CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Reconstruction of Skeleton using Partially Occluded Data for Human Behavior Detection using Generative Adversarial Networks

by

Hassan Nawaz

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2025

Copyright © 2025 by Hassan Nawaz

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

Thanks to Allah Almighty, who gave me opportunity of Higher Studies. Afterthat, I will dedicate my work to my beloved Mother and my siblings and also want to dedicate my work to my Late Father.

CERTIFICATE OF APPROVAL

Reconstruction of Skeleton using Partially Occluded Data for Human Behavior Detection using Generative Adversarial Networks

by

Hassan Nawaz

(MCS233004)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Ahmad Din	FAST-NU,Islamabad
(b)	Internal Examiner	Dr. Syed Saqib Raza Rizvi	CUST, Islamabad
(c)	Supervisor	Dr. Nadeem Anjum	CUST, Islamabad

Dr. Nadeem Anjum
Thesis Supervisor
October, 2025

Dr. Mohammad Masroor Ahmed
Head
Dept. of Computer Science
October, 2025

Dr. M. Abdul Qadir
Dean
Faculty of Computing
October, 2025

Author's Declaration

I, **Hassan Nawaz** hereby state that my MS thesis titled "**Reconstruction of Skeleton using Partially Occluded Data for Human Behavior Detection using Generative Adversarial Networks** " is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Hassan Nawaz)

Registration No: MCS233004

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled **Reconstruction of Skeleton using Partially Occluded Data for Human Behavior Detection using Generative Adversarial Networks**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Hassan Nawaz)

Registration No: MCS233004

Acknowledgement

First of all , I would praise to Allah Almighty for enabling me to complete this work without any delay. Then, I would like to thank my mentor, my supervisor, Dr. Nadeem Anjum, who was my first professor and last professor from whom I get my university's first lecture and finally, he gave me his precious time to complete my thesis. After this, I would like to thanks all of the teachers, staff and other workers of university for providing us a good and learning environment. Then, I would like to special thanks to the friend who come in my life and shared joyful moments with me and also my office colleagues who really helped me out in completing my degree. At last, I would like to thanks my parents and my siblings, who have supported me and pushed me to complete my university on time. Thank you all.

(Hassan Nawaz)

Abstract

Human Action recognition is invovling the detection of 3d skeleton motion data from real word scenerio to detect human activities. The recognition of activities is affected due to the misinformed data i.e. occluded data where several parts of human body get occlude due to the presence of multiple fators, like behind the furniture, or self occlusion of body parts , in which camera has just captured the one arm, one leg from a single camera point of view. So, to tackle this problem, we have focused our research on the reconstruction of 3d human skeleton where some body parts are missing. We have considerd eight occlusion cases, i.e. left arm, right arm, left leg, right leg, left arm and left leg, right arm and right leg, both arms and both legs. We have used a GAN model which consists of two neural network i.e. Generator and Discriminator, in this regard, we have followed an approach in which Generator is based on CRNN+BiLSTM and Discriminator based on LSTM. We have firstly implemented this and results are get improved very well. Afterthat, what we did, we also implemented Transformer as a Generator Network of our GAN Model and the results were improved more than the CRNN+BiLSTM approach. So, our research work consists on these two models CRNN+BiLSTM and Transformers as generator. Reconstruction of Human skeleton from both approached are get the higher results than the benchmarked results. After Reconstruction, we have implemented classification of activities based on simple LSTM based classifier who's function is to classify the activities by providing reconstructed skeletons to it and we have classified the activities such as walk, run, jogging, clapping etc. Overall, our proposed methodolgy have increased the recognition of Human activities.

Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
BiLSTM	Bi Directional Long short term memory
CNNs	Convolutional Neural Networks
CRNN	Convolutional Recurrent Neural Networks
DLL	Deep Learning Models
GAN	Generative Adversarial Networks
HMHH	Hierarchical Motion History Histogram
HCI	Human Computer Interaction
HAR	Human Activity Recognition
JTM	Joint Trajectory Maps
JDM	Joint Distance Maps
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MSE	Mean Square Error
RNN	Recursive Neural Network
SOS	Skeleton Optical Spectra
WA	Weighted Accuracy

Symbols

L_{total}	Total generator loss thickness
L_{adv}	Adversarial loss
$Lambdal$	Weighting factor thickness
L_{MAE}	Mean Absolute Error loss
L_{bone}	Bone-length consistency loss
$L_{temporal}$	Temporal smoothness loss

Chapter 1

Introduction

1.1 Background

With the advancement in computer technology, it is considerable to recognize the Human activities in advance as they plays a vital role in the real world scenario for surveillance of any activities such as playing, walking, talking, looting, dacoity, target killing people in public areas like parks, footpaths, commercial areas where people enjoy, do shopping or engage with other people and also these activities can be performed in late nights or in the non rushy areas.

Recognition of human activities also plays an important role in healthcare. For example, the elder patients who live in their houses and there activities are being monitor by sensors and if any abnormal behavior detected, it will reported to their respective doctors. This helps in predicting health problems earlier and enables doctors to provide more accurate treatment based on patients' observed activities.

In HCI, our focus is on the design and use of computer technologies that facilitate the effective interaction between humans and computer machines. In this context, machines can be controlled or provided with input through hand gestures or brain signals using mind-perceiving headsets, without the need for physical devices or touch. In the sports domain, recognition of players' activities is helpful for coaches in training, and circumstances that may lead to injury can be identified in advance.

In robotics, human activity recognition is also useful because of the reason as robots can respond by perceiving human gestures and can participate in activities where both humans and robots are involved.

3D skeletal data [1] is highly effective in recognizing human activities because it captures both the spatial and temporal dynamics of human movement and provide a robust and precise movement of human actions. The human activities may involved such as walking, running, sitting, standing, eating, reading, or playing sports.

In real-world scenarios, occlusion of body parts (such as limbs, arms, forearms, and legs) is a major challenge when obtaining 3D skeletal data from cameras which installed in the desired locations where human activities are proposed to be detected such as parks, shopping areas, office buildings, hospitals, or roads. Occlusion of skeletal parts significantly affects the recognition of human activities and may result in misclassification.

1.2 Human Activity Recognition Approaches

We have mainly two approaches for Human Activity recognition, and the details are given below and also shown in Figure 1.1.

1.2.1 Sensor-based

1. Data Source: Wearable sensors (accelerometer, gyroscope, magnetometer, smartwatches, smartphones, IMUs).

2. Advantages:

   (a) Works in any environment (indoor/outdoor, day/night).

   (b) Privacy-preserving (no images/videos recorded).

   (c) Low computational cost and power consumption.

(d) Less affected by background noise or lighting.

3. Limitations:

   (a) Requires users to wear or carry devices.

   (b) Limited context (no scene understanding).

   (c) Sensor placement and orientation can affect accuracy.

1.2.2 Camera-based

1. Data Source: Video/RGB images, depth images, infrared, or skeleton joints extracted from cameras (RGB-D, Kinect, etc.).

2. Advantages:

   (a) Rich contextual information (background, interactions, body pose).

   (b) Non-intrusive (no wearable needed).

   (c) Advanced methods (CNNs, Transformers) achieve high accuracy.

3. Limitations:

   (a) Sensitive to lighting, occlusion, and camera angle.

   (b) High computation and storage requirements.

   (c) Privacy concerns (video recording).

   (d) Limited in outdoor/large-scale environments.

1.3 Occlusion

Occlusion refers to the absence of joints or body parts that are concealed for various reasons. For example, when recording human activity, self-occlusion, furniture placed in front of the subject, or obstacles behind which the subject passes may cause certain body parts to become invisible. This can be observed in Figure 1.2.

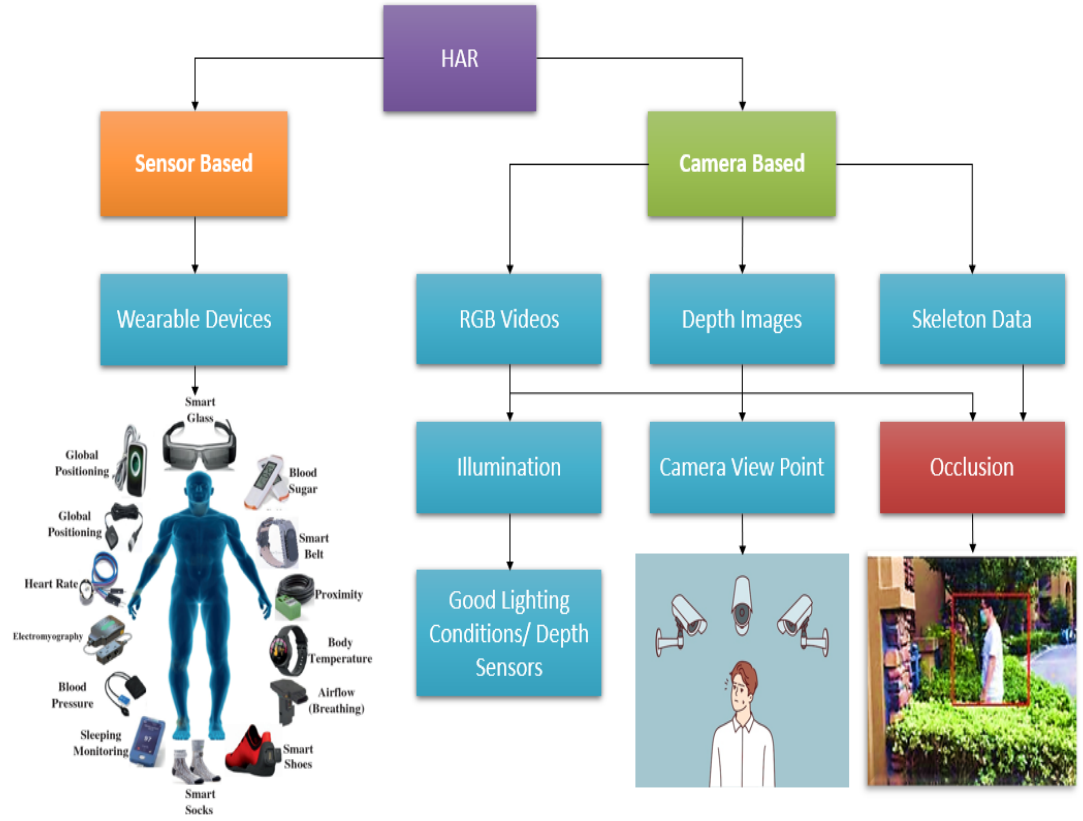

FIGURE 1.1: Human Activity Recognition Approaches

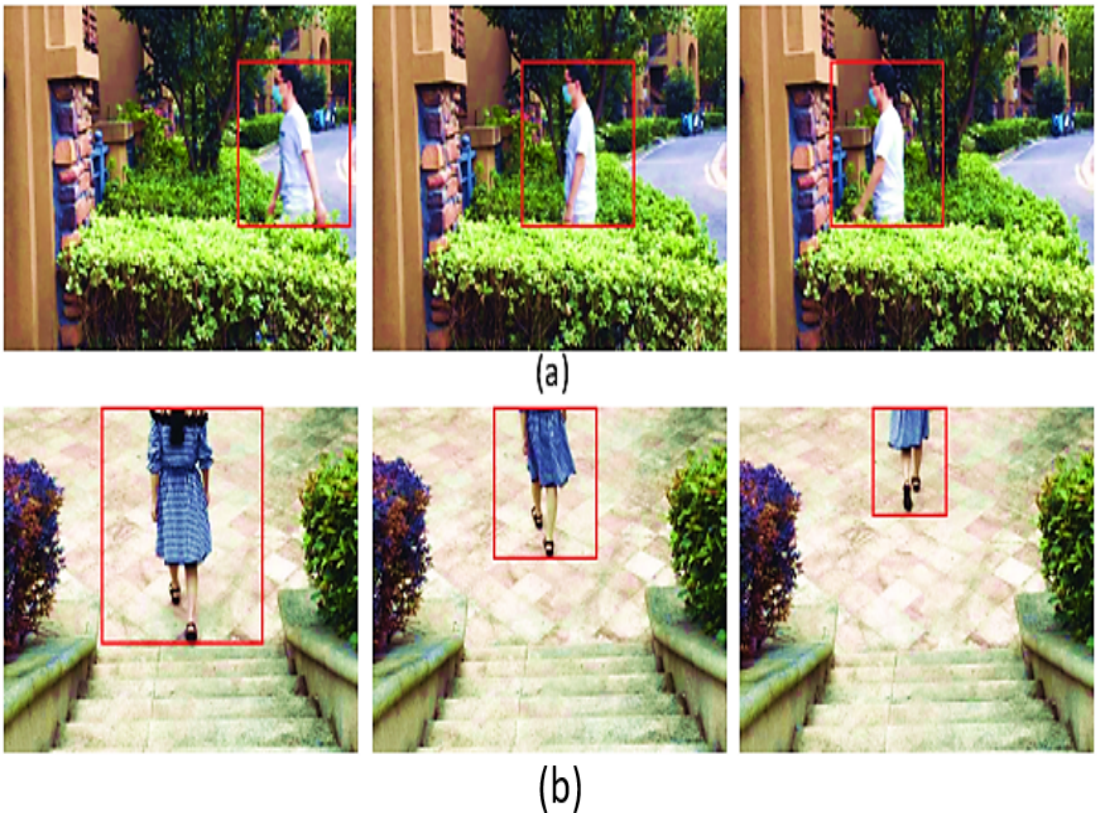

FIGURE 1.2: Occlusion

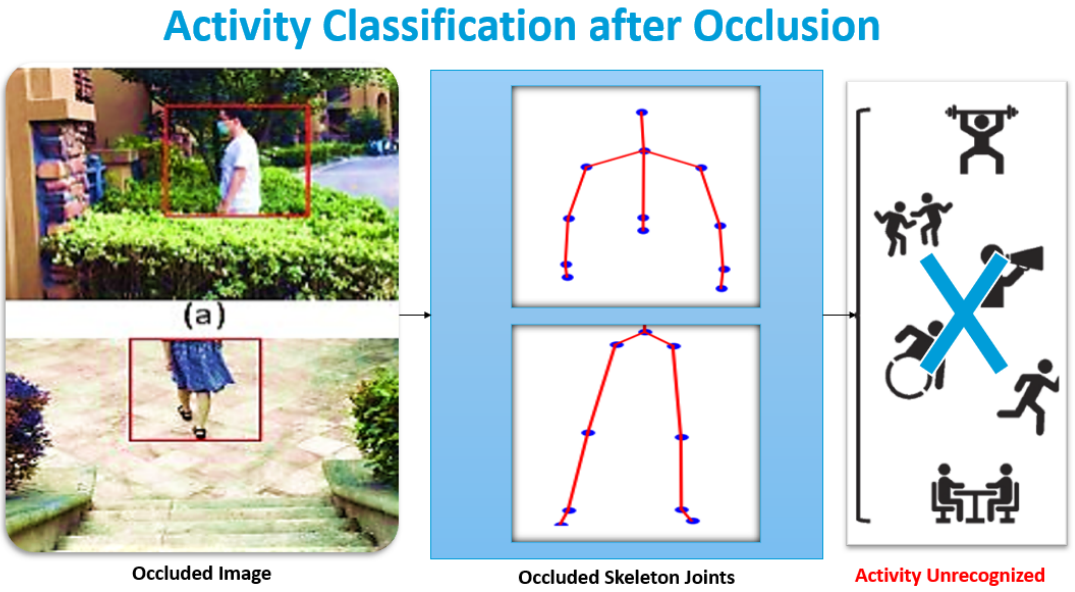

FIGURE 1.3: Occlusion - Classification of HAR

Occlusion affects the classification of activities. It is nearly impossible to recognize which activity is being performed when body parts are missing. Specific activities depend on specific body parts such as clapping depends on the hands. If the hands are occluded, the activity may not be recognized or may be misclassified. However, if other body parts such as the legs are occluded but the hands are visible, clapping can still be correctly recognized. This is illustrated in Figure 1.3, which shows how occlusion affects the classification of activities.

1.4 Problem Statement

In considering the advantages of Human Activity Recognition (HAR), a major problem that can affect the recognition of human activities is the occlusion of body parts, such as missing arms, hands, forearms, legs, feet, or other joints.

Occluded skeleton data can lead to misjudgment in human activity recognition. Occlusion significantly effect the recognition of human activities like missing of body parts (e.g., arms, legs, or other joints) distort the input data and it reduces the classification accuracy of human activities.

Since 3D skeleton data plays an important role in identifying human behaviors during activity performance, many researchers in the past have proposed solutions using various deep learning models, mainly CNNs and RNNs, as well as appearance-based template matching paradigms such as MHI, DMHI (Directional Motion History Image), MMHI, and HMHH (Hierarchical Motion History Histogram).

To address the problem of body-part occlusion, this research aims to reconstruct the missing skeleton joints or body parts using Generative Adversarial Networks (GANs), which consist of two deep neural networks: a Generator and a Discriminator. In this study, we utilize the publicly available UT Kinect 3D dataset.

1.5 Research Objectives

1. Reconstruction of partially occluded skeletons into complete 3D skeletons for improved HAR performance.
2. Addressing the problem of occlusion by formulating the reconstruction of missing skeletal data as a Generative Adversarial Network (GAN) task.
3. Evaluating the effectiveness of the proposed GAN-based skeleton reconstruction approach using publicly available datasets under various occlusion scenarios.

1.6 Motivation

In the presence of occluded body parts, recognition of human activities becomes a challenging task, as missing body parts affect the classification of activities. Therefore, we are strongly motivated by the idea that reconstructing the missing body parts can enhance human activity recognition.

The primary motivation of this research is to develop an effective method for reconstructing missing skeletal data caused by occlusion and to improve the accuracy

of activity recognition from occluded skeleton sequences. This will contribute to enhancing the performance of HAR systems in environments where occlusion is inevitable.

1.7 Research Purpose

In the real world Environment, recognition of human activities is become more challenging task when partial occlusion occurs due to factors like self-occlusion in which a human body part is occluded by its own body part, someone comes in front of a targeted individual, due to furniture like human behind table, desk, chair etc.

To solve this problem, researchers have tackled this problem by implementing digital image processing models MHI, DMHI (Directional Motion History Image), MMHI, HMHH (Hierarchical Motion History Histogram), deep learning models like CNN, RNN etc. [7]

This research going to solve this problem of partially occluded skeleton by reconstructing the skeleton using the novel approach i.e. Generative Adversarial Networks. After that we will be able to recognize the human activities from the reconstructed skeleton which is generated by providing partially occluded data.

1.8 Proposed Methodology

- Generator: A Convolutional Recurrent Neural Network (CRNN) and Transformer act as the generator at a single time each. Its task is formulated as a regression problem, aiming to reconstruct the complete 3D skeleton sequence given an occluded sequence. The input to the CRNN/Transformer is the raw 3D joint positions with the occluded joints removed, and the output is the reconstructed skeleton. The CRNN includes a BiLSTM layer to capture the temporal information in the skeletal data, while the Transformer utilizes self-attention mechanisms to model long-range dependencies across the

temporal sequences more effectively. Both DL architectures are designed to learn spatio-temporal correlations among joints across frames of skeleton sequences, but the Transformer outperforms in capturing global context without recurrence, which makes it particularly more effective in reconstructing complex spatio-temporal motion patterns. The generated 3D skeleton output from both generators separately is then passed to the Discriminator, which evaluates it against real (non-occluded) skeleton sequences and thereby improves reconstruction quality through adversarial learning. The proposed architecture from [1] can be seen in Figure 1.4.

- Discriminator: Long Short-Term Memory (LSTM) network is used as the discriminator. It takes as input both real (non-occluded) and generated (reconstructed) skeleton sequences and learns to distinguish between them. The discriminator's loss is a sigmoid cross-entropy loss of the real and generated sequences.

- Training: A separate GAN (generator and discriminator pair) is trained for each specific case of partial occlusion (e.g., left arm occluded, both legs occluded). The GAN objective function adopted is based on the Pix2Pix GAN framework, combining a conditional GAN loss with an L1 loss to ensure the generated skeleton is close to the real one.

- Classification: After training the GAN, during the evaluation phase, an occluded activity sample is fed into the trained generator (CRNN/Transformer) corresponding to that specific occlusion pattern to reconstruct the missing skeletal data. The reconstructed skeleton sequence is then passed to a separate LSTM classifier (trained exclusively on non-occluded data) to recognize the performed activity. Different classifier architectures (with one or three inputs depending on the number of camera viewpoints in the dataset) are used, featuring LSTM and dense layers.

