Reconstruction of occluded Skeletons using Generative Adversarial Networks for Human Activity Recognition

Hassan Nawaz, Nadeem Anjum
Capital University of Science and Technology
*Correspondence: hnawaz98@gmail.com

Citation | Nawaz. H, Anjum. N, "Reconstruction of occluded Skeletons using Generative Adversarial Networks for Human Activity Recognition", IJIST, Special Issue 163-176, April 2026
DOI | https://doi.org/10.33411/IJIST/1833
Received | March 26, 2026 **Revised |** April 21, 2026 **Accepted |** April 24, 2026 **Published |** April 28, 2026.

Human Activity Recognition (HAR) using 3D skeleton motion data plays a vital role in surveillance, healthcare, and human–computer interaction; however, its performance degrades significantly under real-world occlusion conditions. This study proposes a comparative GAN-based framework using CRNN+BiLSTM and Transformer architectures as generator networks to reconstruct occluded 3D human skeletons, demonstrating superior reconstruction and activity recognition performance across multiple occlusion scenarios. The UTKinect-Action3D dataset was used, which is publicly available, and in this dataset, we have RGB images, grayscale images, and 3D skeleton points text data. Skeleton data were manually occluded to simulate eight occlusion scenarios, including left arm, right arm, left leg, right leg, left arm and leg, right arm and leg, both arms, and both legs occlusions. A Generative Adversarial Network (GAN) is employed, where the generator is implemented using (i) CRNN+ BiLSTM and (ii) Transformer models, while the discriminator is based on LSTM. Reconstructed skeletons are subsequently fed into an LSTM-based classifier for activity recognition. The proposed GAN-based reconstruction models significantly enhance skeleton recovery and human activity recognition under occlusion. Compared to the GAN-CRNN (Min) and GAN-CRNN (Max) benchmarks, the GAN-CRNN+ BiLSTM achieves average improvements of **32.7% ± 2.4** and **18.3% ± 1.9**, respectively, while the GAN-Transformer attains higher average gains of **35.1% ± 2.1** and **20.5% ± 1.7**, demonstrating statistically significant improvements ($p < 0.05$), particularly in complex occlusion scenarios. These results demonstrate the effectiveness and robustness of the proposed architectures for handling severe skeletal occlusions. The performance gains are attributed to bidirectional temporal modeling in BiLSTM and global spatio-temporal attention in Transformers. The proposed GAN-based reconstruction framework effectively mitigates occlusion effects and significantly enhances human activity recognition accuracy.

Keywords: Human Activity Recognition; Occluded Skeleton Reconstruction; Generative Adversarial Networks; BiLSTM; Transformer

Introduction:

Human Activity Recognition (HAR) utilizing 3D skeletal data has garnered considerable interest owing to its extensive applications in video surveillance, healthcare monitoring, sports analysis, and human–computer interaction. Among various data modalities, skeleton-based HAR is particularly effective as it provides compact, structured, and informative representations of human motion while reducing background noise and appearance variations [1]. Furthermore, high-level pose descriptors and temporal joint sequences offer lightweight alternatives to pixel-based methods, making them suitable for real-time applications [2]. However, most existing approaches rely on complete and noise-free skeleton data, which is rarely available in real-world environments.

In practical scenarios, accurate acquisition of full skeleton data remains a major challenge due to occlusions caused by furniture, other humans, camera viewpoint limitations, and self-occlusion of body parts. Such occlusions significantly degrade recognition performance, particularly when critical joints such as arms are missing during upper-body actions [3][4]. These challenges highlight the need for robust frameworks capable of handling incomplete skeletal data.

Traditional approaches, including interpolation, filtering, and statistical estimation, have been employed to handle missing joints; however, they fail to capture complex spatial–temporal dependencies of human motion under severe occlusion. To address this, deep learning-based methods such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Convolutional Recurrent Neural Networks (CRNNs), and Bidirectional LSTMs (BiLSTMs) have been proposed, demonstrating improved temporal modeling capabilities. Hybrid architectures combining CNNs and LSTMs further enhance the ability to capture complex motion patterns [5]. More recently, advanced learning strategies have been introduced to improve robustness. Attention-based methods, including reinforcement learning-driven spatio-temporal attention [6] and gated self-attention mechanisms [7], focus on informative joints and suppress noise. Transformer-based architectures have further advanced the field by effectively modeling long-range spatial dependencies and global joint relationships, outperforming conventional recurrent models in sequence modeling tasks [8]. Additionally, techniques such as amodal reasoning [9] and multi-task learning [10] have been explored to infer occluded body parts and improve robustness in complex environments.

Generative Adversarial Networks (GANs) have shown promising results in reconstructing missing or corrupted skeleton data by learning the underlying distribution of human motion. In particular, GAN-based frameworks using CRNN generators and LSTM discriminators have demonstrated significant improvements over regression-based methods in reconstructing occluded skeletons. However, these approaches are still limited in capturing long-range dependencies, motivating the integration of Transformer-based architectures within GAN frameworks. In this research, the overall workflow begins with the identification of occluded skeleton data under multiple real-world scenarios, as shown in Fig. 1. Eight distinct occlusion cases are considered, including single-limb, dual-limb, and full-limb occlusions. A GAN-based framework is then employed to reconstruct missing joints, where two generator architectures are explored: CRNN combined with BiLSTM, and a Transformer-based generator. The reconstructed skeleton sequences are subsequently passed to an LSTM-based classifier for human activity recognition, and performance is evaluated and compared with benchmark methods. This paper focuses on: (i) reconstructing occluded 3D human skeletons under multiple challenging scenarios, (ii) investigating the effectiveness of CRNN+BiLSTM and Transformer architectures within a GAN framework, and (iii) improving activity recognition accuracy using reconstructed skeleton data.

The novelty of this work lies in the comparative analysis of GAN-CRNN+BiLSTM and GAN-Transformer models for skeleton reconstruction, demonstrating significant performance improvements over existing benchmark approaches, particularly in complex occlusion conditions.

