Kanis, "Neural sign language synthesis: words are our glosses," in *2020 IEEE Winter Conference on Applications of Computer Vision*, Mar. 2020, pp. 3384–3392, doi: 10.1109/WACV45572.2020.9093516.
[20] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: decomposing motion and content for video generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1526–1535, doi: 10.1109/CVPR.2018.00165.
[21] S. Aigner and M. Körner, "FutureGAN: anticipating the future frames of video sequences using spatio-temporal 3D convolutions in progressively growing GANs," Nov. 2018, *arXiv: 1810.01325.*
[22] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable GANs for pose-based human image generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 3408–3416, doi: 10.1109/CVPR.2018.00359.
[23] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: interpretable representation learning by information maximizing generative adversarial nets," in *30th Conference on Neural Information Processing Systems (NIPS 2016)*, 2016, pp. 2180–2188.
[24] H. Zhang *et al.*, "StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks," in *2017 IEEE International Conference on Computer Vision*, Oct. 2017, pp. 5908–5916, doi: 10.1109/ICCV.2017.629.
[25] P. Waghmare and A. Deshpande, "Indian sign language video and text dataset for sentences (ISLVT)," *Mendeley Data*, 2024, doi: 10.17632/98mzk82wbb.1.
[26] R. Elakkiya and B. Natarajan, "ISL-CSLTR: Indian sign language dataset for continuous sign language translation and recognition," *Mendeley Data*, 2021, doi: 10.17632/kcmpdxky7p.1.
[27] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *28th International Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

BIOGRAPHIES OF AUTHORS

Prachi Pramod Waghmare received masters degree from Pune University, India working as an assistant professor in the Department of Electronics and Telecommunication, Maharshi Karve Stree Shikshan Samstha's Cummins College of Engineering for Women, affiliated to Savitribai Phule Pune University, Pune, India. She is currently pursuing her Ph.D. and has made significant contributions to deep learning and sign language recognition, including developing deep learning approaches for Indian sign language recognition and video generation models. Developed the Indian sign language video and text dataset for sentences (ISLVT), a resource designed to enhance the precision and resilience of Indian sign language gesture recognition and generation systems. She actively participates in academic conferences and contributes to scholarly articles, focusing on improving communication accessibility for the hard of hearing community in India. She can be contacted at email: prachi.waghmare@cumminscollege.in.

Ashwini Mangesh Deshpande is an associate professor in the Department of Electronics and Telecommunication, Maharshi Karve Stree Shikshan Samstha's Cummins College of Engineering for Women, affiliated to Savitribai Phule Pune University, Pune, India. Her research interests include signal, image and video processing, computer vision, deep learning, NLP, and remote sensing. She has 26 years of experience in teaching and has published many research papers indexed in SCI, SCIE, Scopus, WoS journals; presented papers in several national and international conferences and has Indian patents with published and granted status. She has received best innovative teaching award and several research excellences awards. She has worked as a PI in various government-funded research projects and co-pi for institute-level research projects. She can be contacted at email: ashwini.deshpande@cumminscollege.in or amdeshpande20@gmail.com.