- Evaluation: The proposed approach is evaluated on one publicly available 3D skeleton dataset (UTKinect-Action3D) by manually simulating different

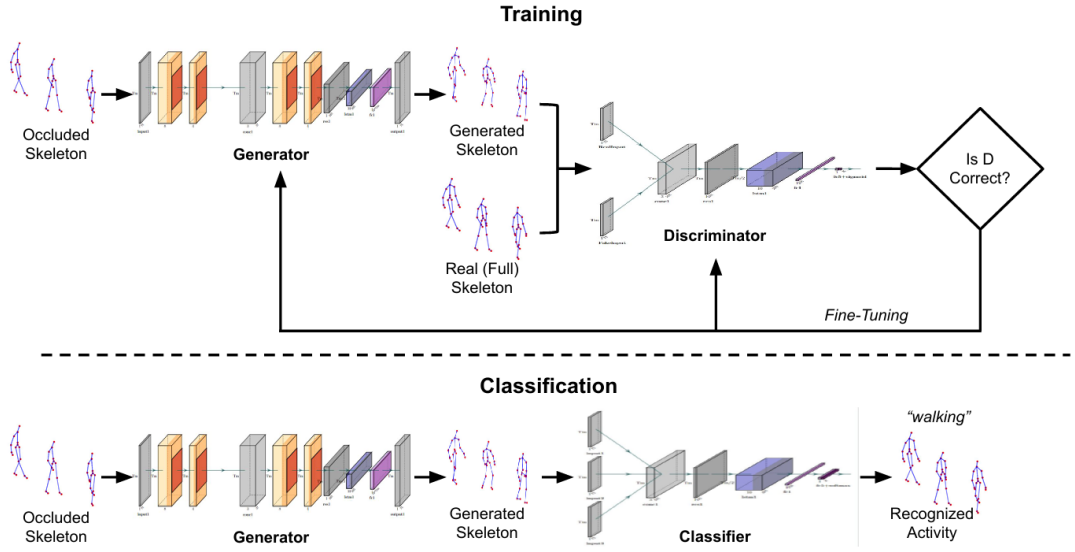

FIGURE 1.4: Model Architecture

partial occlusion scenarios. The performance is compared against a baseline (LSTM trained and evaluated on non-occluded data), a reference case (LSTM trained on non-occluded, evaluated on occluded), a regression-based reconstruction method from previous work, and an augmentation approach using artificially occluded samples. Weighted accuracy, per-class accuracy, and F1-scores are used as evaluation metrics.

1.9 Thesis Organization

There will be five chapters in the thesis which are given below;

- Introduction: This chapter consists of the background, Problem statement, Research Objectives, Research purposes, proposed methodology, thesis organization and contribution.
- Literature Review: Details of the relevant literature (online / offline).
- Proposed Methodology: Proposed Methodology of DLL models etc.
- Results and Disscusions: Reconstructed and classification results.
- Conclusion and Future work: Future enhancements.

1.10 Contribution

The contribution of our work is;

- Improving Results for a Single-Camera Viewpoint Dataset
- Two models were implemented i.e. a Transformer-Based GAN and CRNN + BiLSTM Based GAN.
- Improved results over the previous CRNN-based GAN in occluded cases.

Chapter 2

Literature Review

2.1 Introduction

Human Activity Recognition (HAR) has become a crucial research domain due to its wide range of applications, including healthcare, surveillance, human-computer interaction, and sports analysis. With the availability of RGB, depth, and skeleton data, researchers have explored both traditional and deep learning-based techniques to recognize human actions. Among these modalities, skeleton-based HAR has gained significant attention, as skeleton data provides compact yet informative representations of human motion that are less sensitive to variations in appearance, background, and illumination.

Despite these advantages, skeleton-based HAR faces a major challenge: occlusion. In real-world scenarios, parts of the human body are often occluded due to self-occlusion, interaction with objects, or other environmental factors. Such occlusions degrade recognition performance, as models typically assume complete skeleton availability. Therefore, various methods have been proposed to reconstruct or compensate for missing skeleton joints, ranging from regression-based models to more advanced Generative Adversarial Networks (GANs). This chapter reviews the literature in this field, focusing on approaches that address occlusion

and skeleton reconstruction, and highlighting how these methods contribute to advancing HAR.

2.2 GAN-based Skeleton Reconstruction under Occlusion

The first notable work on handling skeleton occlusion through GANs was introduced in [1]. This study proposed a novel GAN framework specifically for reconstructing occluded skeleton data to improve HAR performance. The architecture consisted of a CRNN-based generator and an LSTM-based discriminator. The generator was responsible for reconstructing missing skeleton joints from partially occluded sequences, while the discriminator evaluated the realism of the reconstructed skeletons. Experiments conducted on multiple benchmark datasets, including PKU-MMD, NTU-RGB+D, SYSU-3D-HOI, and UTKinect, demonstrated that the proposed GAN framework significantly outperformed regression-based approaches. The method achieved improvements in weighted accuracy ranging from 2.2% to 37.5% under different occlusion scenarios. This contribution was pivotal, as it was the first attempt to leverage adversarial learning for skeleton reconstruction. However, the study also suggested future directions, such as applying Transformers for better long-range dependency modeling and validating the approach in real-world scenarios.

2.3 Occlusion Challenges in HAR

Another important study investigated the effects of occlusion on skeleton-based HAR using CNN models [2]. The authors analyzed how recognition accuracy varied depending on which body parts were occluded. They found that occlusion of the arms had a more severe impact compared to occlusion of the legs. The experiments, conducted on PKU-MMD and NTU-RGB+D datasets, showed a clear degradation in performance under occlusion conditions. This work highlighted

the necessity of training models that can tolerate partial occlusion, rather than assuming complete skeleton sequences. Although the study did not propose a reconstruction method, it provided valuable insights into how occlusion influences HAR and motivated subsequent research in occlusion-robust modeling.

2.4 Regression approaches for Skeleton Reconstruction

Regression-based methods represent an earlier line of research for skeleton reconstruction under occlusion. One such study proposed a CRNN-based regression framework that directly predicted missing skeleton joints from available ones [3].

By leveraging temporal dependencies in sequences, the model attempted to recover occluded joints and improve downstream recognition. This regression-based approach marked the first attempt to address skeleton reconstruction under occlusion without adversarial learning. Although it demonstrated improvements over naive imputation, it suffered from limited realism, especially when large portions of the skeleton were missing. This limitation motivated the transition toward GAN-based frameworks, which could generate more realistic and coherent skeleton reconstructions.

2.5 Survey and Reviews on RGB-D based HAR

A comprehensive survey of RGB-D based HAR methods was presented in [4]. The authors systematically categorized approaches into RGB-based, depth-based, skeleton-based, and hybrid RGB-D based methods. For skeleton-based HAR, they further classified techniques according to the underlying models, including CNNs, RNNs, and hybrid architectures.

This review paper also summarized the challenges associated with different modalities, such as viewpoint variations, occlusions, and scalability to real-world applications. While not focused exclusively on occlusion, the survey provided an essential taxonomy of existing approaches and positioned occlusion handling as a critical future research direction.

2.6 Skeleton Encoding Approaches using CNNs

Several studies have focused on encoding skeleton data into image-like representations to exploit the powerful feature extraction capabilities of CNNs. These methods have achieved state-of-the-art results in HAR tasks, though their performance under occlusion remains a challenge.

2.6.1 Joint Trajectory Maps

In [5], skeleton sequences were encoded into Joint Trajectory Maps (JTMs), which transformed spatio-temporal information of joints into 2D representations. These JTMs were then fed into ConvNet classifiers. The approach demonstrated that encoding temporal dynamics into image-like structures allowed CNNs to capture discriminative features more effectively. However, the method assumed complete skeleton sequences and was not explicitly designed to handle missing joints due to occlusion.

2.6.2 Skeleton Optical Spectra

The Skeleton Optical Spectra (SOS) representation, introduced in [6], converted skeleton data into spectral images that encode both spatial and temporal dynamics. The CNN-based model trained on these images achieved accuracies above 94% on datasets such as MSRC-12, G3D, and UTD-MHAD. Despite its high performance, the method struggled with action pairs that involved subtle differences, and it lacked robustness under occlusion scenarios.

2.6.3 Joint Distance Maps

In [7], the authors proposed Joint Distance Maps (JDMs), which encoded pairwise distances between skeleton joints into image matrices. These were then processed by CNNs to recognize activities. The method demonstrated state-of-the-art performance on NTU-RGB+D and UTD-MHAD datasets and was shown to be robust against viewpoint variations. However, like other CNN encoding approaches, JDMs relied on full skeleton availability and were not inherently robust to occlusion.

2.6.4 Spectral Image Transformations

Another approach introduced Discrete Spectrum Transformations (DST) combined with CNNs to address viewpoint invariance [8]. By applying geometric rotation preprocessing, the method enhanced robustness to viewpoint changes and significantly improved cross-view recognition accuracy. Although effective for viewpoint variation, the method did not explicitly address occlusion and assumed complete skeleton sequences as input.

2.6.5 Skeleton Sequence to Image Transformation

In [9], a simple yet effective skeleton-to-image transformation was introduced, where entire skeleton sequences were converted into image representations and processed using CNNs. This end-to-end framework achieved remarkable results, including 100% accuracy on the Berkeley MHAD dataset and over 91% F1-score on ChaLearn. Despite its simplicity and effectiveness, the method was designed for complete data and did not incorporate mechanisms to reconstruct or infer occluded joints.

2.7 Models for Occlusion and Skeleton Reconstruction

Different models have contributed to the problem of skeleton occlusion and reconstruction in distinct ways:

- CNNs: Widely used to encode skeleton sequences into image-like representations (e.g., JTM, SOS, JDM). They excel at spatial feature extraction and action classification but assume complete skeleton availability. As such, CNNs are not robust against occlusion without additional reconstruction mechanisms.

- CRNNs: By combining CNN-based spatial encoding with RNN-based temporal modeling, CRNNs capture both joint relationships and sequential dynamics. They have been applied in regression-based skeleton reconstruction and as GAN generators, where they help predict or reconstruct missing joints.

- GANs: Introduced as the first adversarial learning framework for skeleton reconstruction under occlusion. The CRNN generator reconstructs missing joints, while the LSTM discriminator ensures temporal and structural realism. GANs significantly outperform regression methods in reconstructing realistic skeleton sequences.

- Regression Approaches: Represent early attempts to reconstruct missing skeleton data. While regression with CRNNs improved recognition performance, these approaches lacked the ability to generate realistic or diverse reconstructions, limiting their robustness under severe occlusion.

- Transformers: Recently emerging as powerful alternatives due to their ability to capture long-range dependencies across joints and frames. Transformers hold promise for handling complex occlusion scenarios more effectively than CRNNs, although their application to skeleton reconstruction is still in its infancy.

- LSTM Models: Frequently employed in skeleton-based HAR for temporal modeling. In reconstruction tasks, LSTMs are often used as discriminators in GAN frameworks, assessing the coherence and realism of generated skeleton sequences.

2.8 Summary and Research Gap

The reviewed literature highlights the significant progress made in skeleton-based HAR, as well as the limitations that remain in addressing occlusion. CNN-based methods, such as JTM, SOS, JDM, and sequence-to-image transformations, have demonstrated excellent performance when complete skeleton data is available. However, these methods are highly sensitive to missing joints, as they rely on dense spatial encodings that assume full-body information.

Regression-based approaches marked an important step forward, as they attempted to predict missing joints using CRNNs. While effective to some extent, these models lacked the ability to produce realistic reconstructions under severe occlusion. The introduction of GANs addressed this limitation by leveraging adversarial training to reconstruct skeletons that were not only accurate but also realistic and temporally coherent. Experimental results demonstrated significant improvements over regression, establishing GANs as a promising framework for skeleton reconstruction.

Despite these advancements, several gaps remain. Current GAN approaches are limited to controlled datasets and have not been extensively validated in real-world environments where occlusion patterns are more dynamic and unpredictable. Furthermore, while CRNNs have proven effective, they struggle with long-range dependencies. This opens an opportunity for Transformer models, which are better suited for capturing complex temporal relationships and may offer superior performance in skeleton reconstruction under occlusion. Additionally, most existing studies focus on partial occlusion of isolated joints; future work should consider

more complex occlusions caused by human-object interactions and multi-person scenarios.

These gaps directly motivate the focus of this thesis, which aims to explore advanced GAN and Transformer-based frameworks for robust skeleton reconstruction and HAR under occlusion. By addressing these challenges, the proposed work seeks to bridge the gap between controlled experimental setups and real-world applicability, ultimately improving the robustness and generalizability of HAR systems.

2.9 Comparison of Reviewed Studies

In this section, we have reviewed the literature and disscuss the literature details in the given Table 2.1 and dataset details which are used in all over the literature can be seen in the Table 2.2.

TABLE 2.1: Summary of Related Works on Human Activity Recognition and Occlusion Handling

Ref	**Authors / Year**	**Technique / Dataset**	**Remarks**	**Limitations**
[1]	Vernikos & Spyrou (2025)	GAN/ PKU-MMD, NTU-RGB+D, SYSU-3D-HOI, UTKinect-Action3D	Handles partial occlusion reconstruction using GANs; significant accuracy gain	Requires separate GAN training for each occlusion type; two body parts, lacks real-time adaptability
[2]	Giannakos, Spyrou, Mylonas (2021)	PKU-MMD, NTU-RGB+D	Highlights severe accuracy loss due to occlusion because it affects the HAR performance	Only uses manually simulated occlusions

Table 2.1 (continued)

Ref	Authors / Year	Technique / Dataset	Remarks	Limitations
[3]	YangLi, YiPan (2021)	LSTM, GRU on financial platforms	High ensemble model for forecasting	related to HAR or occlusion research and not more than one part occlusion
[4]	Wang et al. (2018)	CNN, RNN Survey / RGB-D modalities	Categorizes deep learning HAR techniques	specific focus on occlusion or skeleton reconstruction
[5]	Wang et al.(2016)	Joint Trajectory Maps/ MSRC-12, G3D,UTD-MHAD	Real-time ConvNet-based HAR using 2D skeleton images	Struggles with viewpoint sensitivity and visually similar actions
[6]	Hou et al. (2018)	Skeleton Optical Spectra (SOS) / MSRC-12, G3D, UTD-MHAD	Encodes skeleton motion into color texture images for CNNs	Fails to distinguish similar-looking actions; encoding choices affect results
[7]	Wang et al. (2017)	Joint Distance Maps (JDM) / NTU RGB+D, UTD-MHAD	Encodes pairwise joint distances for robust CNN-based recognition	Performance drops for similar joint-distance actions
[8]	Mathe et al. (2020)	3D skeleton → DFT-transformed images / PKU-MMD (ADLs)	Uses spectral domain transformations and geometric pre-processing	Performance depends on view alignment ann rotation changes, not tested under real occlusion

Table 2.1 (continued)

Ref	Authors / Year	Technique / Dataset	Remarks	Limitations
[9]	Li et al. (2016)	Skeleton Transformer CNN / NTU RGB+D, PKU-MMD	Introduces transformer module for focusing on key joints	Assumes full skeleton input; limited occlusion robustness
[10]	Kwon, T., Tekin, B (2022)	Context-aware sequence alignment using 4D skeletal augmentation	Preserves spatial-temporal info by converting sequence to image	CASA suffers from errors due to reliance on pose estimators

TABLE 2.2: Summary of Datasets Used in Skeleton-based Human Activity Recognition

Dataset	Type	Modalities	Viewpoints	Classes	Notes
PKU-MMD	Multi-view, Large-scale	RGB, Depth, IR, Skeleton	3 views	51	Used in occlusion and ADL-focused studies; Kinect V2 sensor used
NTU RGB+D	Multi-view, Large-scale	RGB, Depth, IR, Skeleton	3 views	60	Benchmark dataset, cross-subject and cross-view splits widely used
SYSU-3D-HOI	Single-view	Skeleton	Single view	12	Focused on human-object interaction

Table 2.2 (continued)

Dataset	Type	Modalities	Viewpoints	Classes	Notes
UTKinect-Action3D	Single-view	Skeleton	Single view	10	Compact dataset; used for basic motion and action classification
MSRC-12 Kinect	Single-view	Skeleton	Single view	12	Gesture dataset captured with Kinect V1; good for testing CNN encodings
G3D	Single-view	Skeleton	Single view	20	Contains gaming actions; used for JTMs and SOS studies
UTD-MHAD	Single-view	RGB, Depth, Inertial, Skeleton	Single view	27	Multi-modal dataset; CNN encodings like SOS, JDM used here
Berkeley MHAD	Multi-modal	Video, Audio, Skeleton	Multi-sensor setup	11	Used in CNN-based end-to-end recognition models

Chapter 3

Proposed Methodology

3.1 Overview of the Proposed Framework

The proposed architectures are designed to reconstruct complete 3D human skeleton sequences from partially occluded data by leveraging Generative Adversarial Networks (GANs). In GANS, there two neural networks, one is named as a Generator, in this we have used CRNN+ BiLSTM and Transformers models separately. While second one is the discriminator which is based on LSTM and is the same in both. The reconstructed sequences are later utilized for human behavior activities recognition. The framework comprises three primary components:

(i) Data preprocessing and occlusion simulation to prepare and simulate real-world data challenges.

(ii) Two GAN-based (CRNN, Transformer) skeleton reconstruction models were used to restore missing joint data. Model Diagram can be observed in Figure 3.1.

(iii) A behavior classification model to identify performed actions from reconstructed sequences. As shown in Figure 3.2.

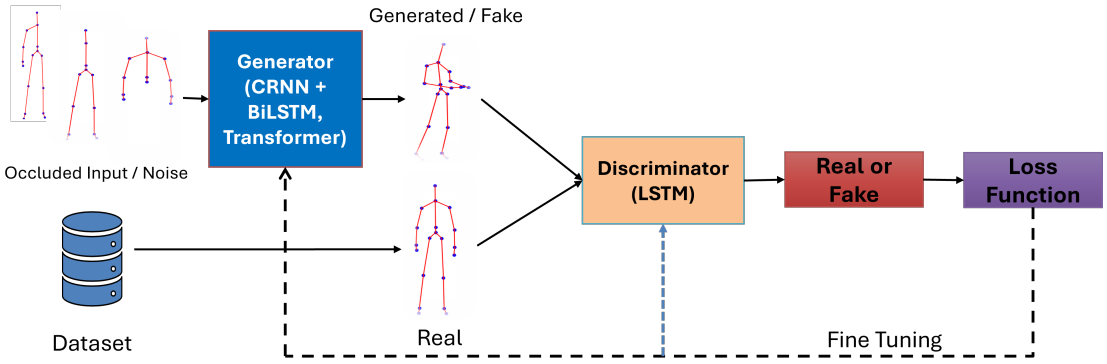

FIGURE 3.1: Generative Adversarial Network Model Diagram

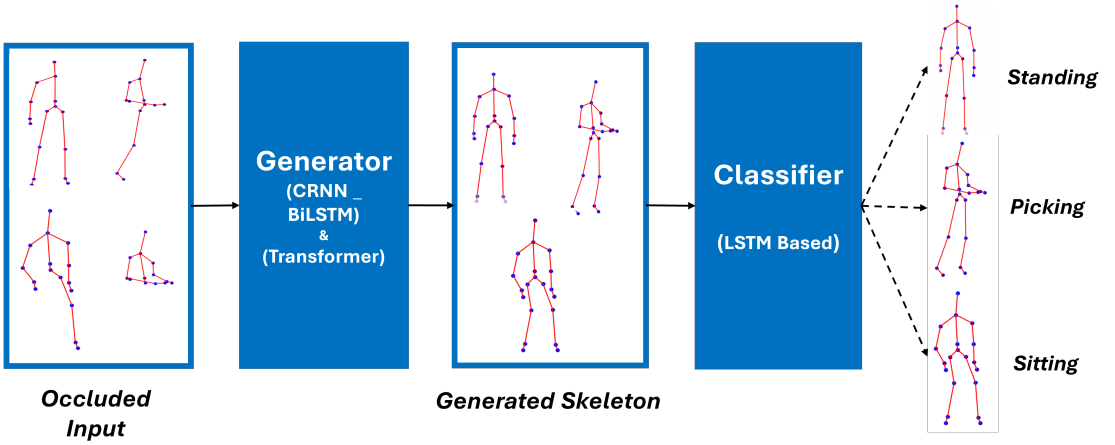

FIGURE 3.2: Methodology Block Diagram

Each component addresses a specific aspect of the problem, from handling noisy or incomplete skeleton data to learning spatio-temporal dependencies for robust action recognition.

The process begins with the extraction and formatting of the skeleton sequences to obtain a consistent representation. Various occlusion types, such as self-occlusion or sensor dropout, are simulated to emulate real-world scenarios. A conditional GAN architecture is employed to reconstruct missing joint data by learning temporal patterns from non-occluded training sequences. Finally, the reconstructed sequences are fed into a temporal classifier to detect and classify the performed actions accurately. The reconstruction of skeleton from GAN model can be observed in Figure 3.1 and overall process from occluded skeleton to the classification of activities can be observed in Figure 3.2.