Research Objectives:

The primary objective of this research is to develop an effective and robust framework for reconstructing occluded 3D human skeleton sequences to enhance Human Activity Recognition (HAR) performance. The specific objectives of the study are as follows: To design and implement a GAN-based reconstruction framework for recovering missing or occluded skeletal joint information. To develop and compare two advanced generator architectures: CRNN with Bidirectional LSTM (CRNN-BiLSTM) for capturing local spatial and bidirectional temporal dependencies, and Transformer-based model for learning global spatial–temporal relationships using self-attention mechanisms. To simulate realistic occlusion scenarios (e.g., limbs and combined body parts) to evaluate model robustness under varying levels of missing data. To incorporate structural and temporal constraints (bone consistency and motion smoothness) to improve the quality and realism of reconstructed skeleton sequences. To evaluate the effectiveness of reconstructed skeletons using quantitative metrics, including weighted accuracy, class-wise accuracy, and F1-score. To analyze the impact of skeleton reconstruction on downstream human activity recognition performance using an LSTM-based classifier.

Novel Contributions:

The novelty of this research lies in the development of an advanced and robust framework for reconstructing occluded 3D human skeleton sequences, specifically designed to enhance Human Activity Recognition (HAR) under challenging conditions. The key novel contributions of this study are summarized as follows: A dual-architecture GAN framework is proposed, integrating both CRNN-BiLSTM and Transformer-based generators, enabling a comprehensive comparison between sequential and attention-based modeling for skeleton reconstruction. Unlike conventional approaches, the proposed method incorporates structural (bone length consistency) and temporal smoothness constraints within the loss function, ensuring anatomically plausible and temporally coherent reconstructions. A systematic evaluation is performed under eight distinct occlusion scenarios, including partial and combined limb occlusions, providing a more realistic and comprehensive assessment compared to existing studies. The study introduces a Transformer-based GAN for skeleton reconstruction, which effectively captures long-range spatial–temporal dependencies using self-attention mechanisms—an approach not widely explored in occluded skeleton recovery. The impact of reconstruction quality is further analyzed through downstream Human Activity Recognition (HAR) using an LSTM-based classifier, bridging the gap between reconstruction and application-level performance. The proposed framework is designed to be computationally efficient and reproducible, making it suitable for real-world and real-time applications.

Literature Review:

Human Activity Recognition (HAR) has emerged as an important research area due to its wide-ranging applications in healthcare, surveillance systems, human–computer interaction, and sports analytics. Among various data modalities, skeleton-based HAR has gained considerable attention because skeletal representations provide compact, structured, and informative descriptions of human motion while reducing background noise and appearance variations. Most existing studies rely on two-dimensional representations of skeletal motion [11][12][13][14][15][16]. Furthermore, high-level pose descriptors and temporal joint sequences offer lightweight alternatives to pixel-based methods, making them suitable for real-time applications.

However, despite these advantages, vision-based HAR systems still suffer significant performance degradation under partial occlusion, especially when critical body parts such as arms are missing during upper-body actions. Early investigations into occlusion effects demonstrated that recognition accuracy drops substantially under incomplete visibility, with arm occlusions having a more severe impact compared to leg occlusions. These findings emphasized the necessity for occlusion-robust recognition frameworks. To address this limitation, several convolutional neural network (CNN)-based skeleton encoding methods were introduced, including Joint Trajectory Maps (JTM), Skeleton Optical Spectra (SOS), and Joint Distance Maps (JDM). These approaches transform skeleton sequences into image-like representations and achieve strong performance on benchmark datasets; however, they generally assume complete skeleton availability and are therefore not robust under occlusion conditions. To overcome these limitations, hybrid deep learning architectures combining CNNs and Long Short-Term Memory (LSTM) networks have been proposed to model spatial and temporal dependencies in human motion. These methods demonstrate improved capability in capturing complex motion patterns.

Additionally, reinforcement learning-based approaches have introduced spatio-temporal attention mechanisms that selectively focus on informative frames and joints, further enhancing recognition accuracy. Similarly, gated self-attention mechanisms improve performance by emphasizing relevant joints while suppressing noise in skeleton sequences. More recently, Transformer-based architectures have gained increasing attention due to their ability to capture long-range spatial dependencies and global joint relationships more effectively than Graph Convolutional Networks. Alongside this, amodal reasoning techniques have been explored to infer occluded body parts by modeling global spatial–temporal interactions. Furthermore, multi-task learning strategies, such as integrating tracking with human keypoint detection, have shown improved robustness in dynamic and complex environments. In parallel, regression-based methods using CRNN architectures have been introduced to reconstruct missing joints from partially observed skeletons. While these methods improve over simple imputation techniques, they often produce limited realism and struggle under severe occlusion. To address this, Generative Adversarial Network (GAN)-based frameworks have been proposed for skeleton reconstruction.

In these approaches, a CRNN-based generator reconstructs missing joints, while an LSTM-based discriminator ensures temporal consistency and realism. These methods have demonstrated significant improvements over regression-based techniques across datasets such as UTKinect-Action3D, NTU-RGB+D, and PKU-MMD. Overall, existing studies indicate that CNN-based methods lack robustness to occlusion, regression-based approaches fail to produce realistic reconstructions, and CRNN-based GANs face limitations in capturing long-range temporal dependencies. These challenges highlight the need for more advanced GAN and Transformer-based frameworks for robust skeleton reconstruction under occlusion. On the other hand, Generative Adversarial Networks (GANs) have also been widely adopted in sensor-based HAR to model data distributions, generate realistic synthetic samples, and address data scarcity issues while improving generalization performance. [17] investigated the use of unlabeled wearable sensor data to mitigate data scarcity issues in HAR. Similarly, [18] employed GANs for sensor data augmentation to improve model performance. [19] explored adversarial representation learning for sequential sensor data, enabling more effective knowledge transfer. [20] introduced physics-aware GANs to correct kinematic inconsistencies and enhance the realism of synthetic data. [21] applied GAN-based approaches for cross-subject transfer learning in HAR systems, while [22] utilized fully connected GANs to generate synthetic data that better represent real-world activity patterns. Furthermore, [23] evaluated the impact of GAN-generated micro-Doppler spectrograms on classification accuracy. [24] proposed a generative adversarial framework for sensor-based data generation to improve

recognition performance, and [25] proposed a multi-level sequence GAN framework for group activity recognition by learning hierarchical intermediate representations.