FIGURE 3.3: RGB Images of Dataset

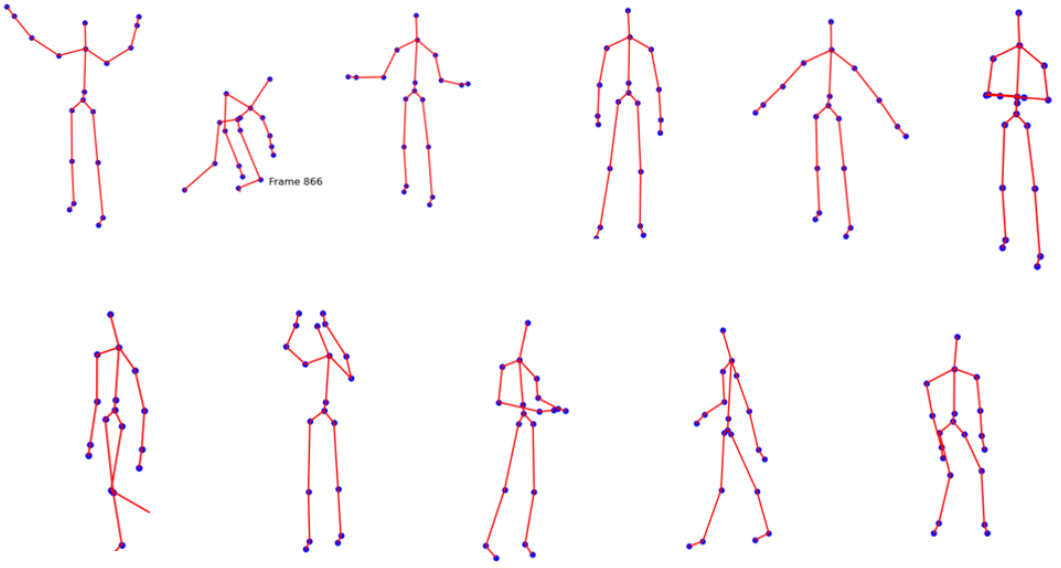

FIGURE 3.4: 3d Skeleton Dataset

3.2 Dataset Details

We have used UTKInect 3d dataset which is publicly available. This dataset consists of 10 actions and the type of data was RGB images as shown in Figure 3.3, Depth Images and 3d Skeleton joints indices values as shown in Figure 3.4.

3.3 Data Preprocessing

Skeletal data consist of 3D coordinates (x, y, z) for 20 joints captured in sequential frames. Each frame is represented as a 61-dimensional vector, where the first value denotes the frame number, and the remaining 60 values correspond to the (x, y, z) coordinates of the 20 joints. The preprocessing pipeline includes the following steps:

- Normalization: Joint coordinates are scaled using min-max normalization to ensure uniformity across different sequences and subjects.
- Sequence Formatting: Sequences are padded or interpolated to a fixed maximum duration, denoted as Tm, to maintain consistent input dimensions for the model.
- Occlusion Simulation: Specific joints are manually occluded through:
  - Zeroing out their coordinates to simulate complete data loss.
  - Introducing noise or masking to mimic realistic occlusion scenarios.
- Occlusion Cases: To test robustness, eight occlusion scenarios are simulated, I have given multiple figures which are representing the occlusion cases involving different activities. All are shown in Figures 3.5 to 3.8:
  - *Single-limb occlusions*: left arm, right arm, left leg, right leg
  - *Multi-limb occlusions*: both arms, both legs, left arm + leg, right arm + leg

  For each case, the corresponding occluded and real sequences are loaded, interpolated, and reshaped before training.

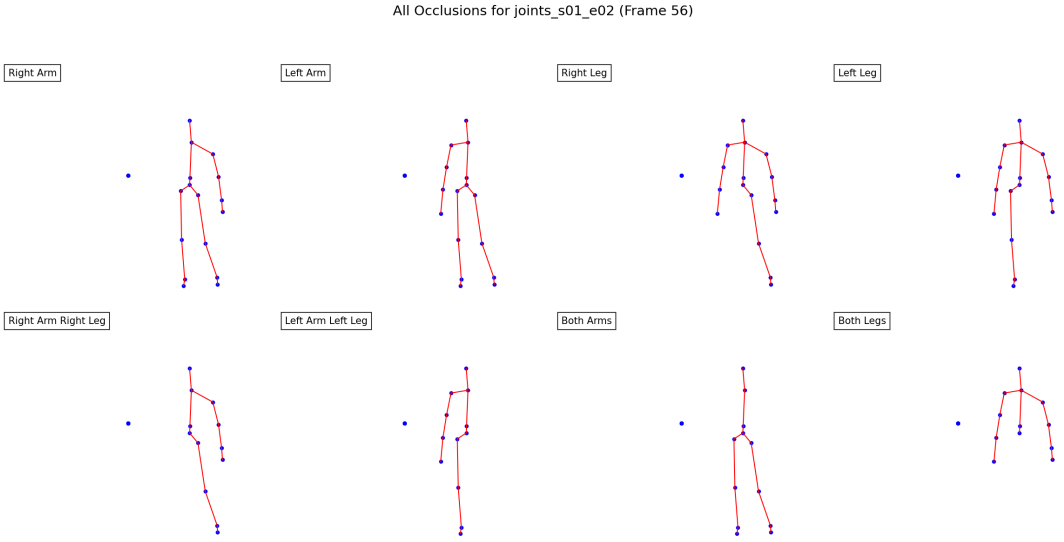

FIGURE 3.5: Occluded Cases- Standing

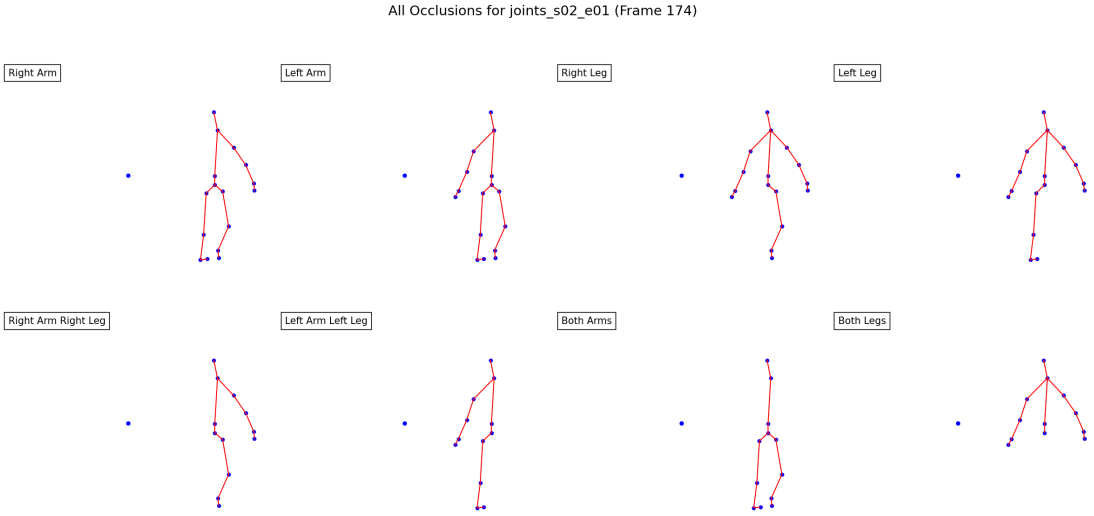

FIGURE 3.6: Occluded Case- Dancing

3.4 Proposed Generator Models

The proposed approach included two generetor models for skeleton reconstruction of occluded skeleton sequences. In the first method, Generator of GAN Model is based on CRNN for reconstructing partially occluded 3D human skeleton sequences. The architecture incorporates a CRNN-based generator and a Transformer-inspired discriminator trained with a multi-component loss function to generate realistic and temporally consistent skeleton reconstructions.

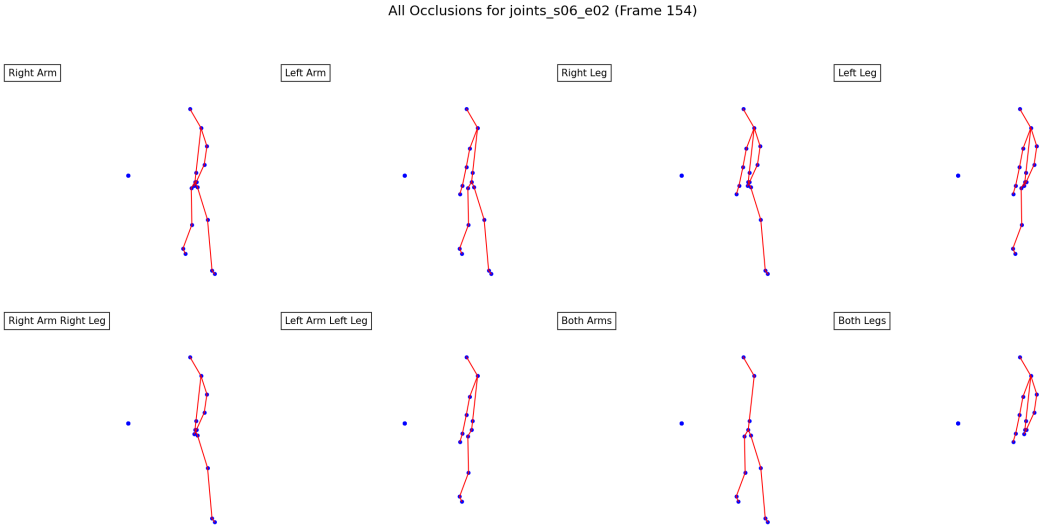

FIGURE 3.7: Occluded Case- Walking

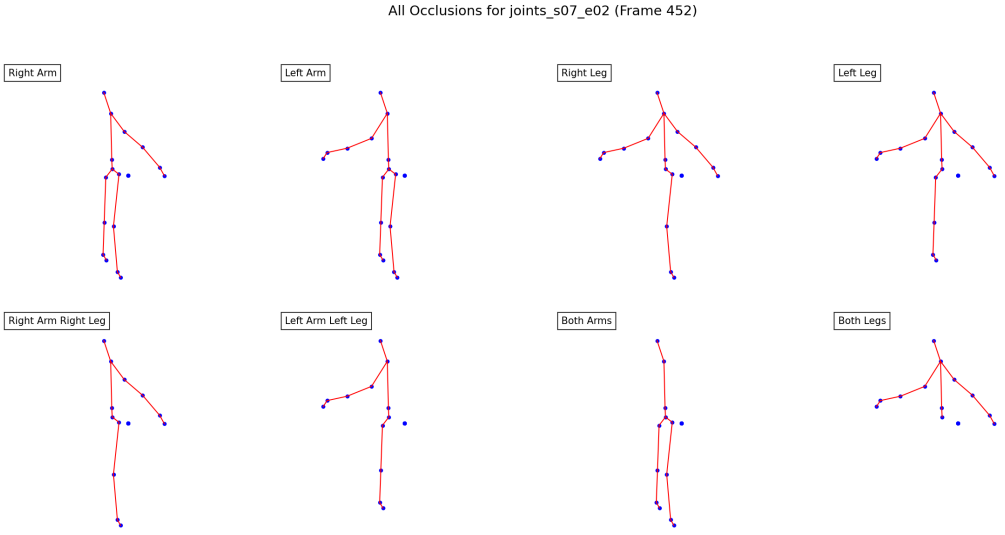

FIGURE 3.8: Occluded Case- Hugging

3.4.1 CRNN Based GAN Framework

The proposed GAN architecture consists of the following key components and can be seen in Figure 3.9.

- CRNN-based Generator: Reconstructs the full skeleton from the occluded sequence.
- LSTM based Discriminator: Differentiates between real and generated skeletons.

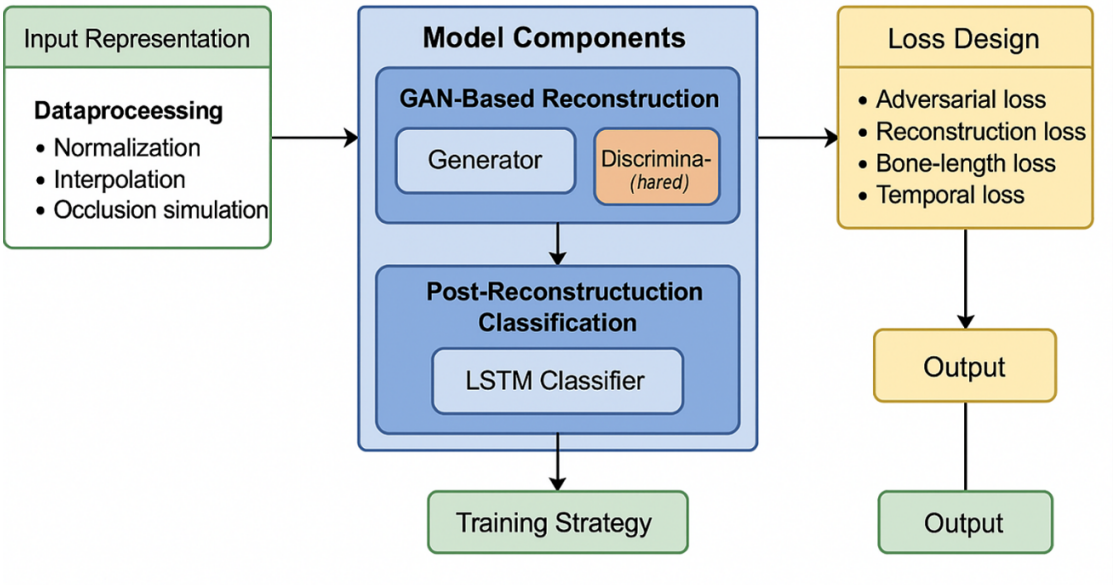

FIGURE 3.9: General GAN Implementation

- Composite Loss Function: Enforces anatomical correctness, temporal smoothness, and reconstruction fidelity.

Each model is trained per occlusion case (e.g., missing left arm, right leg), allowing the generator to specialize in handling specific types of occlusions.

3.4.2 Generator Architecture

The generator follows a Convolutional Recurrent Neural Network (CRNN) design. The input to the generator is a 3D skeleton sequences in which each sequence has an input of 61 float values where first one is the frame number and remaining 60 are built up with 20 joints consists of 3d x, y, z coordinates to represent skeleton joints indices in 3d spatial dimensions (x, y, z) and tensor of shape (Tm, J, 3), where Tm = 30. The architecture includes:

- Time Distributed Flattening: Applies flattening per time step.
- Two 1D Convolutional Layers: Extracts temporal features with kernel size 3 and *Leaky ReLU* activation.

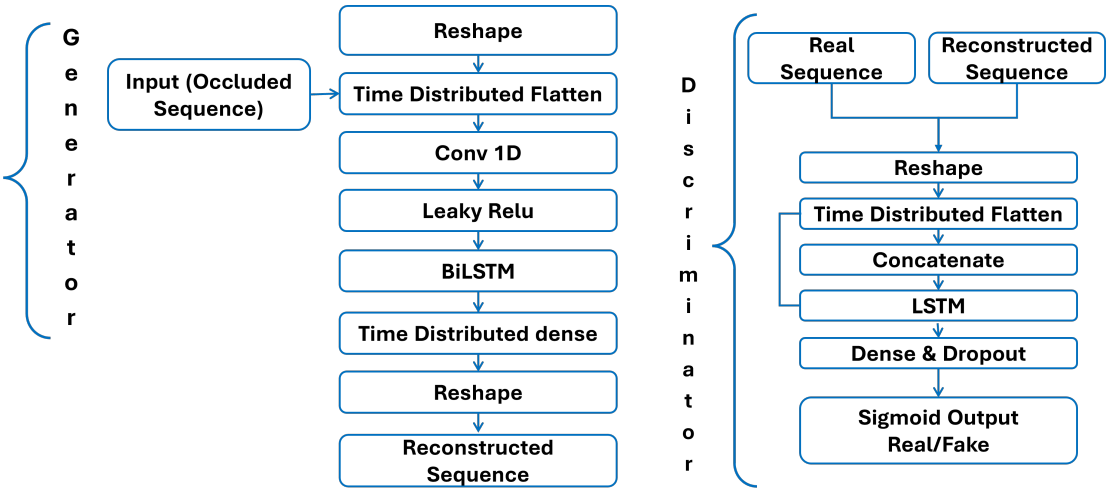

FIGURE 3.10: CRNN Model

- Bidirectional LSTM: Models forward and backward temporal dependencies using 256 hidden units.
- Time Distributed Dense: Projects outputs back to $(J \times 3)$ coordinates per frame.
- Reshaping Layer: Converts output back to $(T_m, J, 3)$.

Here, the detailed architecture of CRNN model can be observed in Figure 3.10.

3.4.3 Discriminator Architecture

The discriminator is designed to evaluate the realism of the generated skeleton compared to the ground truth. Its design can be observed in the Figure 3.10.

- Dual Input Layers: Receives both the real and generated skeleton sequences.
- Flattening and Concatenation: Flattens and merges the real and generated sequences.
- Fully Connected Layers: Uses dense layers with ReLU activation and dropout for regularization.
- Output Layer: A sigmoid activation function outputs the probability that the generated sequence is real.

3.4.4 Loss Function Design

In the proposed CRNN-based GAN architecture, the generator is trained using a composite loss function that incorporates multiple components to enforce spatial consistency, temporal coherence, and adversarial realism. These loss functions guide the generator to reconstruct occluded skeleton sequences that are not only numerically accurate, but also anatomically plausible and temporally smooth.

3.4.4.1 Adversarial Loss

To align with the GAN framework, an adversarial loss is incorporated to ensure the generator produces outputs indistinguishable from real data. The binary cross-entropy loss is used between the discriminator's prediction and the ground-truth label (real = 1):

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{(Y_{\text{real}}, Y_{\text{gen}})} \left[-\log D(Y_{\text{real}}, Y_{\text{gen}}) \right] \tag{3.1}$$

This encourages the generator to synthesize realistic skeletons that can fool the discriminator.

3.4.4.2 Reconstruction Loss

To minimize the numerical difference between the reconstructed and ground-truth skeletons, the Mean Absolute Error (MAE) is used:

$$\mathcal{L}_{\text{MAE}} = \mathbb{E} \left[\| Y_{\text{real}} - Y_{\text{gen}} \|_1 \right] \tag{3.2}$$

This term ensures per-joint coordinate accuracy between generated and actual skeleton sequences.

3.4.4.3 Bone-Length Consistency Loss

To preserve anatomical structure, a bone-length consistency loss is employed. It compares the Euclidean distances between pairs of connected joints (bones) in both the real and generated sequences:

$$\mathcal{L}_{\text{bone}} = \sum_{(i,j)\in \text{BONES}} \mathbb{E}\left[\left(\|Y^i - Y^j\| - \|\hat{Y}^i - \hat{Y}^j\|\right)^2\right] \tag{3.3}$$

This term penalizes unnatural deformations in the skeleton by enforcing bone-length preservation.

3.4.4.4 Temporal Smoothness Loss

To avoid jitter and abrupt transitions between frames, the temporal smoothness loss penalizes discrepancies in joint velocity across time steps:

$$\mathcal{L}_{\text{temporal}} = \mathbb{E}\left[\left\|(Y_t - Y_{t-1}) - (\hat{Y}_t - \hat{Y}_{t-1})\right\|^2\right] \tag{3.4}$$

This term helps produce smooth and realistic motion sequences over time.

3.4.4.5 Total Generator Loss

The overall loss used to optimize the generator is a Pix2Pix-inspired weighted combination of the above components. In the implementation, all reconstruction-related losses (MAE, bone-length, and temporal smoothness) are grouped and scaled by a common factor λ, and combined with the adversarial loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \lambda \cdot (\mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{bone}} + \mathcal{L}_{\text{temporal}}) \tag{3.5}$$

Where $\lambda = 50$ is an empirically chosen hyperparameter that balances the generator's focus between realism (via adversarial loss) and accuracy (via reconstruction losses).

This total loss ensures that the generated skeleton sequences are both perceptually realistic and structurally accurate with respect to the real motion data.

3.4.5 Training Protocol

The training process is divided into two stages:

- Pretraining Generator: The generator is first trained independently using L1 loss for 30 epochs to stabilize its output.
- Adversarial Training: The full GAN is trained for 150 epochs using a batch size of 10. The discriminator and generator are jointly optimized using binary cross-entropy and the composite loss.

An Adam optimizer with a learning rate of 1×10^{-4} is used for both training phases.

3.4.6 Experimental Pipeline

The experimental loop performs the following steps per occlusion case:

1. Load real and occluded skeleton sequences.
2. Preprocess and interpolate all sequences to a fixed length.
3. Train the generator and the discriminator.
4. Evaluate training, validation, and test splits.
5. Store results in tabular form for cross-case comparison.

The resulting performance metrics are compiled into a comprehensive table summarizing the GAN's effectiveness across all occlusion conditions.

3.5 Transformer-Based Generator Architecture

The Transformer-based generator is designed to effectively reconstruct missing or occluded joint sequences in human skeletal data. Unlike conventional recurrent models such as LSTMs or GRUs, the Transformer leverages self-attention mechanisms to model long-range temporal dependencies in skeleton motion sequences. This is particularly advantageous in the context of human motion reconstruction, where occluded joint trajectories often depend on spatial and temporal cues across the entire sequence.

3.5.1 Input Representation

The input to the generator consists of an occluded 3D skeleton sequence of temporal length Tm = 30, with each frame containing 20 joints in three dimensions (x, y, z), resulting in an input shape of (30, 60). Before feeding into the Transformer, the data is passed through a dense layer to project it to a higher- dimensional embedding space (typically 256 dimensions), enabling the model to represent complex patterns in motion data.

3.5.2 Positional Encoding

Since Transformers lack inherent awareness of sequence order, a learnable *Positional Encoding* is added to the embedded input. This encoding injects temporal order into the input sequence, allowing the model to distinguish between frames and understand motion continuity across time.

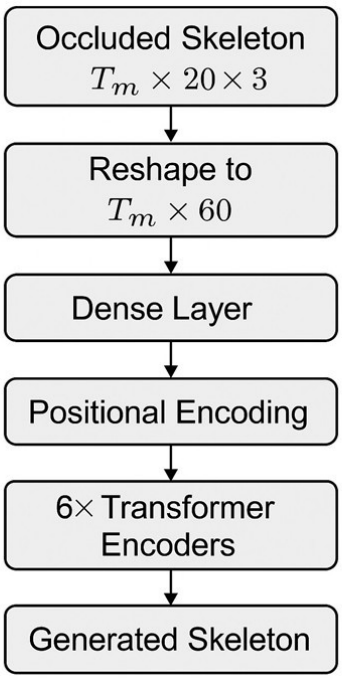

FIGURE 3.11: Transformer Based - GAN

3.5.3 Transformer Encoder Layers

The generator comprises multiple stacked Transformer encoder blocks (typically six). The Transformer can be seen in Figure 3.11 and each encoder block includes:

- Multi-Head Attention: This layer enables the model to attend to different parts of the sequence simultaneously, capturing both local and global dependencies.
- Feed-Forward Network: A two-layer MLP that processes the output of the attention layer to enhance non-linear representations.
- Residual Connections and Layer Normalization: Used after each sub-layer to stabilize training and preserve gradient flow.

These layers collectively allow the generator to infer missing joint trajectories by effectively attending to context from unoccluded frames in both the past and future.

3.5.4 Post-Transformer Processing

Following the encoder stack, the output is passed through a 1D convolutional layer to refine temporal features and then through a fully connected layer to regress the final output in the original joint-space dimension. Finally, the output is reshaped back to $(30, 20, 3)$ to represent the reconstructed 3D skeleton sequence.

3.5.5 Integration into GAN Framework

The generator is trained adversarially within a GAN setup. It is paired with a PatchGAN-style discriminator, which distinguishes between real (non-occluded) and generated (reconstructed) skeleton sequences. The training process optimizes a composite objective function comprising:

- Adversarial Loss: Encourages the generator to produce realistic skeletons.
- L1 Loss: Penalizes absolute reconstruction error between predicted and real sequences.
- Bone Length Loss: Preserves anatomical correctness by maintaining consistent bone lengths across frames.
- Temporal Smoothness Loss: Encourages coherent motion dynamics across time by penalizing unnatural frame-to-frame changes.

Detailed architecture of Transformer model be observed in Figure 3.12.

3.5.6 Loss Function Design for Transformer-Based GAN

To train the Transformer-based generator in the context of 3D skeleton reconstruction, a composite loss function is employed. This design ensures that the generated sequences are not only numerically accurate but also biomechanically

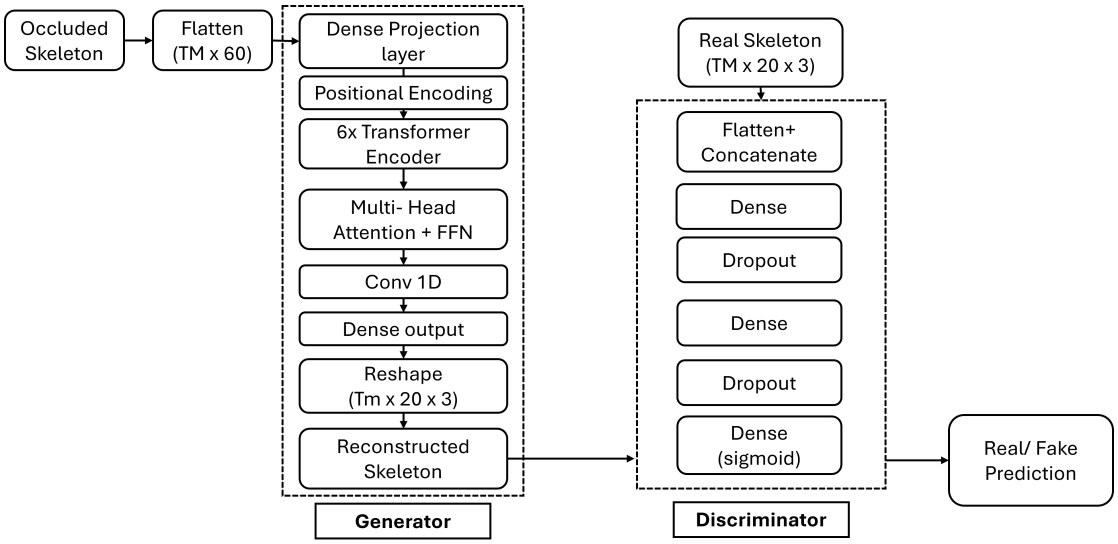

FIGURE 3.12: Detailed Transformer Model Architecture

plausible and temporally coherent. The final loss consists of three main components: L1 reconstruction loss, bone-length consistency loss, and temporal smoothness loss. These are combined into a single objective function used to supervise the generator during GAN training.

3.5.6.1 L1 Reconstruction Loss

The reconstruction loss measures the mean absolute error between the predicted and ground-truth skeletons:

$$\mathcal{L}_{\text{MAE}} = \mathbb{E}\left[\|Y_{\text{real}} - Y_{\text{gen}}\|_1\right] \tag{3.6}$$

This loss penalizes joint-wise differences between the real and generated coordinates. It encourages the generator to produce outputs that are numerically close to the ground truth for every joint and frame.

3.5.6.2 Bone-Length Consistency Loss

This term enforces anatomical realism by preserving the distances (i.e., lengths) between physically connected joints:

$$\mathcal{L}_{\text{bone}} = \sum_{(i,j)\in \text{BONES}} \mathbb{E}\left[\left(\|Y^i - Y^j\| - \|\hat{Y}^i - \hat{Y}^j\|\right)^2\right] \tag{3.7}$$

For each bone defined as a pair of joints (i, j), this loss penalizes the difference in bone lengths between the real and generated skeletons. This helps ensure anatomical plausibility in the reconstructed output.

3.5.6.3 Temporal Smoothness Loss

To promote motion continuity and reduce jitter across frames, temporal smoothness loss penalizes differences in frame-to-frame joint displacement:

$$\mathcal{L}_{\text{temporal}} = \mathbb{E}\left[\left\|(Y_t - Y_{t-1}) - (\hat{Y}_t - \hat{Y}_{t-1})\right\|^2\right] \tag{3.8}$$

This loss encourages the generator to maintain consistent motion patterns over time by matching the velocity of joints between real and generated sequences.

3.5.6.4 Total Generator Loss

The total generator loss combines all three terms using predefined weighting coefficients λ, λ_{bone}, and $\lambda_{\text{temporal}}$:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{MAE}} + \lambda_{\text{bone}} \cdot \mathcal{L}_{\text{bone}} + \lambda_{\text{temporal}} \cdot \mathcal{L}_{\text{temporal}} \tag{3.9}$$

In the implementation, these weights are set as follows: $\lambda = 100$, $\lambda_{\text{bone}} = 10$, and $\lambda_{\text{temporal}} = 5$.

3.5.6.5 Adversarial Learning Context

Although the `combined_loss` is applied specifically to the generator's output in the GAN training loop, the full GAN model also includes an adversarial discriminator

trained using binary cross-entropy. The overall loss passed to the GAN model during compilation is:

$$\mathcal{L}_{\text{GAN}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{total}} \tag{3.10}$$

Here, $\mathcal{L}_{\text{BCE}}$ refers to the binary cross-entropy loss from the discriminator, which learns to differentiate real from generated skeleton sequences. $\mathcal{L}_{\text{total}}$ is the generator's composite loss, defined as a weighted sum of reconstruction, bone-length, and temporal smoothness losses. This adversarial learning setup encourages the generator to produce skeleton sequences that are not only structurally accurate and temporally smooth but also indistinguishable from real data in the discriminator's perspective.

Summary

This combination of data fidelity, structural accuracy, and temporal regularization ensures that the generator produces high-quality skeleton sequences that are realistic, smooth, and biologically plausible. The Transformer-based architecture, when trained with this carefully designed loss function, is well-suited for handling long-range dependencies in temporal skeleton data.

3.6 Post-Reconstruction Classification

After reconstruction, the resulting skeleton sequences are used for behavior classification. The classifier is a stacked LSTM model with the following specifications:

- Input: Reconstructed 3D joint sequences.
- Hidden Layers: LSTM units with LeakyReLU activation to model temporal patterns.
- Output: A softmax classification layer over 10 behavior classes.

- Dropout: A dropout rate of 0.25 is applied to prevent overfitting.
- Training: The classifier is trained exclusively on real (non-occluded) sequences to ensure it learns accurate action patterns.

3.7 Evaluation Protocols

The performance of both reconstruction quality and classification accuracy is assessed using the following protocols:

- Baseline: Direct classification on occluded sequences without reconstruction.
- Reference: Classification on real (non-occluded) sequences.
- Regression: Classification on sequences reconstructed using interpolation or regression-based models.
- GAN- CRNN+BiLSTM: Classification on sequences reconstructed by the GAN- CRNN+BiLSTM framework.
- GAN- Transformer: Classification on sequences reconstructed by the GAN-Transformer framework.

The evaluation metrics include:

- Reconstruction Metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE), calculated frame-wise.
- Classification Metrics: Weighted Accuracy (WAcc) and per-action performance metrics.

3.8 Summary

This chapter outlined the methodology for reconstructing occluded human skeleton sequences using GANs and classifying behaviors from the reconstructed sequences.

By integrating temporal modeling with adversarial learning, the system effectively restores realistic motion patterns and maintains high recognition accuracy, even in the presence of significant missing data.

Chapter 4

Results and Discussions

This chapter presents the results of the proposed GAN models for 3D human skeleton reconstruction under partial occlusion. All the results like from reconstruction of skeleton to the classification of activities. In recostruction of skeleton, we have proposed two separate deep learning models which are used under the Generator of GAN i.e, CRNN+ BiLSTM based generator and Transformer based Genrator. Both gives the enhanced results as compared to the previous results which was given in the paper which we have followed.

Moreover, It includes the model training, evaluation metrics, reconstructed and classification results and performance analysis across various occlusion scenarios. Comparison between real, occluded, and generated skeletons plots are shown of our proposed models separately. Meanwhile, quantitaive results are also shown.

4.1 Training Configuration Comparison

The proposed models follow a GAN training approach with a generator and discriminator but configurations are slightely changed such as optimizer settings, loss components, and training strategy.

Table 4.1 summarizes the key configuration differences between the two proposed models i.e. one is GAN based on CRNN and LSTM, second Transformer based GAN model .

TABLE 4.1: Training Configurations: CRNN–based vs Transformer–based GAN

Configuration	**CRNN–based GAN**	**Transformer–based GAN**
Epochs	150	150
Batch Size	10	10
Optimizer (G and D)	Adam (lr = 0.001)	Adam (lr = 0.0001)
Normalization	**Yes** (MinMaxScaler $[-1, 1]$)	**No**
GAN Loss Function	BCE + L1 + Bone + Temporal	BCE + MAE + Bone + Temporal
Loss Weights (λ)	$\lambda = 50$ (L1)	$\lambda = 100, 10, 5$ (MAE, Bone, Temporal)
Dropout Rate	0.25	0.3
Training Strategy	Manual GradientTape loop	Keras `.fit()` per epoch
GAN Labels	1 (real), 0 (fake)	1 (real), 0 (fake)
Evaluation Metrics	MSE, MAE, Weighted Accuracy	MSE, MAE, Weighted Accuracy

4.2 Training and Inference Flow

The GAN models are trained using only complete (non-occluded) skeleton sequences. The flow is summarized below:

4.2.1 Training Phase

1. For input, we have full skeleton sequences $\mathcal{S}_{real}$.

2. The Generator (G) learns to reconstruct $\mathcal{S}_{real}$ from the occluded or noisy versions.

3. The Discriminator (D) receives pairs of real and generated sequences and tries to classify them as real or fake.

4. The loss function includes adversarial (BCE), L1 (or MAE), bone length, and temporal smoothness terms.

5. Both G and D are trained in alternating steps for 150 epochs.

4.2.2 Inference Phase

1. After training, G is used independently.

2. Occluded sequences $\mathcal{S}_{occ}$ are input to G.

3. G reconstructs missing joints, producing $\mathcal{S}_{gen}$.

4. $\mathcal{S}_{gen}$ is compared against ground truth $\mathcal{S}_{real}$ to compute evaluation metrics.

4.3 Comparative Model - Reconstruction Results

4.3.1 Baseline

Trained and tested only on non-occluded skeleton samples.

As shown in Figure 4.1.

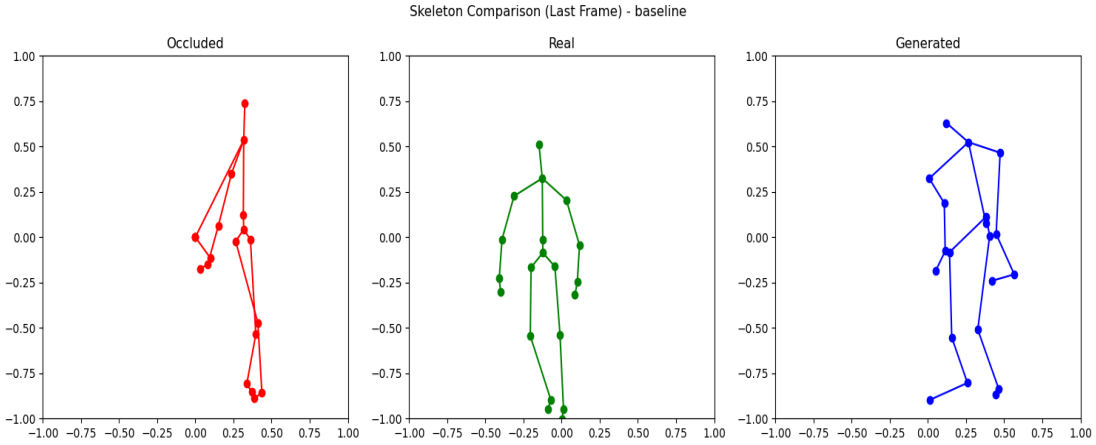

FIGURE 4.1: Baseline Right Arm

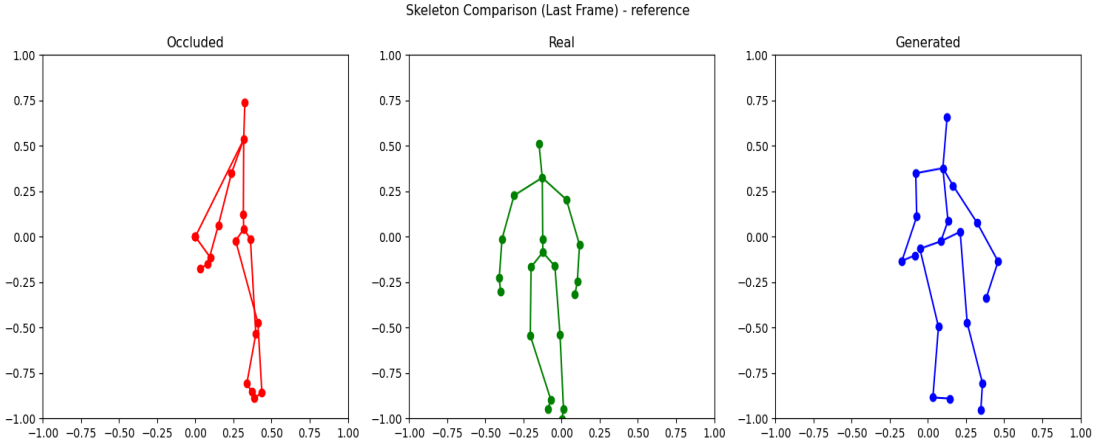

FIGURE 4.2: Reference Right Arm

4.3.2 Reference

Trained on non-occluded samples but tested on occluded skeleton samples. As shown in Figure 4.2.

4.3.3 Augmented

Trained on both original non-occluded data and synthetically occluded (augmented) skeletons, then tested on occluded samples.

As shown in Figure 4.3.

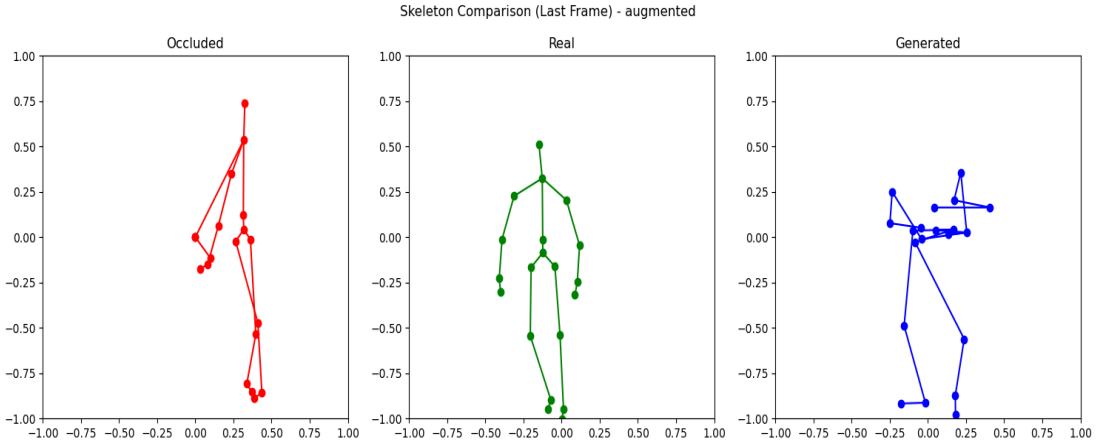

FIGURE 4.3: Augmented Right Arm

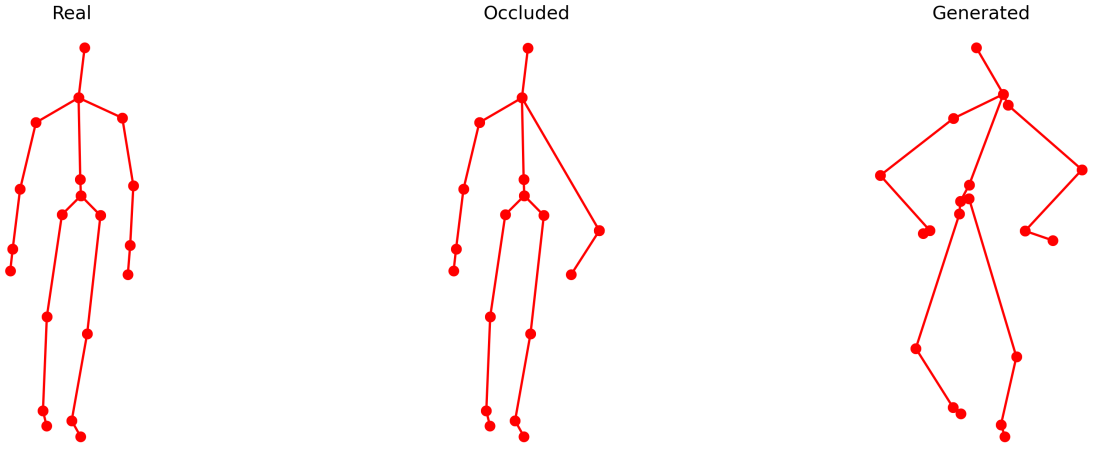

FIGURE 4.4: CRNN Based Right Arm

4.3.4 CRNN Based Reconstructions

In the following section, the reconstruction of occluded skeletons from the CRNN-based GAN network is shown. The results of different occlusion cases are presented in following Figures 4.4 to 4.11.

4.3.5 Transformer Based Reconstructions

In the following section, the reconstruction of occluded skeletons from the Transformer based GAN network has been showed and can be observed in Figures 4.12 to 4.17.