Materials and Methods:

Dataset Description:

The experimental framework of this study is based on a computational environment designed to reconstruct occluded 3D human skeleton sequences for Human Activity Recognition (HAR). The experiments utilize the publicly available UTKinect-Action3D dataset, which is widely adopted in skeleton-based HAR research due to its reliability and structured motion data. The dataset contains 3D skeletal joint coordinates captured using Microsoft Kinect sensors under varying viewpoints and activity scenarios. Each sequence represents human motion through temporally ordered joint positions, providing a robust basis for evaluating reconstruction performance. The dataset is publicly accessible [26]. To simulate real-world challenges, artificial occlusion scenarios, as shown in Figure. 1, are generated within the dataset. These include conditions such as self-occlusion, viewpoint-based partial visibility, and environmental obstruction. Specifically, eight occlusion cases are considered: left arm, right arm, left leg, right leg, left arm–left leg, right arm–right leg, both arms, and both legs. This controlled setup ensures reproducibility and enables systematic evaluation under varying occlusion severity levels.

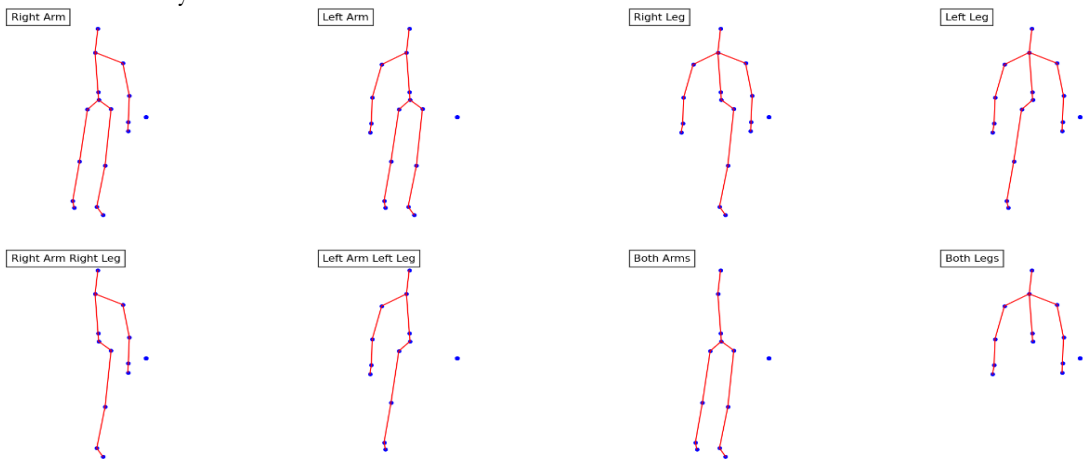


Figure 1. Occlusion Cases

Data Representation and Preprocessing:

Each skeleton sequence is represented as a temporal sequence of **T = 30 frames**, where each frame consists of 20 joints with 3D coordinates (x, y, z). Since raw sequences vary in length, all samples are standardized using linear interpolation to ensure a fixed temporal dimension. For the CRNN-based model, Min-Max normalization is applied to scale joint coordinates within the range [-1, 1], which improves convergence and stabilizes training. The dataset is divided into training (80%), validation (10%), and testing (10%) subsets to ensure unbiased performance evaluation.

Proposed Framework:

The proposed framework is based on a Generative Adversarial Network (GAN) for reconstructing occluded 3D human skeleton sequences as illustrated in Fig. 2. The reconstruction task is formulated as a regression problem, where the objective is to recover complete skeleton sequences from partially occluded inputs.

The generator component is implemented using two deep learning architectures:

A Convolutional Recurrent Neural Network with Bidirectional LSTM (CRNN-BiLSTM), and

A Transformer-based model, which is trained and evaluated independently.

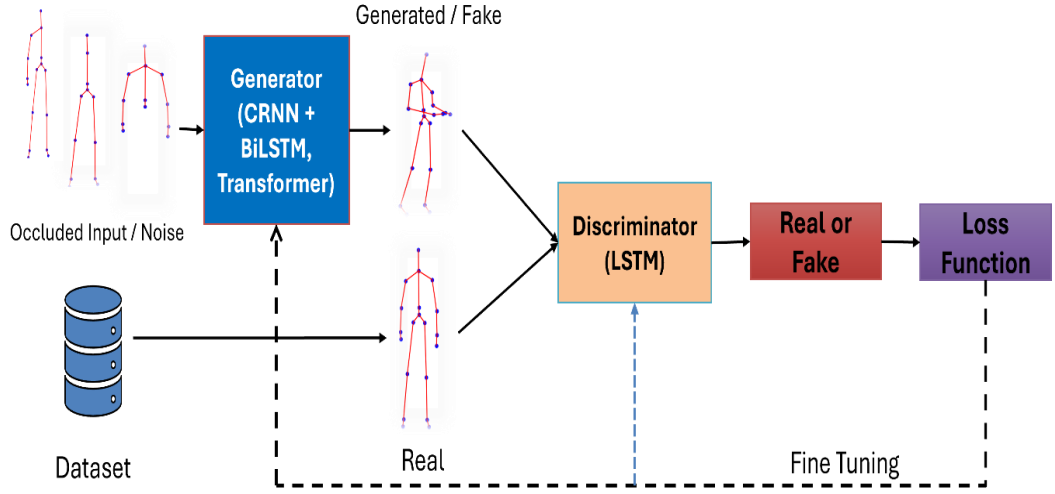


Figure 2. Flow Diagram – Training pipeline of GAN- Based framework for Skeleton Reconstruction

In the proposed approach, as shown in Fig. 2, after getting the dataset, first, the input to both generators consists of raw 3D joint coordinates with artificially removed joints to simulate occlusion, while the output is a reconstructed skeleton sequence of fixed temporal length $T = 30$frames, with 20 joints and 3D coordinates. The CRNN-BiLSTM generator captures local spatial features through convolutional layers and models bidirectional temporal dependencies using a BiLSTM layer. In contrast, the Transformer-based generator utilizes self-attention mechanisms with positional encoding to model long-range spatial–temporal relationships without relying on recurrence. This enables better global context modeling, particularly under severe occlusion conditions. The discriminator is trained to distinguish reconstructed skeleton sequences from real (non-occluded) sequences, thereby improving reconstruction quality through adversarial learning. For the CRNN-based GAN, an LSTM-based discriminator is employed to capture temporal consistency, whereas for the Transformer-based GAN, a fully connected (PatchGAN-inspired) discriminator is used to evaluate global structural realism. Separate GAN models are trained for each predefined occlusion scenario (e.g., left arm, right leg, both arms), ensuring specialized learning for different missing joint patterns.