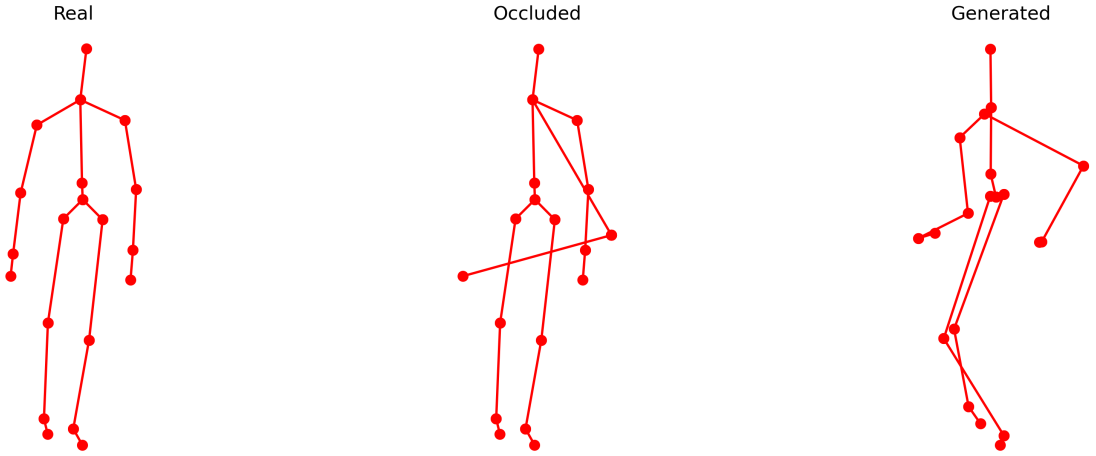

FIGURE 4.5: CRNN Based Left Arm

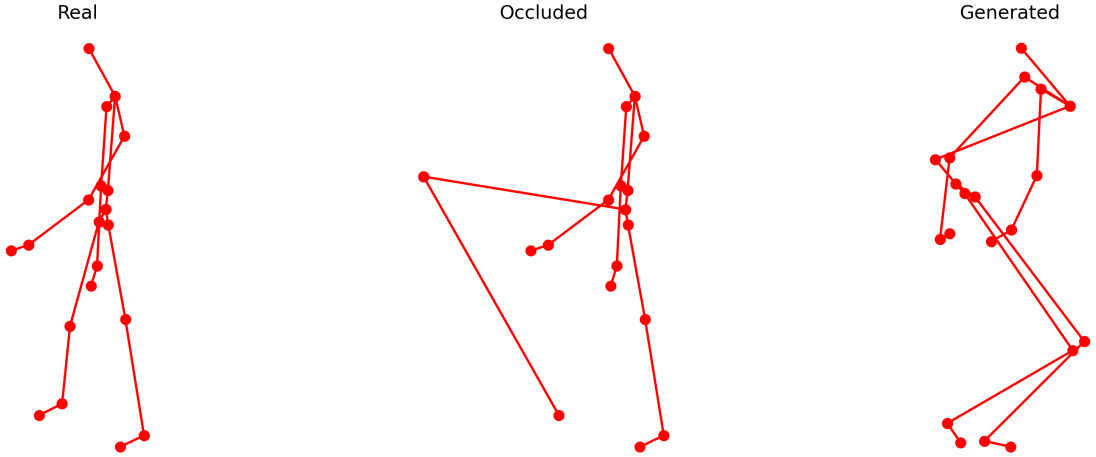

FIGURE 4.6: CRNN Based Right Leg

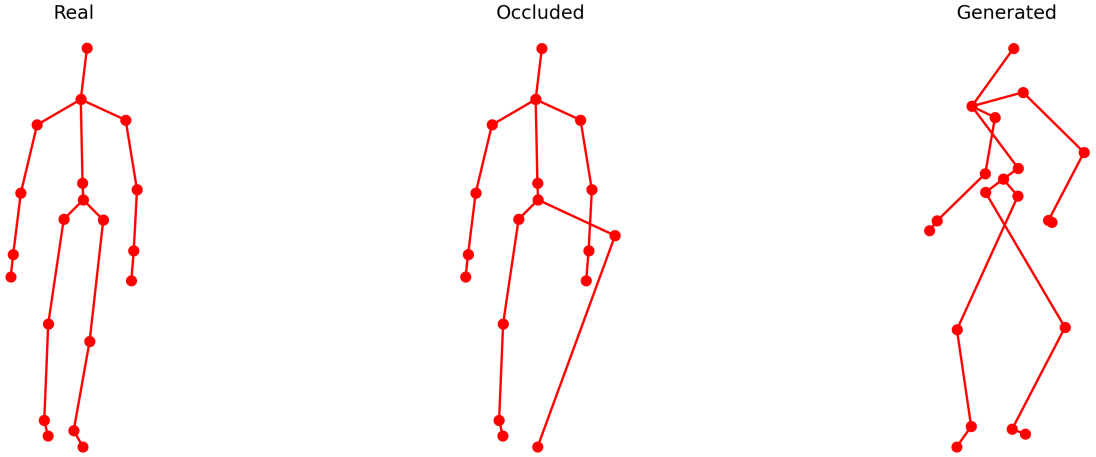

FIGURE 4.7: CRNN Based Left Leg

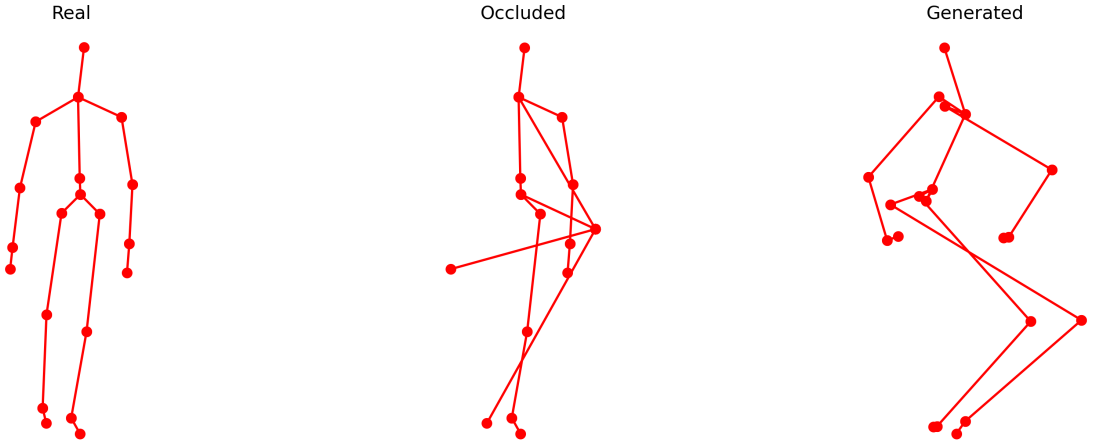

FIGURE 4.8: CRNN Based Left Arm and Left Leg

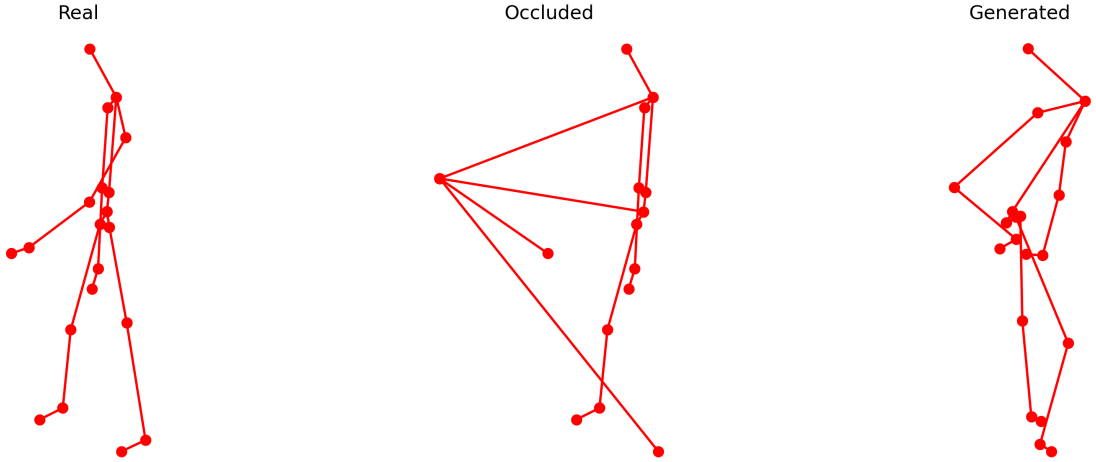

FIGURE 4.9: CRNN Based Right Arm and Right Leg

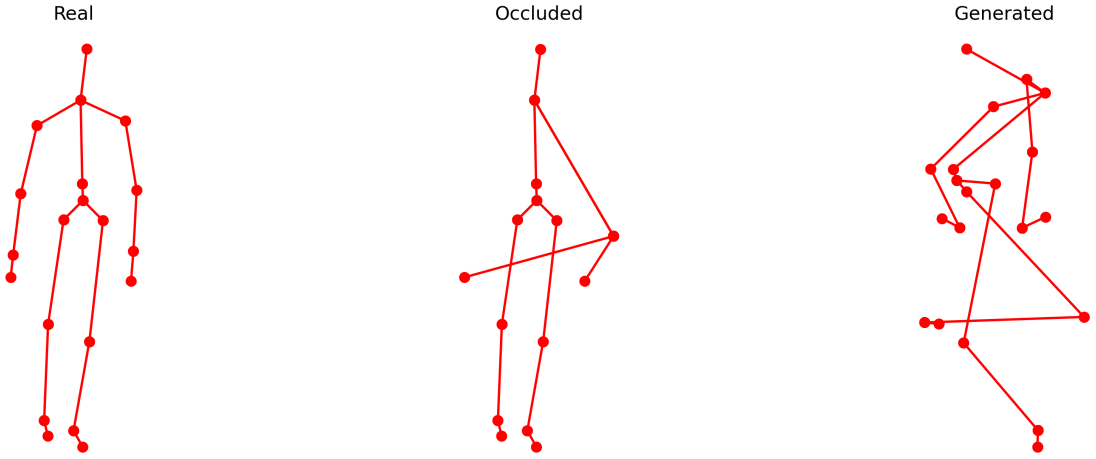

FIGURE 4.10: CRNN Based Both Arms

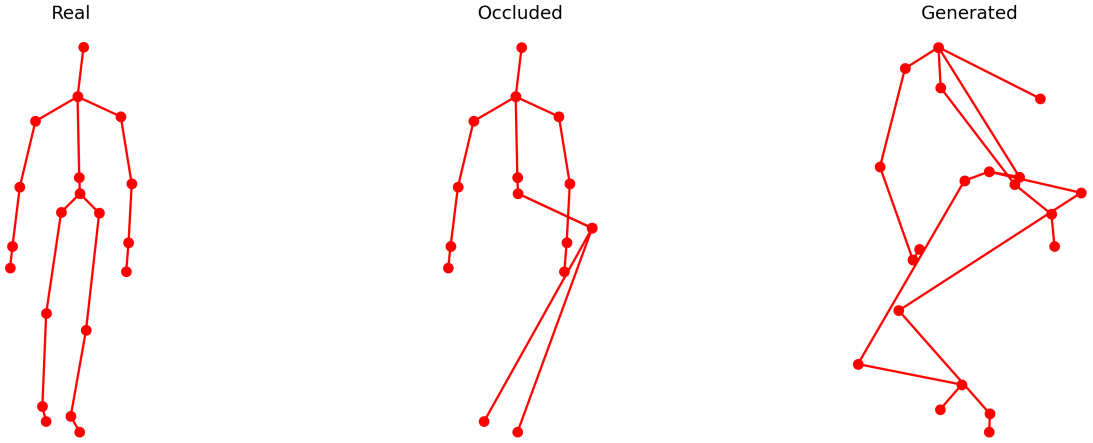

FIGURE 4.11: CRNN Based Both legs

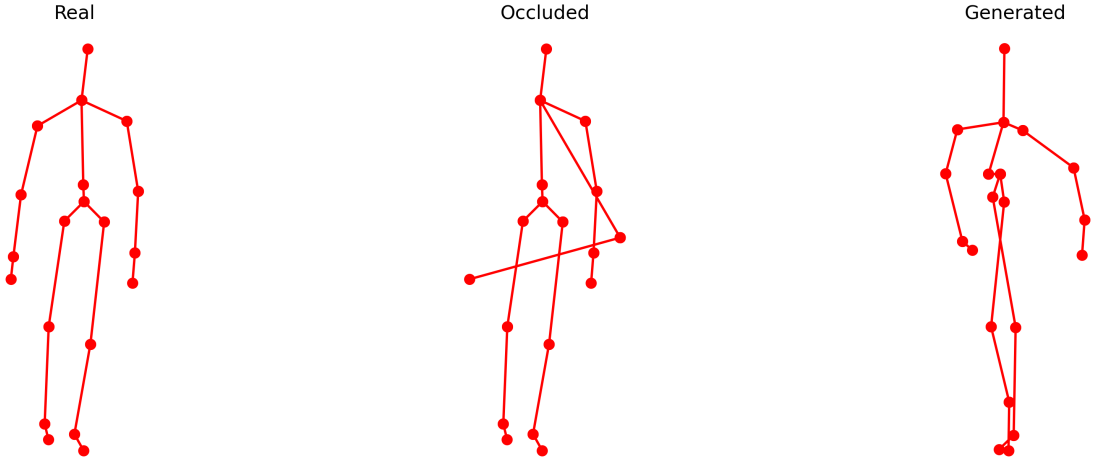

FIGURE 4.12: Transformer Based Left Arm

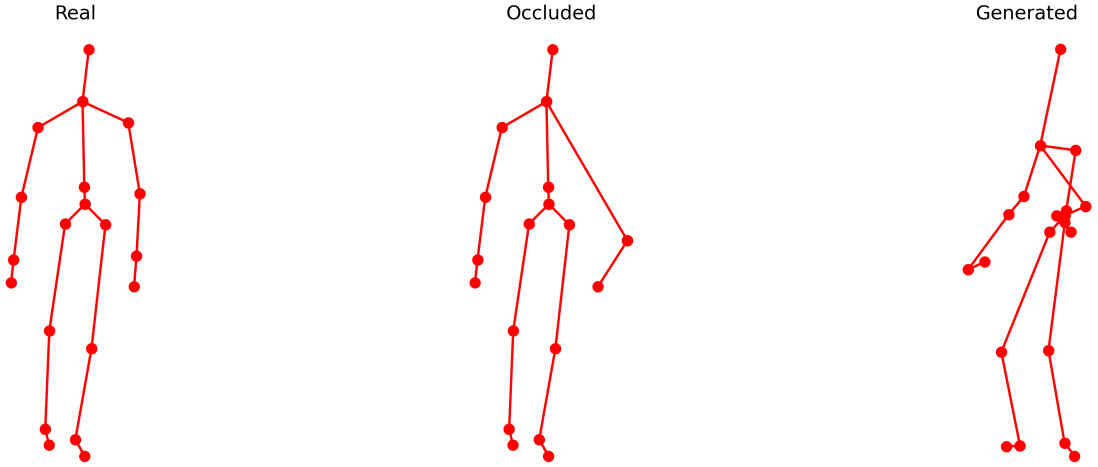

FIGURE 4.13: Transformer Based Right Arm

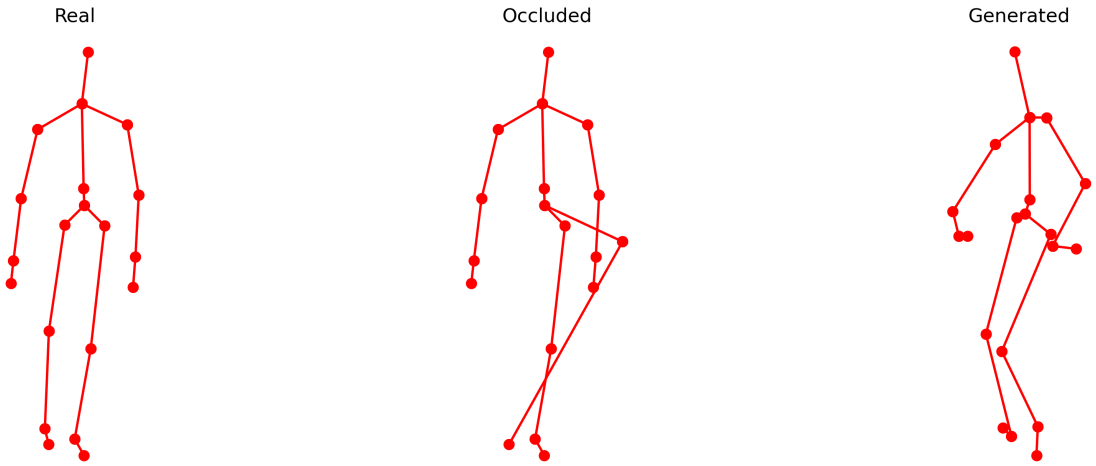

FIGURE 4.14: Transformer Based Left Leg

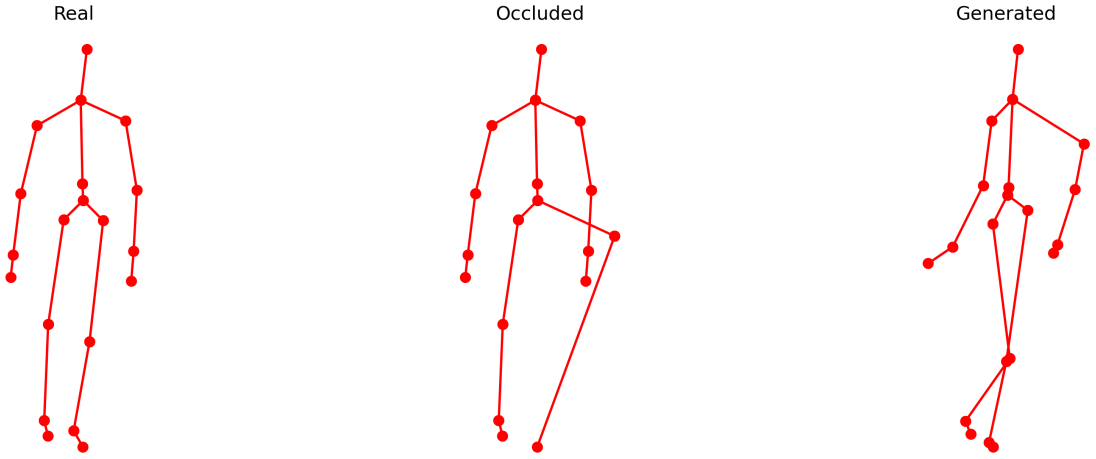

FIGURE 4.15: Transformer Based Right Leg

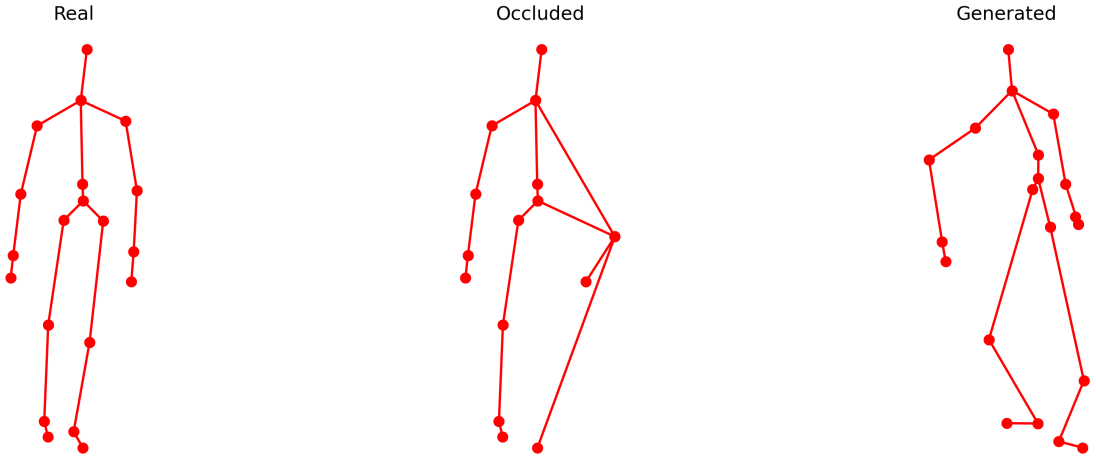

FIGURE 4.16: Transformer Based Right Arm and Right Leg

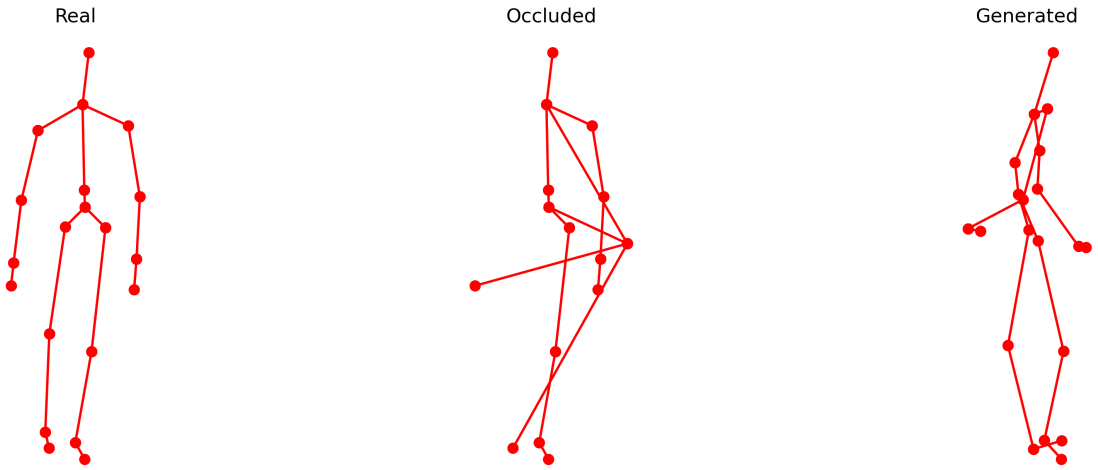

FIGURE 4.17: Transformer Based Left Arm and Left Leg

4.4 Comparative Model - Quantitative Results

In this section, we present the results of our proposed model. MAE (Mean Absolute Error) was used to measure the error between the joints of the reconstructed skeleton and the joints of the real skeleton.

Similarly, MSE (Mean Squared Error) was used to measure the error between the reconstructed frame with the full reconstructed skeleton and the real skeleton.

4.4.1 Mean Absolute Error Results

Mean Absolute Error used to measure the distance between the reconstructed joint position as compared to the the real joint position. As shown in Table 4.2.

TABLE 4.2: Reconstruction Error: GAN-CRNN vs GAN-Transformer

Occlusion Case	**GAN-CRNN**			**GAN-Transformer**		
	Train	**Val.**	**Test**	**Train**	**Val.**	**Test**
left_arm	0.0974	0.1161	0.1093	**0.0691**	**0.0874**	**0.0811**
right_arm	**0.0786**	**0.0967**	**0.0882**	0.0736	0.1013	0.0910
both_arms	0.1225	0.1667	0.1449	**0.0681**	**0.0945**	**0.0806**
left_leg	0.1037	0.1336	0.1201	**0.0664**	**0.0921**	**0.0795**

Table 4.2 (continued)

Occlusion Case	GAN-CRNN			GAN-Transformer		
	Train	**Val.**	**Test**	**Train**	**Val.**	**Test**
right_leg	0.0865	0.1112	0.1001	**0.0623**	**0.0851**	**0.0748**
both_legs	0.0705	0.1109	0.0981	**0.0642**	**0.0883**	**0.0790**
left_arm_leg	0.0967	0.1175	0.1086	**0.0665**	**0.0892**	**0.0777**
right_arm_leg	0.1464	0.1907	0.1715	**0.0778**	**0.1035**	**0.0927**

It is easy to understand that our proposed models performed well while Transformer had low error which was quite helpful in classification activities.

4.4.2 Mean Squared Error Results

Mean Squared Error, which was used to measure the error between the reconstructed frame of full reconstructed skeleton and the real skeleton. Like it is the error between the whole frames. As shown in Table 4.3.

TABLE 4.3: MSE: GAN-CRNN vs GAN-Transformer

Occlusion Case	GAN-CRNN			GAN-Transformer		
	Train	**Val.**	**Test**	**Train**	**Val.**	**Test**
left_arm	0.0184	0.0240	0.0212	**0.0099**	**0.0151**	**0.0121**
right_arm	**0.0128**	**0.0166**	**0.0147**	0.0101	0.0178	0.0127
both_arms	0.0294	0.0475	0.0377	**0.0097**	**0.0168**	**0.0122**
left_leg	0.0189	0.0276	0.0245	**0.0093**	**0.0165**	**0.0124**
right_leg	0.0154	0.0215	0.0181	**0.0079**	**0.0147**	**0.0102**
both_legs	0.0116	0.0194	0.0169	**0.0094**	**0.0134**	**0.0114**
left_arm_leg	0.0168	0.0236	0.0201	**0.0096**	**0.0151**	**0.0117**
right_arm_leg	0.0387	0.0574	0.0490	**0.0117**	**0.0182**	**0.0150**

4.4.3 Weighted Accuracy Results

Weighted Accuracy (WAcc) measures skeleton reconstruction by giving more importance to key joints like hips and shoulders, while down-weighting less critical ones such as wrists or ankles. This makes the evaluation more realistic, as it reflects how well the overall body structure and motion are preserved. In this table, we have GAN - CRNN and GAN - Transformer results with train (80%), validate (10%), test (10%) As shown in Table 4.4.

TABLE 4.4: Comparison of WAcc: GAN-CRNN vs GAN-Transformer

Occlusion Case	**GAN-CRNN**			**GAN-Transformer**		
	Train	**Val.**	**Test**	**Train**	**Val.**	**Test**
left_arm	0.7773	0.7191	0.7459	**0.7982**	**0.7514**	**0.7755**
right_arm	**0.8124**	**0.7575**	**0.7804**	0.7853	0.7411	0.7576
both_arms	0.7386	0.6678	0.7110	**0.8039**	**0.7446**	**0.7742**
left_leg	0.7620	0.7104	0.7362	**0.8143**	**0.7667**	**0.7903**
right_leg	0.7850	0.7332	0.7555	**0.8231**	**0.7558**	**0.7866**
both_legs	0.8032	0.7339	0.7601	**0.8120**	**0.7484**	**0.7751**
left_arm_leg	0.7706	0.7051	0.7372	**0.8206**	**0.7583**	**0.7889**
right_arm_leg	0.6997	0.6418	0.6706	**0.7814**	**0.7435**	**0.7690**

4.4.4 Observations and Analysis

The Transformer-based GAN consistently achieves lower reconstruction errors (MSE and MAE) and higher weighted accuracy across all occlusion cases. Its ability to attend over all temporal steps enhances its performance in scenarios with complex or multiple occlusions. The CRNN-based model still performs well but is less effective in recovering from larger missing parts like both arms or combined limb occlusions.

4.5 Results Comparison with Evaluation Protocols

The following evaluation protocols are used to assess the effectiveness of the reconstruction:

1. Baseline The classifier is directly applied on occluded skeletons without reconstruction.After training, G is used independently.
2. Reference Skeletons are reconstructed using interpolation and passed to the classifier.
3. Regression A non-adversarial regression model is used to reconstruct missing joints.
4. Occluded Classifier is trained and tested on occluded data directly.
5. GAN-CRNN Skeletons are reconstructed using the CRNN-based GAN and evaluated.
6. GAN-Transformer Skeletons are reconstructed using the Transformer-based GAN and evaluated.

The dataset was split into training, validation, and testing sets using an 80%–10%–10% split via Scikit–learn's `train_test_split` method.

Evaluation was performed using the following metrics:

4.5.1 Weighted Accuracy

Weighted Accuracy (WAcc) evaluates skeleton reconstruction by giving higher weight to key joints (hips, shoulders) and lower weight to less critical ones (wrists, ankles), providing a more realistic measure of overall body structure and motion. The results can be seen in Table 4.5.

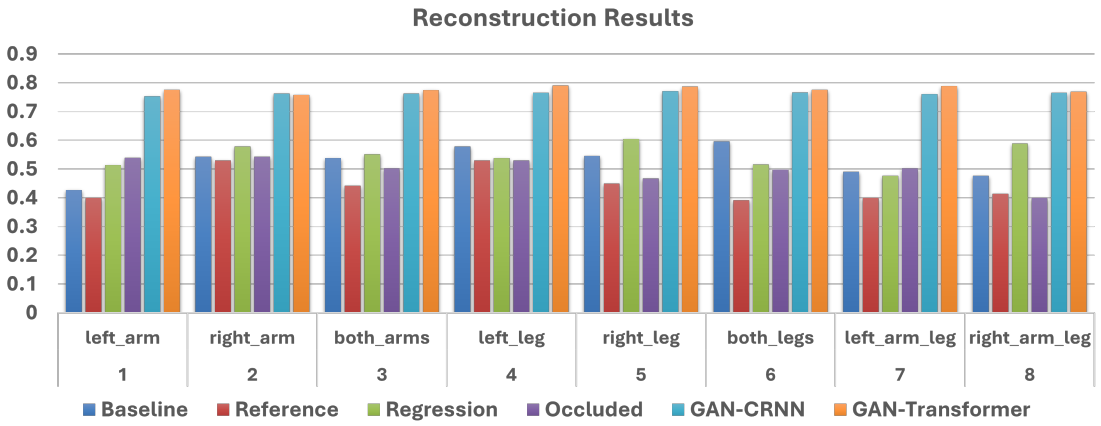

FIGURE 4.18: Weighted Accuracy of Occlussion Cases

TABLE 4.5: Weighted Accuracy Comparison Across Evaluation Protocols

Case	Base.	Ref.	Regr.	Occlud.	CRNN	Transformer
left_arm	0.4269	0.4000	0.5134	0.5386	0.7459	0.7755
right_arm	0.5427	0.5292	0.5784	0.5427	0.7804	0.7576
both_arms	0.5380	0.4421	0.5515	0.5023	0.7110	0.7742
left_leg	0.5784	0.5292	0.5380	0.5298	0.7362	0.7903
right_leg	0.5459	0.4491	0.6047	0.4672	0.7555	0.7866
both_legs	0.5959	0.3912	0.5158	0.4982	0.7601	0.7751
left_arm_leg	0.4908	0.4000	0.4760	0.5023	0.7372	0.7889
right_arm_leg	0.4760	0.4135	0.5889	0.4000	0.6706	0.7690

Figure 4.18, illustrates the weighted accuracy achieved by both the CRNN-based and Transformer-based GAN models across all occlusion types. The Transformer model consistently outperforms CRNN, particularly in more complex occlusions such as "both arms" or "right arm and leg," due to its superior temporal modeling through self-attention.

4.5.2 Evaluation MSE

Mean squared error which tells the distance between the real and the reconstructed frame. The results of MSE are presented in Table 4.6 and shown in Figure 4.19.

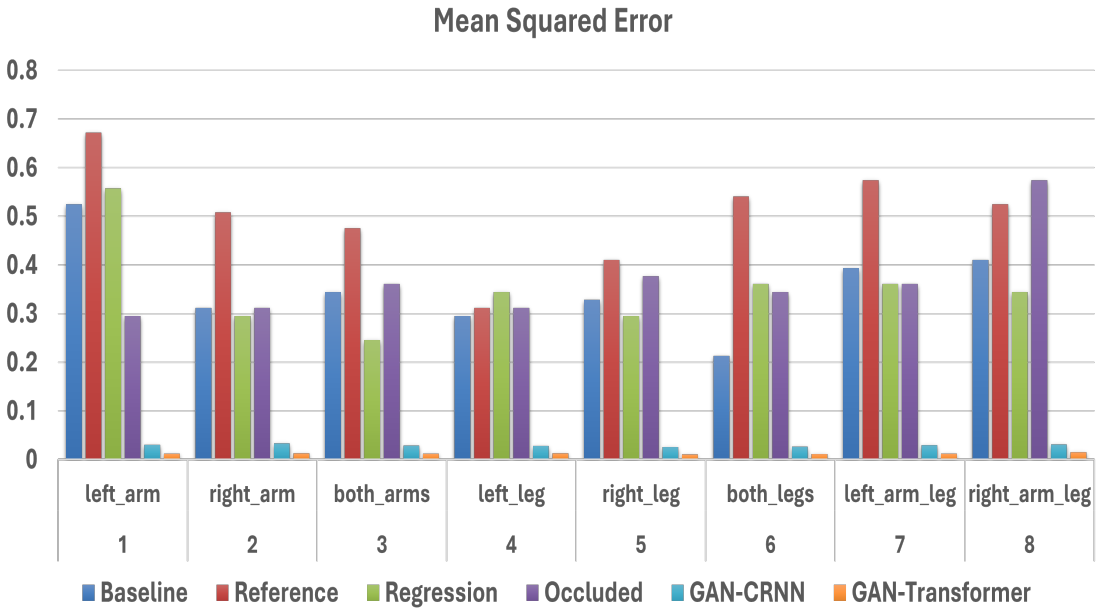

FIGURE 4.19: MSE across Evaluation Protocols

TABLE 4.6: MSE Comparison Across Evaluation Protocols

Case	Base.	Ref.	Regr.	Occlud.	GAN-CRNN	GAN-Transformer
left_arm	0.5246	0.6721	0.5574	0.2951	0.0212	0.0121
right_arm	0.3115	0.5082	0.2951	0.3115	0.0147	0.0127
both_arms	0.3443	0.4754	0.2459	0.3607	0.0377	0.0122
left_leg	0.2951	0.3115	0.3443	0.3115	0.0245	0.0124
right_leg	0.3279	0.4098	0.2951	0.3770	0.0181	0.0102
both_legs	0.2131	0.5410	0.3607	0.3443	0.0169	0.0114
left_arm_leg	0.3934	0.5738	0.3607	0.3607	0.0201	0.0117
right_arm_leg	0.4098	0.5246	0.3443	0.5738	0.0490	0.0150

4.5.3 Evaluation MAE

Mean absolute error which tells the distance between the real and the reconstructed joints. The results are shown in Table 4.7.

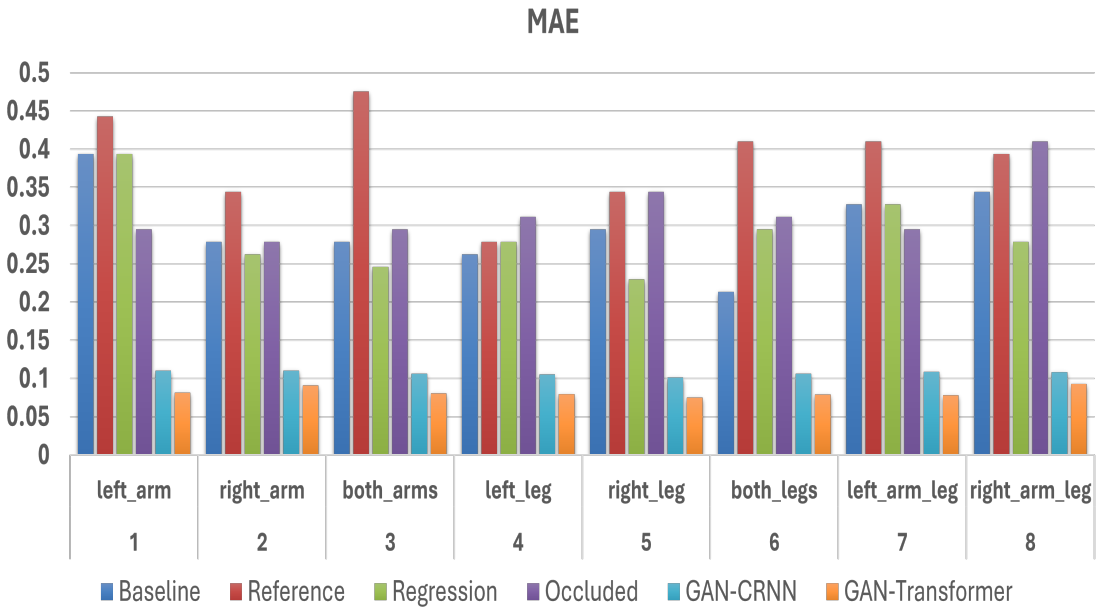

FIGURE 4.20: MAE of occluded Cases

TABLE 4.7: MAE Comparison Across Protocols

Case	**Base.**	**Ref.**	**Regr.**	**Occlud.**	**GAN-CRNN**	**GAN-Transformer**
left_arm	0.3934	0.4426	0.3934	0.2951	0.1093	0.0811
right_arm	0.2787	0.3443	0.2623	0.2787	0.0882	0.0910
both_arms	0.2787	0.4754	0.2459	0.2951	0.1449	0.0806
left_leg	0.2623	0.2787	0.2787	0.3115	0.1201	0.0795
right_leg	0.2951	0.3443	0.2295	0.3443	0.1001	0.0748
both_legs	0.2131	0.4098	0.2951	0.3115	0.0981	0.0790
left_arm_leg	0.3279	0.4098	0.3279	0.2951	0.1086	0.0777
right_arm_leg	0.3443	0.3934	0.2787	0.4098	0.1715	0.0927

As shown in Figure 4.20, occlusion severity impacts MAE values significantly. Multi-limb occlusions like "right arm and leg" and "both arms" lead to the highest reconstruction errors. However, the Transformer-based GAN still maintains lower MAE, indicating stronger generalization to severe occlusion.

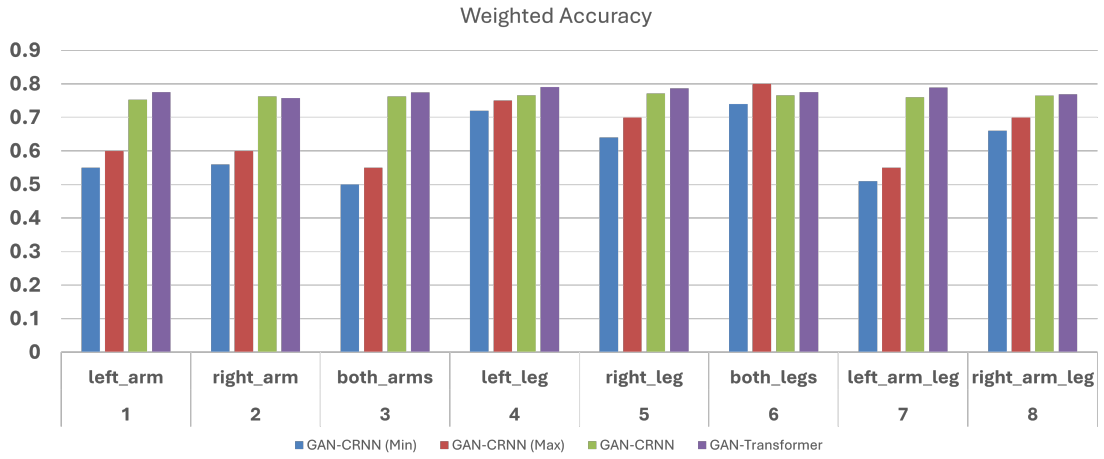

FIGURE 4.21: Proposed vs Paper Reconstruction Results

4.5.4 Comparison with Paper Results

The weighted accuracy of the paper [1] model (GAN-CRNN+LSTM (MIN and MAX)) and our proposed models (GAN-CRNN+BiLSTM and GAN-Transformer) are compared in the Table 4.8. The comparison focuses on how well the GAN-based models performed in each of the eight occlusion cases (columns GAN-CRNN and GAN-Transformer in our findings, and GAN-CRNN (Min) and GAN-CRNN (Max) in the benchmark). As shown in Figure 4.21, this chart is basically showing the results comparison between our proposed models vs the paper models.

In eight occlusion scenarios for skeletal reconstruction, our proposed GAN models, which use CRNN + BiLSTM and transformer–based architectures, significantly outperformed the benchmarked GAN–CRNN+LSTM (Min) and GAN–CRNN+LSTM (Max) models. Improvements ranging 2.59% to 26.24% (average 15.32%) are achieved by implementing the GAN-CRNN model, with the greatest benefits occurring in difficult instances such as both_arms (0.7624 vs. 0.5000). The GAN-Transformer model performed exceptionally well in left_arm_leg scenarios, increasing by -2.48% to 23.89% (average 12.09%) (0.7889 vs. 0.5500). In contrast to the benchmark's unidirectional LSTM, the BiLSTM's bidirectional temporal processing, the transformer's global spatial-temporal modeling through self-attention, and the transformer-based discriminator's improved training enable superior handling of occlusions.

TABLE 4.8: Performance Comparison Across Models

Case No.	**Occlusion Case**	**GAN-CRNN-LSTM (Min)**	**GAN-CRNN-LSTM (Max)**	**GAN-CRNN-BiLSTM**	**GAN-Transformer**
1	left_arm	0.55	0.60	0.752667	**0.775511**
2	right_arm	0.56	0.60	**0.762630**	0.757655
3	both_arms	0.50	0.55	0.762401	**0.774202**
4	left_leg	0.72	0.75	0.765555	**0.790305**
5	right_leg	0.64	0.70	0.771065	**0.786622**
6	both_legs	0.74	0.80	0.765883	**0.775159**
7	left_arm_leg	0.51	0.55	0.760265	**0.788886**
8	right_arm_leg	0.66	0.70	**0.765261**	0.769029

4.6 Classification Results

After the reconstruction of skeleton, we have feed the reconstructed skeleton data to the our classifier Model i.e. based on LSTM. We have got followoing confusion matrix, in which there were total 10 action classes i.e. showing which class was recognized in all the occlusion cases.

To evaluate the classification performance of reconstructed skeletons, confusion matrices were generated for both CRNN-based and Transformer-based GAN outputs. These matrices help visualize how accurately the reconstructed sequences preserve activity-specific spatial-temporal patterns that are critical for correct classification.

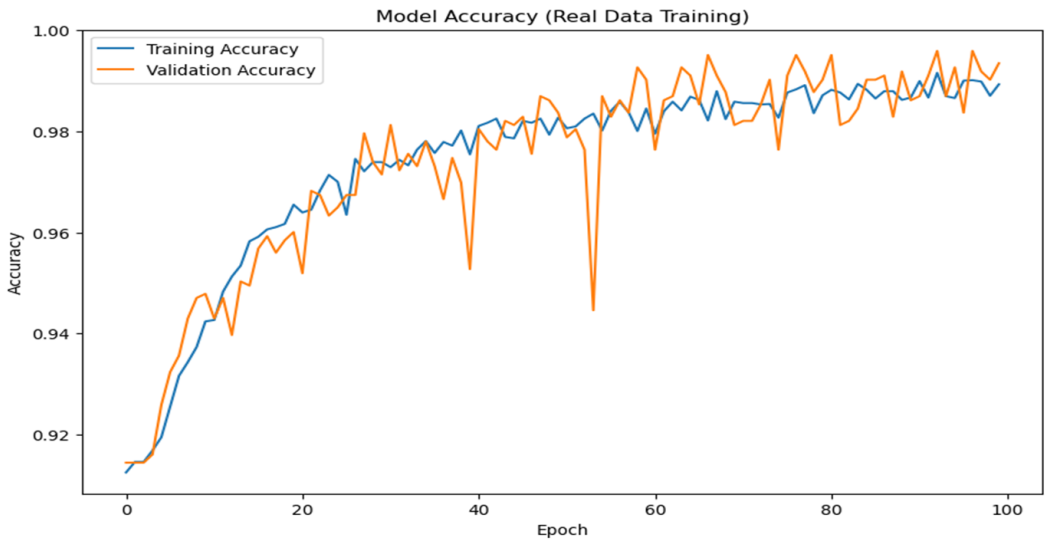

FIGURE 4.22: Model Accuracy

4.6.1 Model Accuracy

This graph is showing how well our model trained for classification and can be observed in Figure 4.22.

4.6.2 CRNN vs Transformer

So, in this section, I have shared the classification results of our proposed model, we have done following settings like Training: 72% Real, 8% Validated & Testing on 20% Reconstructed. The classification results on CRNN and Transformer based reconstructed skeleton dataset can be seen in Figures 4.23 to 4.38.

4.7 Observations and Analysis

The Transformer-based GAN consistently outperforms the CRNN-based GAN across all occlusion types and evaluation metrics (MSE, MAE, and Weighted Accuracy). This suggests that attention mechanisms in Transformers offer better temporal modeling than recurrent structures, especially for complex occlusion patterns such as both arms or right arm and leg. However, CRNN still provides reasonably competitive results at a lower computational cost.