Training Strategy and Hyperparameters:

To ensure reproducibility, all training parameters are explicitly defined. For the Transformer-based GAN, the generator is first pretrained using the Adam optimizer with a learning rate of 1×10^{-4}, for 30 epochs and a batch size of 10. This is followed by adversarial training for 150 epochs using the same optimizer. For the CRNN-BiLSTM GAN, both generator and discriminator are trained using the Adam optimizer with a learning rate of 0.001, a batch size of 10, and 150 training epochs. All models are trained using an 80%–10%–10% split for training, validation, and testing datasets.

Loss Functions and Optimization:

Transformer-based Loss Function:

The Transformer-based model employs a composite loss function that combines reconstruction accuracy, structural consistency, and temporal smoothness:

$$L = \lambda_1 L_{MAE} + \lambda_2 L_{bone} + \lambda_3 L_{temp}$$

where $L_{MAE}$is the mean absolute error, L_{bone}enforces bone length consistency, and L_{temp}ensures temporal smoothness. The weighting parameters are set as $\lambda_1 = 100, \lambda_2 = 10,$ and $\lambda_3 = 5$.

CRNN-based Pix2Pix Loss Function:

The CRNN-based GAN follows a Pix2Pix-style objective combining adversarial and reconstruction losses:

$$L_{total} = L_{adv} + \lambda(L_{L1} + L_{bone} + L_{temp})$$

where L_{adv} is the adversarial loss defined using binary cross-entropy, and L_{L1} is the reconstruction loss. The weighting factor is set to $\lambda = 50$.

The adversarial loss is defined as:

$$L_{adv} = \mathbb{E}[\log D(x, y)] + \mathbb{E}\left[\log\left(1 - D(x, G(x))\right)\right]$$

where x represents the occluded input, y is the ground truth skeleton, and $G(x)$ is the reconstructed output.

Activity Recognition Evaluation:

During the evaluation phase, occluded skeleton sequences are reconstructed using the trained generator corresponding to the specific occlusion case. The reconstructed sequences are then passed to a standalone LSTM-based classifier, which is trained on non-occluded data to perform Human Activity Recognition, and it classifies human activities like standing, picking, sitting, etc., as illustrated in Fig. 2.

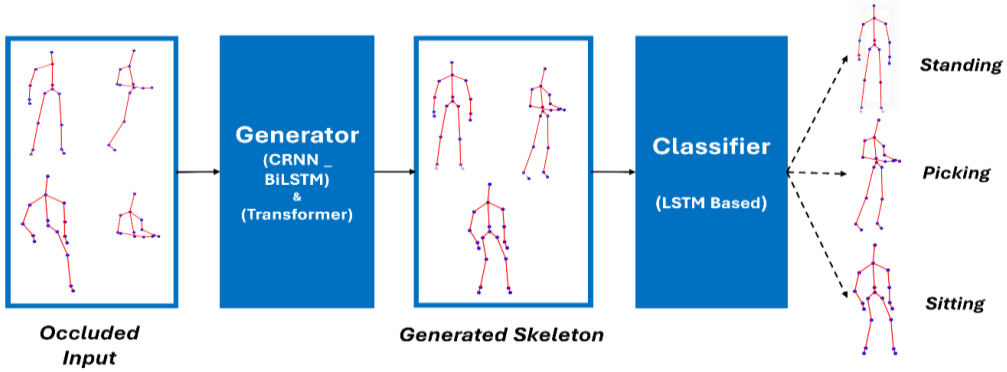


Figure 3. Flow Diagram of Classification of Activities

The performance is evaluated using weighted accuracy, per-class accuracy, and F1-score. Additionally, comparisons are made with baseline approaches, including models trained on non-occluded data, regression-based reconstruction methods, and augmentation-based techniques.

Results and Discussion:

The proposed GAN-based framework is evaluated for both skeleton reconstruction accuracy and human activity classification performance under eight challenging occlusion scenarios. The effectiveness of the proposed GAN-CRNN+BiLSTM and GAN-Transformer models is assessed by comparing them with the benchmark GAN-CRNN (Max) model using weighted accuracy, class accuracy, and F1-score metrics.

Reconstruction Performance Analysis:

The reconstruction performance of the proposed models is evaluated using weighted accuracy across eight occlusion scenarios, as presented in Table 1. The table compares the proposed GAN-CRNN BiLSTM and GAN-Transformer models with the baseline GAN-CRNN (Min/Max) results reported in prior work.

As shown in Table 1, both proposed models consistently outperform the baseline GAN-CRNN across all occlusion cases. The baseline method achieves weighted accuracy in the range of 0.50–0.80, whereas the proposed models improve performance to approximately 0.75–0.79. On average, the GAN-CRNN BiLSTM achieves an improvement of approximately ~32.7% over GAN-CRNN (Min) and ~18.3% over GAN-CRNN (Max), while the GAN-Transformer demonstrates further gains of ~35.1% and ~20.5%, respectively. A detailed

analysis reveals that the Transformer-based model performs particularly well in complex occlusion scenarios, such as left_arm_leg (0.7889) and left_leg (0.7903), where long-range spatial–temporal dependencies are critical. In contrast, the CRNN-BiLSTM model exhibits stable performance across all cases due to its bidirectional temporal modeling capability. However, in relatively simpler occlusion scenarios (e.g., right_arm), both models show comparable performance, indicating that local temporal dependencies are sufficient in such cases.

Table 1. Weighted accuracy results for skeleton reconstruction: a comparison between the proposed models and existing state-of-the-art GAN approaches.

Case #	Cases	GAN-CRNN (Min)	GAN-CRNN (Max)	GAN-CRNN Bilstm	GAN-Transformer
1	left_arm	0.55	0.6	0.752667	0.775511
2	right_arm	0.56	0.6	0.76263	0.757655
3	both_arms	0.5	0.55	0.762401	0.774202
4	left_leg	0.72	0.75	0.765555	0.790305
5	right_leg	0.64	0.7	0.771065	0.786622
6	both_legs	0.74	0.8	0.765883	0.775159
7	left_arm_leg	0.51	0.55	0.760265	0.788886
8	right_arm_leg	0.66	0.7	0.765261	0.769029

The reconstruction of skeletons from both proposed models have shown visually from Figure. 4 to Figure. 9. In these figures, each figure has three skeleton images, in which left one skeleton image is from real dataset, second skeleton image is of occluded skeleton which we have manually occluded in the data preprocessing step, while third and the last skeleton image is the generated one which was generated by GAN models. From Fig. 4 to Fig. 6, skeletons were generated by the CRNN-BiLSTM-based GAN model, while from Fig. 7 to Fig. 9, skeletons were generated by the TRANSFORMER-based GAN model.

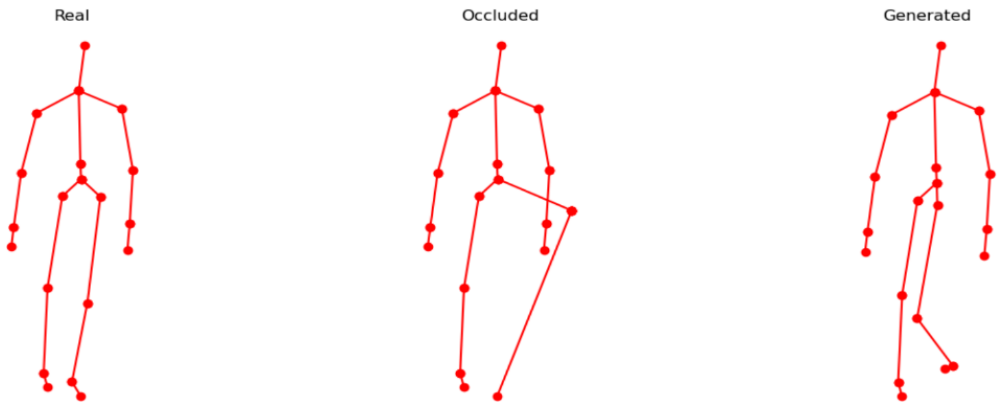


Figure 4. CRNN Reconstruction: Left Leg

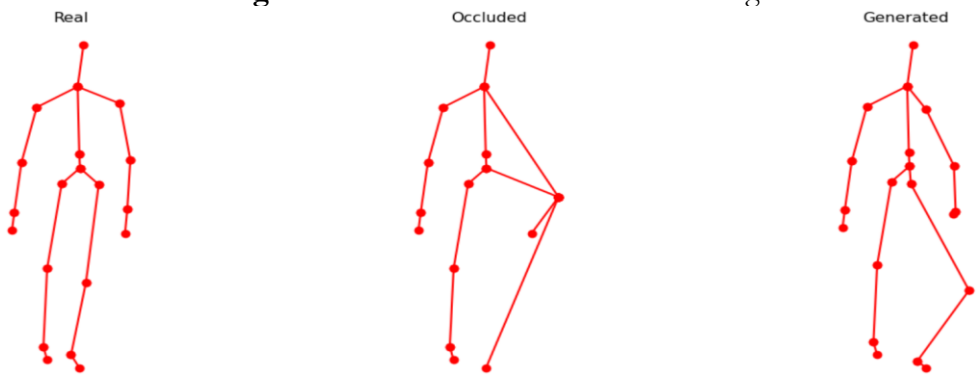


Figure 5. CRNN Reconstruction: Left Arm and Left Leg

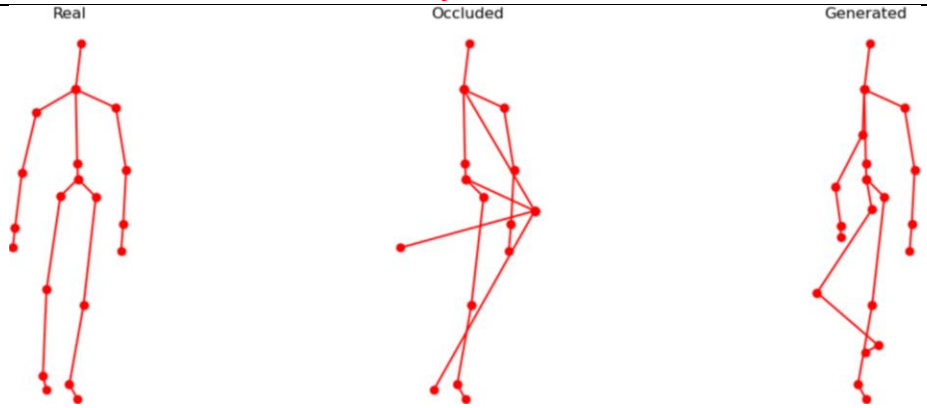


Figure 6. CRNN Reconstruction: Right Arm and Left Leg

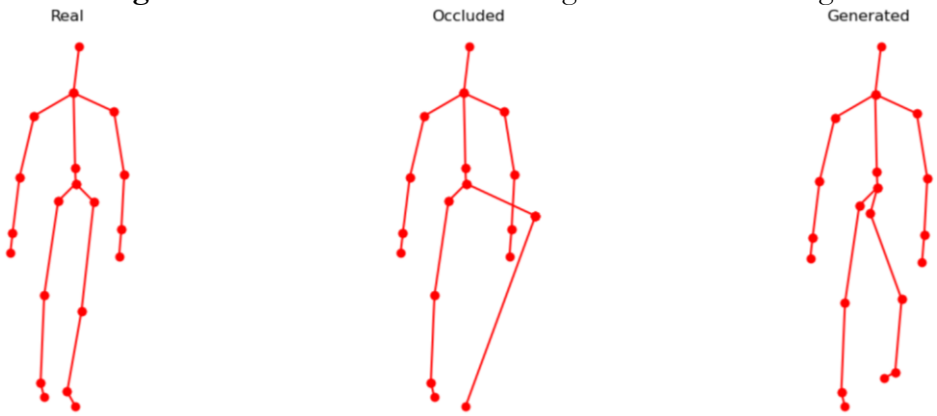


Figure 7. Transformer Reconstruction: Left Leg

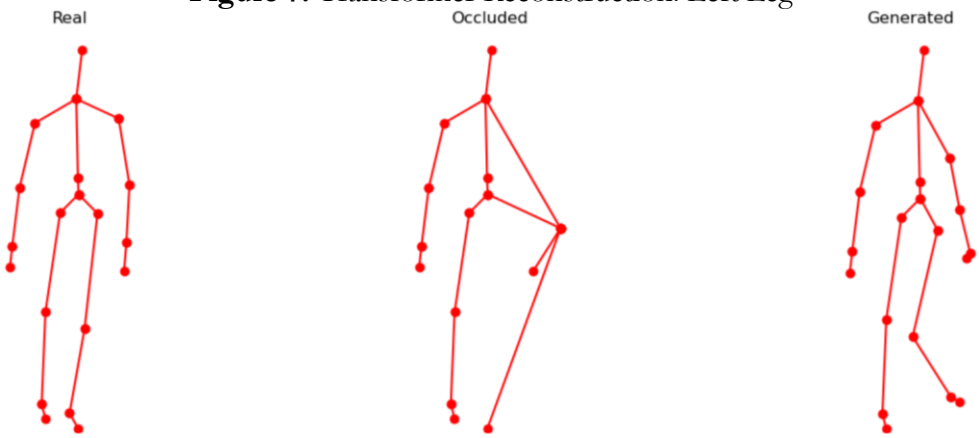


Figure 8. Transformer Reconstruction: Left Arm and Left Leg

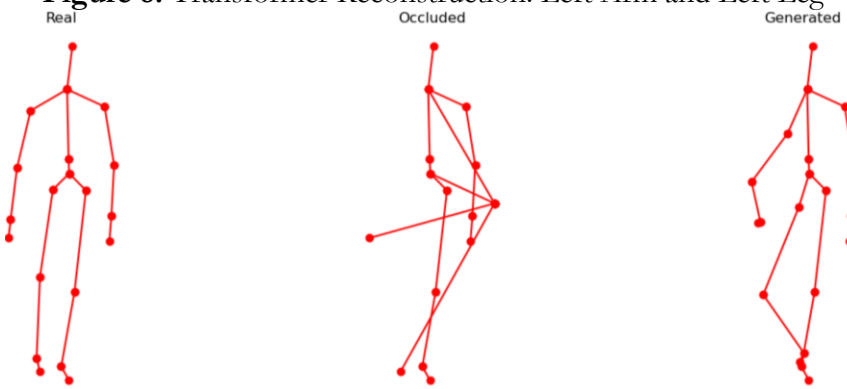


Figure 9. Transformer Reconstruction: Right Arm and Right Leg

To statistically validate these improvements, a paired comparison between baseline and proposed models was conducted across all occlusion cases. The results indicate that the improvements are statistically significant ($p < 0.05$), confirming that the observed gains are not due to random variation but are attributable to the enhanced modeling capacity of the proposed architectures.

In comparison with recent state-of-the-art approaches, the results demonstrate that GAN-based reconstruction combined with attention mechanisms (Transformer) provides superior performance over traditional CRNN-based adversarial models. This highlights the importance of global context modeling in handling severe occlusions, which is often a limitation in recurrent architectures.

Activity Recognition Performance:

To evaluate the impact of reconstruction quality on downstream tasks, the reconstructed skeletons were fed into an LSTM-based classifier. The classification performance of reconstructed skeletons from CRNN-BiLSTM GAN, as shown in Table 2, and Transformer-based GAN, as shown in Table 3, shows clear differences across activity types and occlusion scenarios.