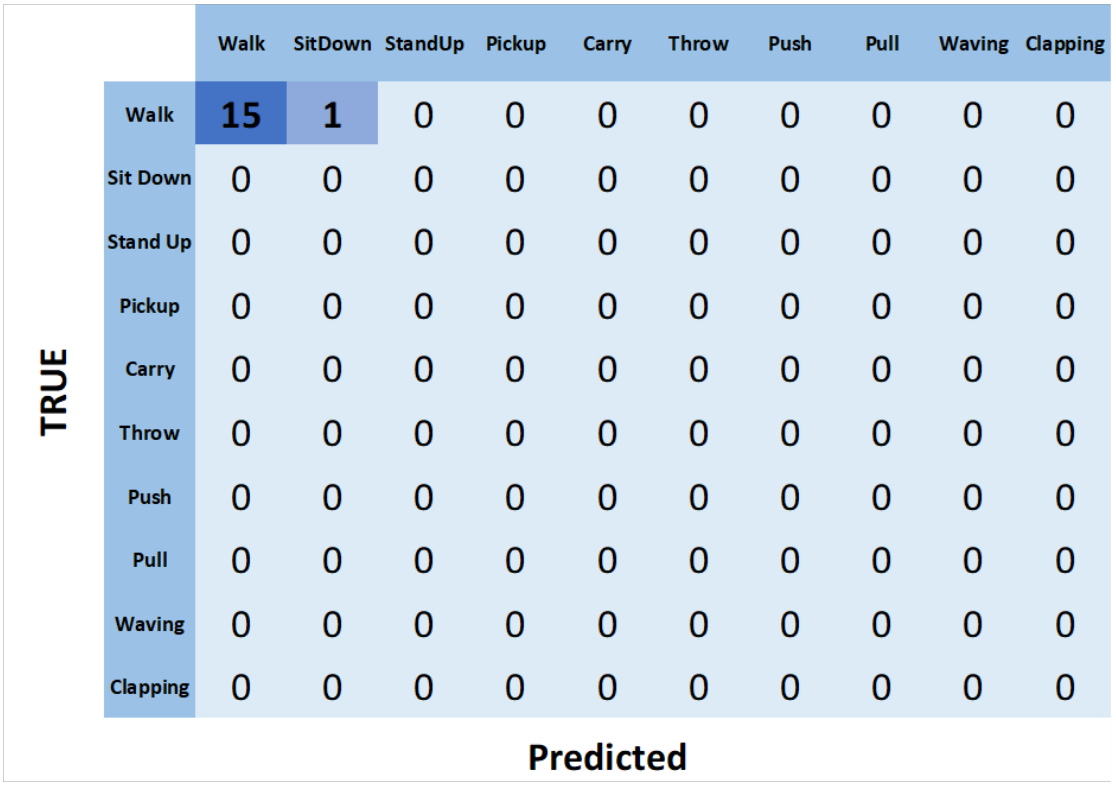

FIGURE 4.23: CRNN Left ARM

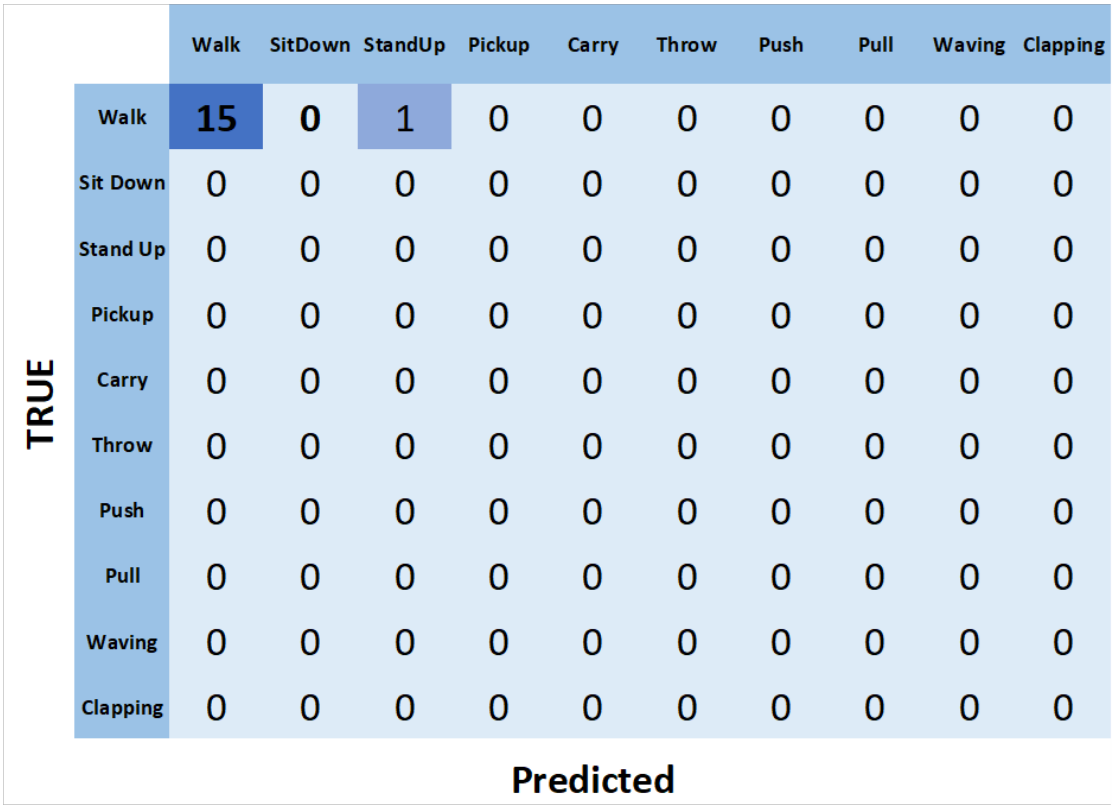

FIGURE 4.24: Transformer Left Arm

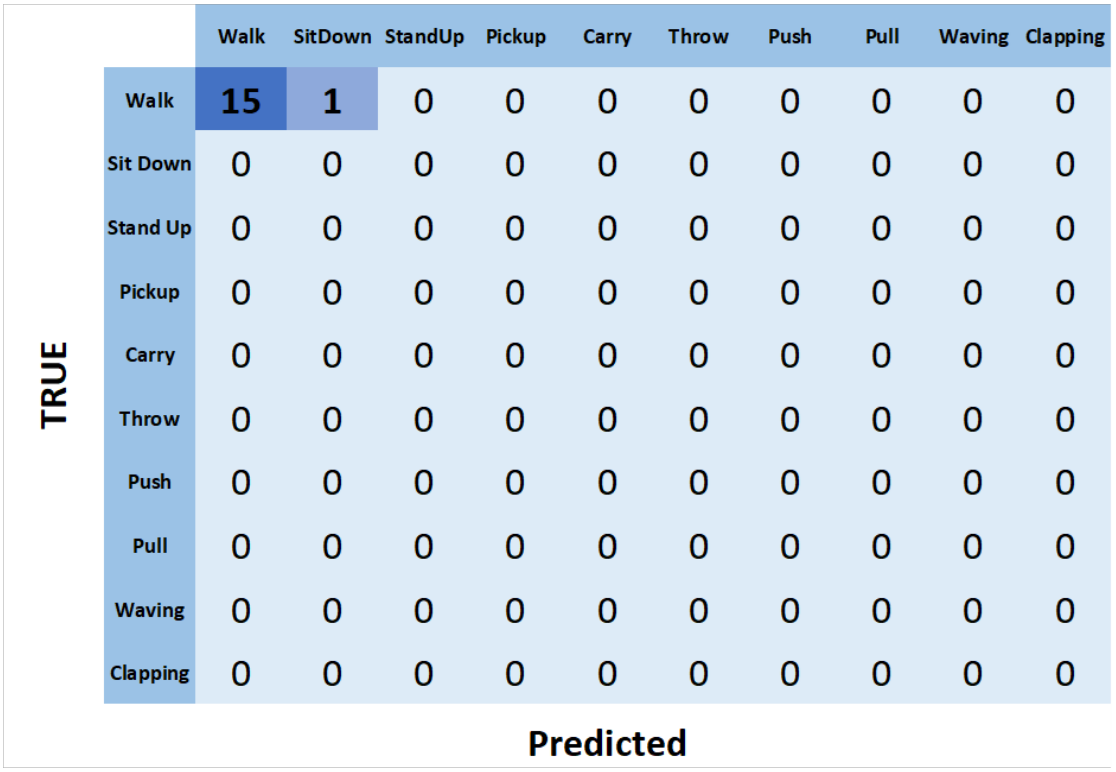

FIGURE 4.25: CRNN Right ARM

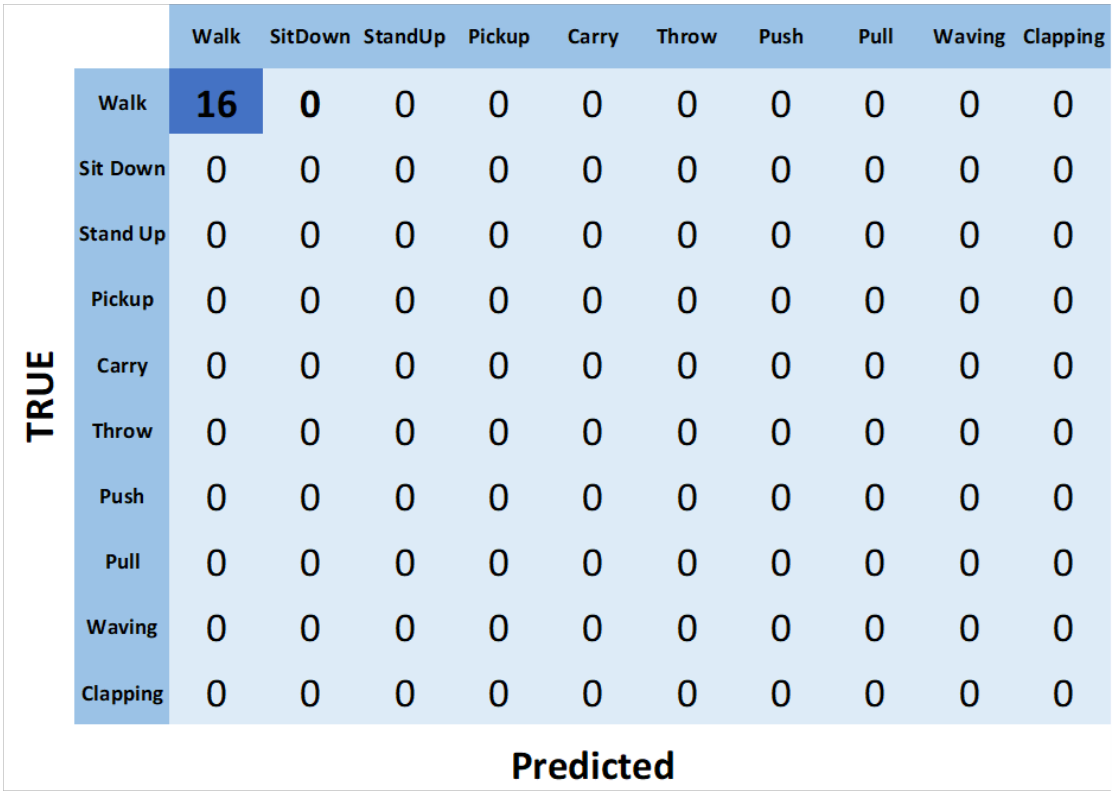

FIGURE 4.26: Transformer Right ARM

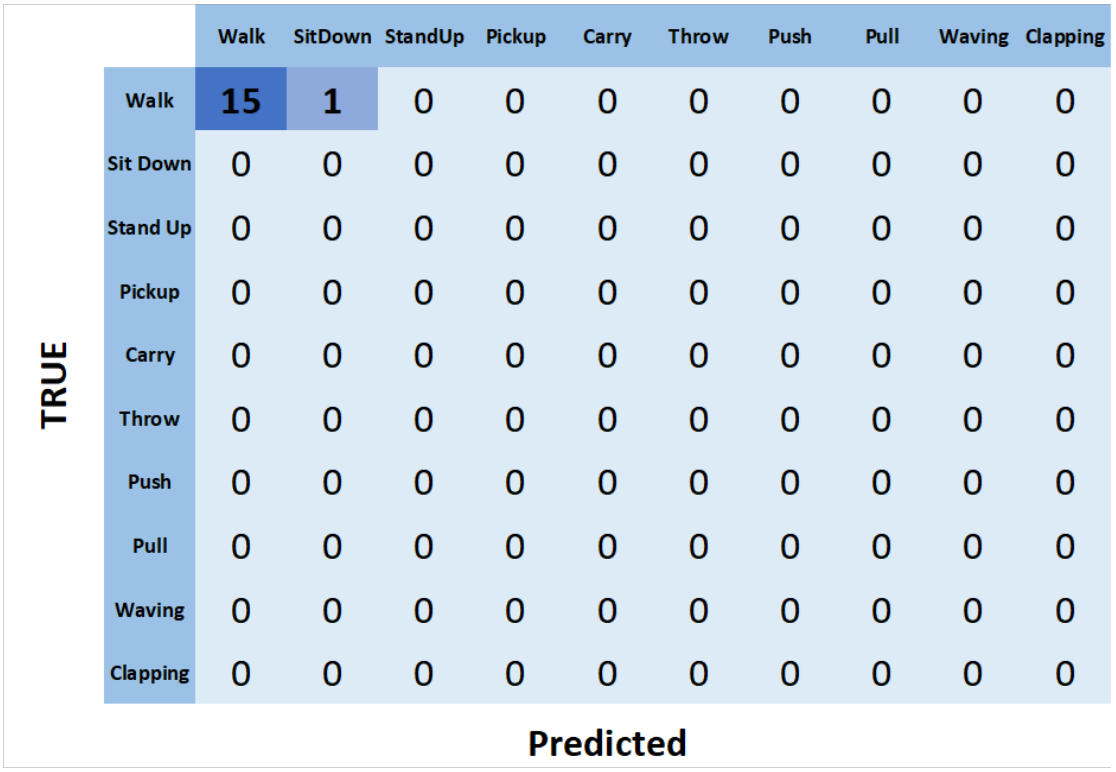

FIGURE 4.27: CRNN Left Leg

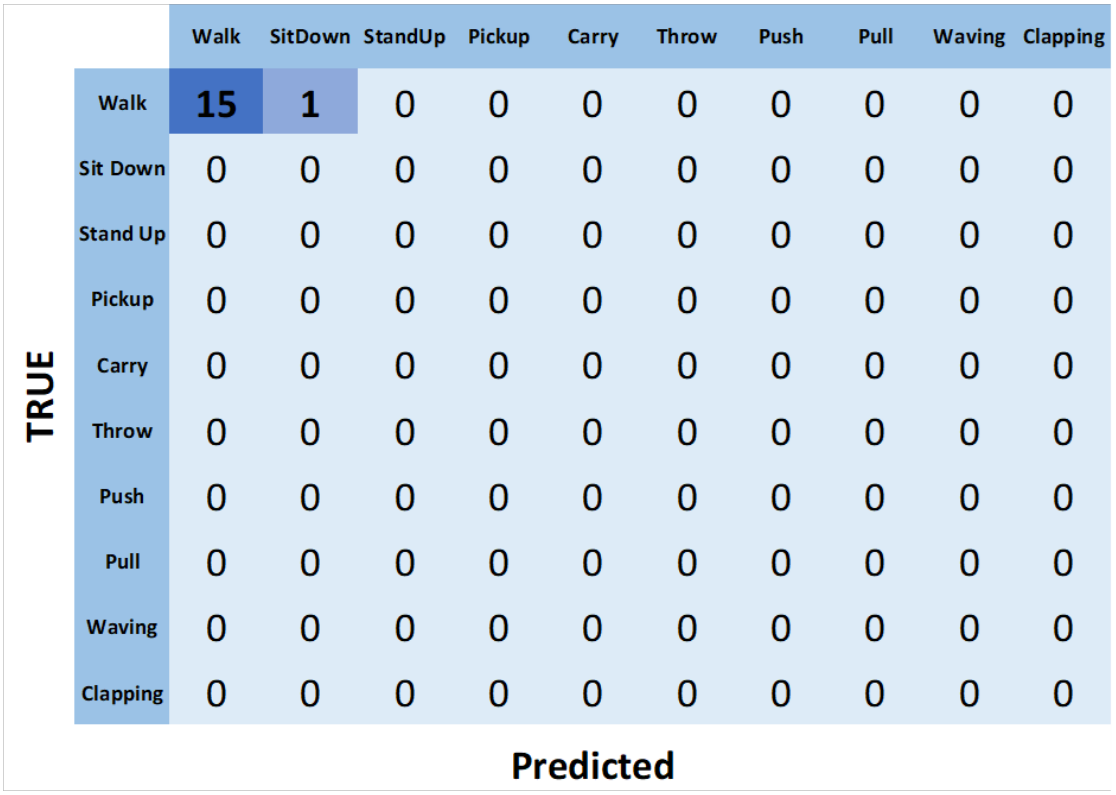

FIGURE 4.28: Transformer Left Leg

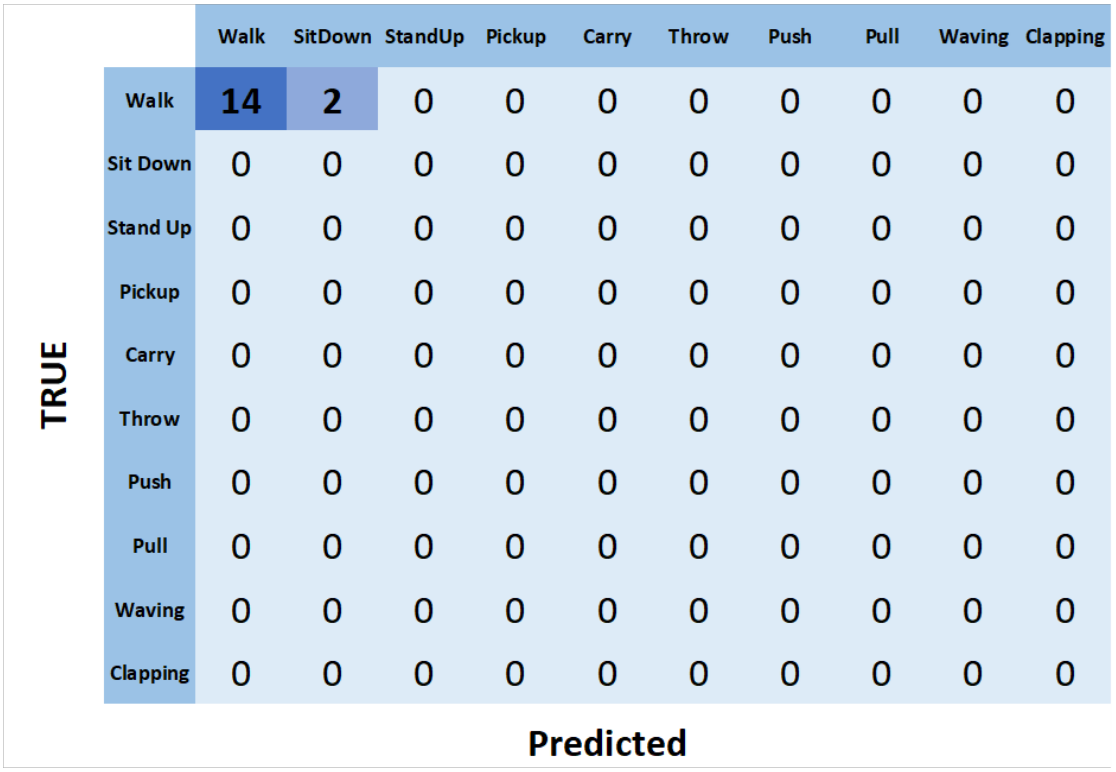

FIGURE 4.29: CRNN Right Leg

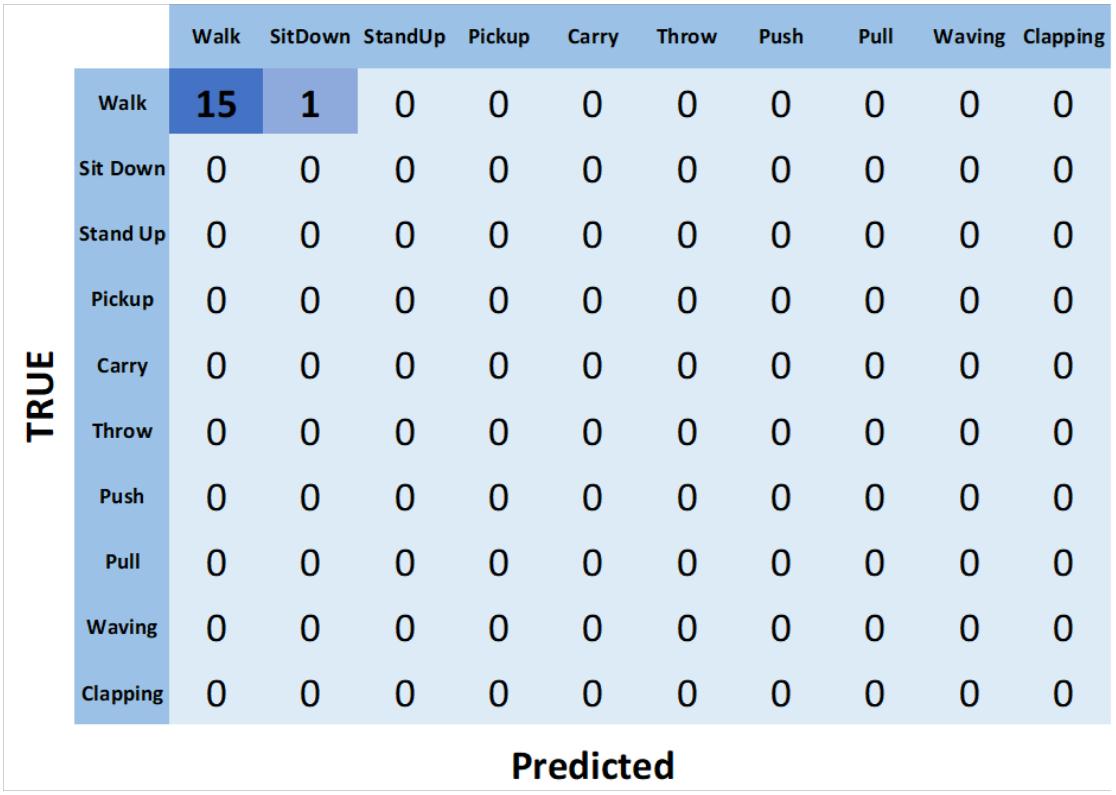