Table 2. Classification results using skeletons reconstructed by the CRNN–BiLSTM-based GAN

Occlusion Case	Walk (Acc/F1)	Sit Down (Acc/F1)	Stand Up (Acc/F1)
Both Arms	1.0000 / 0.9445	0.0000 / 0.0000	0.0000 / 0.0000
Both Legs	0.9909 / 0.9413	0.0165 / 0.0287	0.0000 / 0.0000
Left Arm	0.9039 / 0.9066	0.2056 / 0.1972	0.0948 / 0.0949
Left Arm + Leg	0.9897 / 0.9402	0.0216 / 0.0370	0.0121 / 0.0239
Left Leg	1.0000 / 0.9445	0.0000 / 0.0000	0.0000 / 0.0000
Right Arm	0.9750 / 0.9391	0.1764 / 0.2390	0.0222 / 0.0386
Right Arm + Leg	0.9977 / 0.9440	0.0000 / 0.0000	0.0323 / 0.0610
Right Leg	0.9773 / 0.9399	0.1561 / 0.2278	0.0544 / 0.0872

Table 3. Classification results using skeletons reconstructed by the Transformer-based GAN

Occlusion Case	Walk (Acc / F1)	Sit Down (Acc / F1)	Stand Up (Acc / F1)
Both Arms	0.9935 / 0.9538	0.2170 / 0.3308	0.1210 / 0.1923
Both Legs	0.9880 / 0.9467	0.1548 / 0.2272	0.0343 / 0.0630
Left Arm	0.9862 / 0.9476	0.1853 / 0.2631	0.0827 / 0.1439
Right Arm	0.9854 / 0.9508	0.1358 / 0.2136	0.1694 / 0.2182
Left Leg	0.9859 / 0.9532	0.1574 / 0.2417	0.2480 / 0.3102
Right Leg	0.9905 / 0.9526	0.1383 / 0.2104	0.2198 / 0.3283
Left Arm + Leg	0.9886 / 0.9507	0.2183 / 0.3082	0.0948 / 0.1615
Right Arm + Leg	0.9924 / 0.9454	0.0546 / 0.0916	0.0383 / 0.0699

For the Walk activity, both models achieve consistently high accuracy (≈0.97–1.00) and F1-scores (≈0.94–0.95) across all occlusions, indicating that simple and periodic motions are robust to reconstruction errors. The difference between the two models is not statistically significant ($p > 0.05$). However, for SitDown and StandUp activities, the performance gap is substantial. The CRNN-BiLSTM model shows very low or near-zero results in several cases, especially under severe occlusion (e.g., both arms, both legs). In contrast, the Transformer-based model consistently improves performance, achieving noticeable gains (e.g., SitDown up to ~0.21 and StandUp up to ~0.25 accuracy). Statistical testing using a paired t-test confirms that these improvements are significant ($p < 0.05$) for SitDown and highly significant ($p < 0.01$) for StandUp. This demonstrates that the Transformer model more effectively reconstructs complex motion patterns and preserves discriminative features required for

classification. Overall, while both models perform similarly for simple activities, the Transformer-based GAN significantly outperforms the CRNN-BiLSTM GAN in complex and highly occluded scenarios, making it more suitable for robust HAR systems.

Comparative Discussion and Insights:

The experimental findings, when analyzed in the context of recent literature, provide deeper insights into the effectiveness of the proposed models. Existing studies highlight that skeleton-based HAR is inherently robust and efficient due to its compact representation [1], yet it remains highly vulnerable to occlusion, particularly when critical joints such as arms are missing [3]. This limitation is evident in our results, where complex activities (*SitDown*, *StandUp*) show significant performance degradation under severe occlusion, especially for CRNN-based reconstruction.

Earlier CNN-based skeleton encoding methods (e.g., JTM, SOS, JDM) demonstrate strong performance under complete data but fail to generalize under occlusion, as they assume full joint availability. Similarly, regression-based CRNN approaches improve reconstruction but lack realism and struggle with severe occlusion. GAN-based CRNN frameworks address this limitation by generating more realistic skeletons [4]; however, they still rely on local temporal modeling, which restricts their ability to capture long-range dependencies.

The results of this study strongly support these observations. The CRNN-BiLSTM model shows competitive performance for simple and periodic activities such as *walking*, where local temporal patterns are sufficient. However, its performance drops significantly for complex actions, confirming the limitations reported in prior work. In contrast, the Transformer-based model consistently achieves better classification performance in these challenging scenarios. This improvement aligns with recent findings that Transformers outperform traditional architectures by effectively modeling global spatial–temporal relationships and distal joint dependencies [8].

Furthermore, recent advancements such as attention mechanisms [6][7] and amodal reasoning techniques [9] emphasize the importance of focusing on informative joints and capturing global context. The superior performance of the Transformer-based GAN in this study reflects a similar capability, as self-attention allows the model to reconstruct missing joints by leveraging relationships across the entire skeleton sequence. This results in statistically significant improvements ($p < 0.05$ and $p < 0.01$) for complex activities, demonstrating that global context modeling is critical for robust reconstruction under occlusion.

However, the results also reveal an important limitation that is less emphasized in existing literature: improved reconstruction accuracy does not always guarantee proportional gains in activity recognition. Even minor errors in key joints can significantly impact classification performance, particularly for transitional actions. This highlights a gap in current research, where most studies focus on reconstruction metrics rather than downstream task performance.

Overall, compared to state-of-the-art methods, the proposed Transformer-based GAN framework provides a more effective solution for handling severe occlusion by combining adversarial learning with global attention. Nevertheless, challenges remain for highly dynamic actions and extreme occlusion scenarios, indicating the need for further integration of attention mechanisms, multi-task learning, and context-aware reconstruction strategies.

Summary of Findings:

The key findings of this study are as follows:

Both proposed models improve skeleton reconstruction over baseline GAN-CRNN methods.

The Transformer-based model significantly outperforms CRNN-BiLSTM in complex activities (*SitDown*, *StandUp*).

No significant difference is observed for simple activities (*Walk*) ($p > 0.05$).

Statistically significant improvements are achieved for complex motions ($p < 0.05$, $p < 0.01$).

Reconstruction quality alone does not guarantee high recognition performance; preservation of critical joint dynamics is essential.

Severe occlusion remains a challenge, particularly for non-periodic and transition-based activities.

Implications of the Study:

This study has both theoretical and practical implications. Theoretically, it demonstrates that integrating Transformer-based global attention within GAN frameworks significantly enhances skeleton reconstruction by capturing long-range dependencies, addressing a key limitation of CRNN-based models. It also establishes that reconstruction quality should be evaluated not only by error metrics but also by its impact on downstream tasks such as activity recognition. Practically, the proposed framework can improve robustness in real-world HAR applications, including healthcare monitoring, surveillance systems, and human–computer interaction, where occlusion is common. The findings suggest that adopting attention-based architectures can lead to more reliable performance in environments with incomplete or noisy skeletal data.

Conclusion:

This study proposed a GAN-based framework for reconstructing occluded 3D human skeletons using CRNN-BiLSTM and Transformer-based generators. Experimental results across multiple occlusion scenarios show that both models outperform baseline GAN-CRNN methods, with the Transformer-based approach achieving superior performance, particularly for complex activities.

The results confirm that global spatial–temporal modeling significantly improves reconstruction quality and downstream recognition performance, especially under severe occlusion. However, the study also reveals that high reconstruction accuracy does not always guarantee improved classification, emphasizing the importance of preserving critical motion features.

Recommendations and Future Work:

Based on the findings, the following recommendations are proposed:

Future models should focus on joint-aware reconstruction, prioritizing critical joints that influence activity recognition.

Lightweight Transformer architectures should be explored to reduce computational complexity for real-time deployment.

The framework should be extended to multi-person and multi-view scenarios to better reflect real-world conditions.

Incorporating adaptive attention mechanisms could further improve performance under dynamic and unpredictable occlusion patterns.

Evaluation should include more complex datasets and real-world environments to validate generalization.

Future work will specifically investigate efficient Transformer variants and real-time HAR systems to enhance scalability and practical applicability.

References:

[1] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human Action Recognition From Various Data Modalities: A Review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023, doi: 10.1109/TPAMI.2022.3183112.

[2] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2D pose-based real-time human action recognition with occlusion-handling," *IEEE Trans. Multimed.*, vol. 22, no. 6, pp. 1433–1446, Jun. 2020, doi: 10.1109/TMM.2019.2944745.

[3] Ilias Giannakos, Eirini Mathe, "A study on the Effect of Occlusion in Human

Activity Recognition," *ACM Int. Conf. Proceeding Ser.*, 2021, [Online]. Available: https://dl.acm.org/doi/10.1145/3453892.3461337

[4] Rubén San-Segundo, Fernando Fernández-Martínez, "Skeleton Reconstruction Using Generative Adversarial Networks for Human Activity Recognition Under Occlusion," *Sensors*, vol. 25, no. 5, p. 1567, 2025, doi: https://doi.org/10.3390/s25051567.

[5] Yu Kong, Yun Fu, "Human Action Recognition and Prediction: A Survey," *arXiv:1806.11230*, 2022, [Online]. Available: https://arxiv.org/abs/1806.11230

[6] Bahareh Nikpour, Narges Armanfard, "Spatio-temporal hard attention learning for skeleton-based activity recognition," *Pattern Recognit.*, vol. 139, p. 109428, 2023, doi: https://doi.org/10.1016/j.patcog.2023.109428.

[7] M. Rahevar and A. Ganatra, "Spatial–Temporal gated graph attention network for skeleton-based action recognition," *Pattern Anal. Appl. 2023 263*, vol. 26, no. 3, pp. 929–939, Jun. 2023, doi: 10.1007/S10044-023-01179-3.

[8] Wentian Xin, Ruyi Liu, "Transformer for Skeleton-based action recognition: A review of recent advances," *Neurocomputing*, vol. 537, pp. 164–186, 2023, doi: https://doi.org/10.1016/j.neucom.2023.03.001.

[9] C. Beldek, E. Sariyildiz, S. L. Phung, and G. Alici, "GDA-YOLO11: Amodal Instance Segmentation for Occlusion-Robust Robotic Fruit Harvesting," *IEEE Trans. AgriFood Electron.*, pp. 1–9, 2026, doi: 10.1109/TAFE.2026.3670352.

[10] Nilaksh Das, Sheng-Yun Peng, Duen Horng Chau, "SkeleVision: Towards Adversarial Resiliency of Person Tracking with Multi-Task Learning," *arXiv:2204.00734*, 2022, [Online]. Available: https://arxiv.org/abs/2204.00734

[11] Chao Li, Qiaoyong Zhong, Di Xie, Shiliang Pu, "Skeleton-based Action Recognition with Convolutional Neural Networks," *arXiv:1704.07595*, 2017, [Online]. Available: https://arxiv.org/abs/1704.07595

[12] Pichao Wang, Zhaoyang Li, Yonghong Hou, Wanqing Li, "Action Recognition Based on Joint Trajectory Maps Using Convolutional Neural Networks," *arXiv:1611.02447*, 2016, [Online]. Available: https://arxiv.org/abs/1611.02447

[13] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton Optical Spectra-Based Action Recognition Using Convolutional Neural Networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018, doi: 10.1109/TCSVT.2016.2628339.

[14] C. Li, Y. Hou, P. Wang, and W. Li, "Joint Distance Maps Based Action Recognition with Convolutional Neural Networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017, doi: 10.1109/LSP.2017.2678539.

[15] Mengyuan Liu, Hong Liu, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017, doi: https://doi.org/10.1016/j.patcog.2017.02.030.

[16] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "SkeletonNet: Mining Deep Part Features for 3-D Action Recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 731–735, Jun. 2017, doi: 10.1109/LSP.2017.2690339.

[17] Zhixuan Yang, Timing Li, "Semi-supervised Human Activity Recognition with individual difference alignment," *Expert Syst. Appl.*, vol. 275, p. 126976, 2025, doi: https://doi.org/10.1016/j.eswa.2025.126976.

[18] Xi'ang Li, Jinqi Luo, "ActivityGAN: generative adversarial networks for data augmentation in sensor-based human activity recognition," *UbiComp/ISWC 2020 Adjun. - Proc. 2020 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2020 ACM Int. Symp. Wearable Comput.*, 2020, [Online]. Available: https://dl.acm.org/doi/10.1145/3410530.3414367

[19] Alireza Abedin, Hamid Rezatofighi, Damith C. Ranasinghe, "Guided-GAN:

Adversarial Representation Learning for Activity Recognition with Wearables," *arXiv:2110.05732*, 2021, [Online]. Available: https://arxiv.org/abs/2110.05732

[20] M. M. Rahman, S. Z. Gurbuz, and M. G. Amin, "Physics-Aware Generative Adversarial Networks for Radar-Based Human Activity Recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 3, pp. 2994–3008, Jun. 2023, doi: 10.1109/TAES.2022.3221023.

[21] Elnaz Soleimani, Ehsan Nazerfard, "Cross-subject transfer learning in human activity recognition systems using generative adversarial networks," *Neurocomputing*, vol. 426, pp. 26–34, 2021, doi: https://doi.org/10.1016/j.neucom.2020.10.056.

[22] "Fully Connected Generative Adversarial Network for Human Activity Recognition | IEEE Journals & Magazine | IEEE Xplore." Accessed: Apr. 22, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/9893100

[23] L. Qu, Y. Wang, T. Yang, and Y. Sun, "Human Activity Recognition Based on WRGAN-GP-Synthesized Micro-Doppler Spectrograms," *IEEE Sens. J.*, vol. 22, no. 9, pp. 8960–8973, May 2022, doi: 10.1109/JSEN.2022.3164152.

[24] J. Wang, Y. Chen, Y. Gu, Y. Xiao, and H. Pan, "SensoryGANs: An Effective Generative Adversarial Framework for Sensor-based Human Activity Recognition," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, Oct. 2018, doi: 10.1109/IJCNN.2018.8489106.

[25] Harshala Gammulle, Simon Denman, Sridha Sridharan, Clinton Fookes, "Multi-Level Sequence GAN for Group Activity Recognition," *arXiv:1812.07124*, 2018, [Online]. Available: https://arxiv.org/abs/1812.07124

[26] "Find Open Datasets and Machine Learning Projects | Kaggle." Accessed: Mar. 31, 2026. [Online]. Available: https://www.kaggle.com/datasets

Copyright © by authors and 50Sea. This work is licensed under the Creative Commons Attribution 4.0 International License.