FIGURE 4.30: Transformer Right LEG

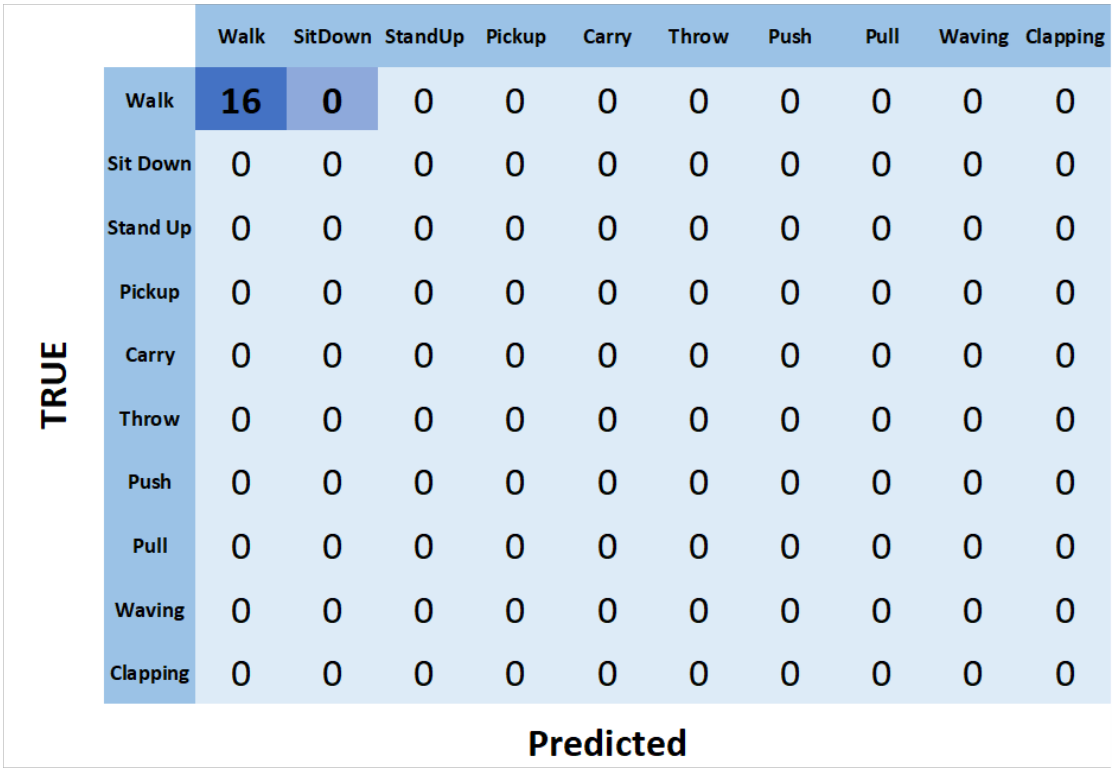

FIGURE 4.31: CRNN Both Arms

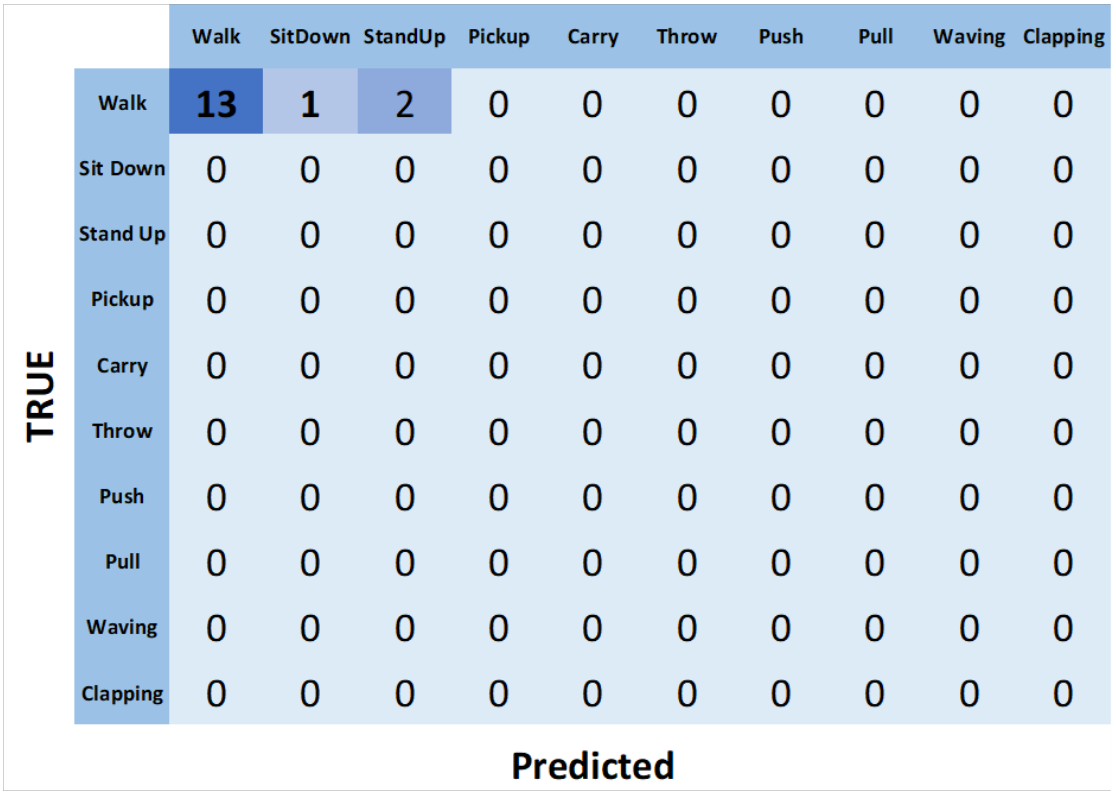

FIGURE 4.32: Transformer Both Arms

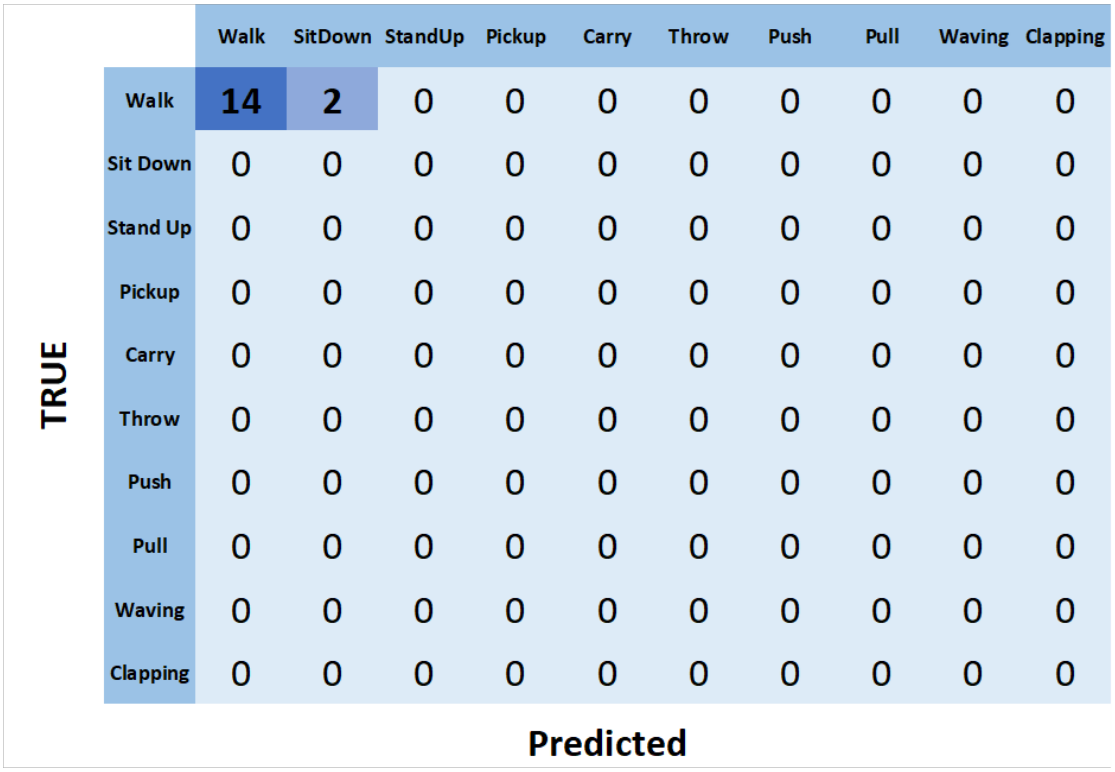

FIGURE 4.33: CRNN Both Legs

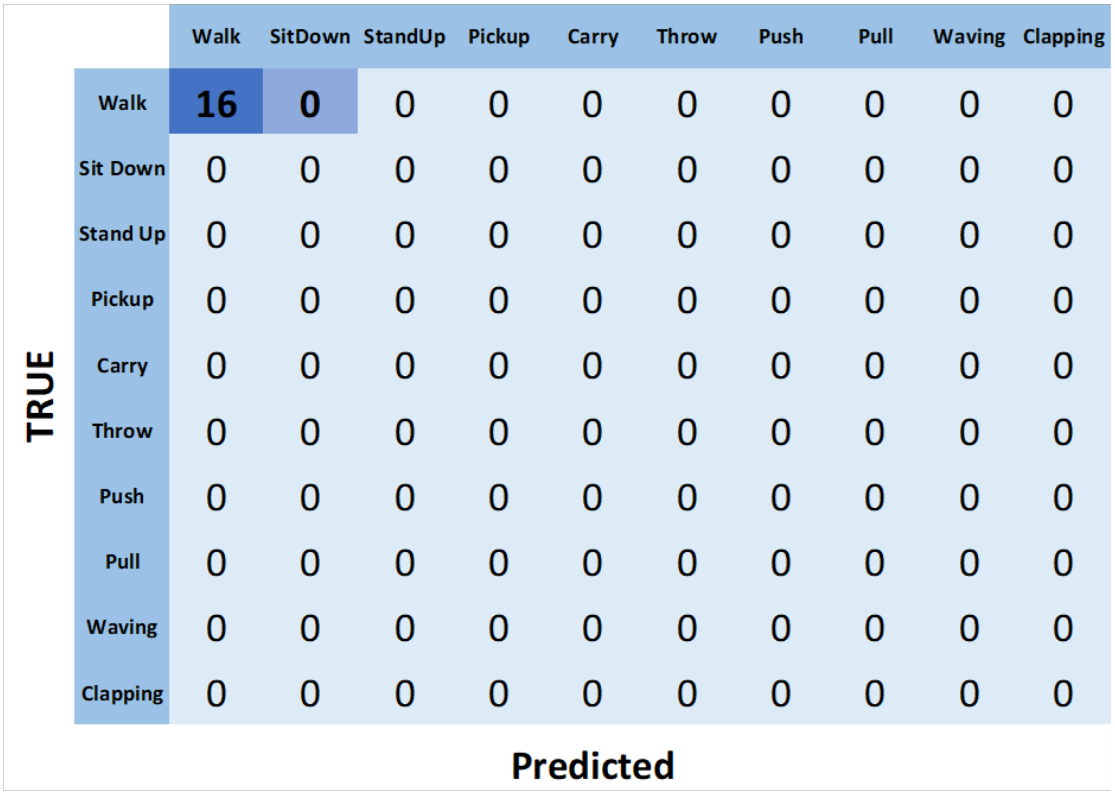

FIGURE 4.34: Transformer Both Legs

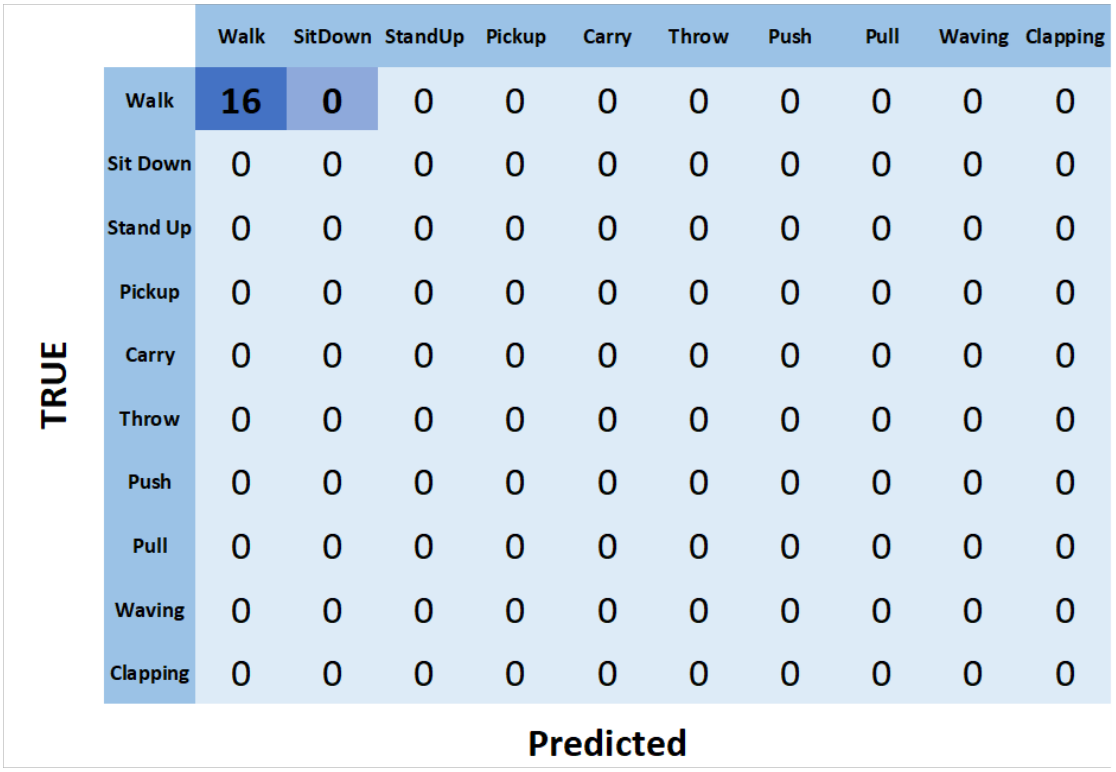

FIGURE 4.35: CRNN Left Arm and Leg

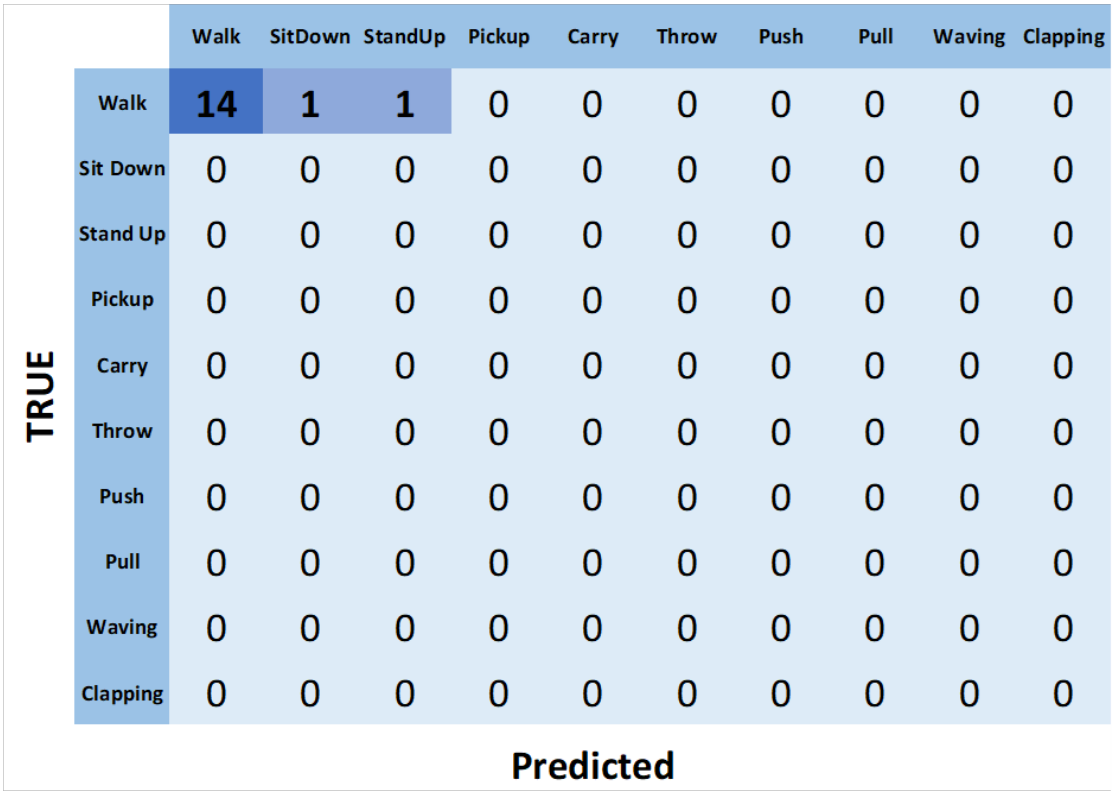

FIGURE 4.36: Transformer Left Arm and Leg

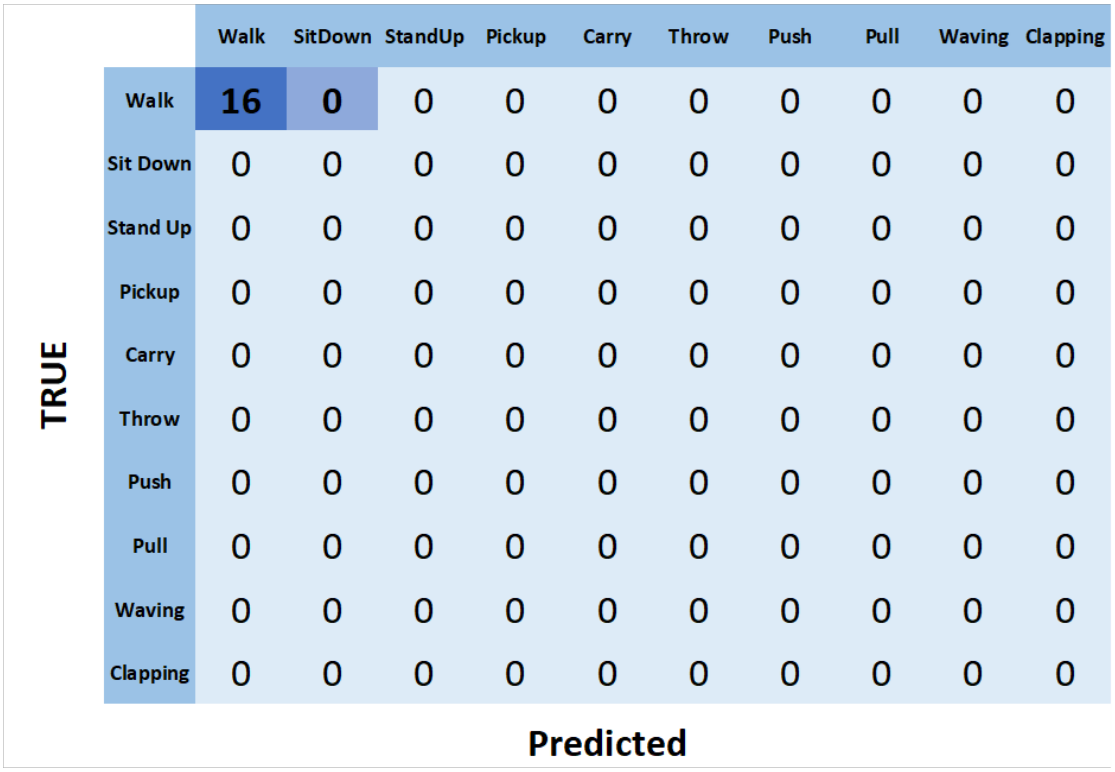

FIGURE 4.37: CRNN Right Arm and Leg

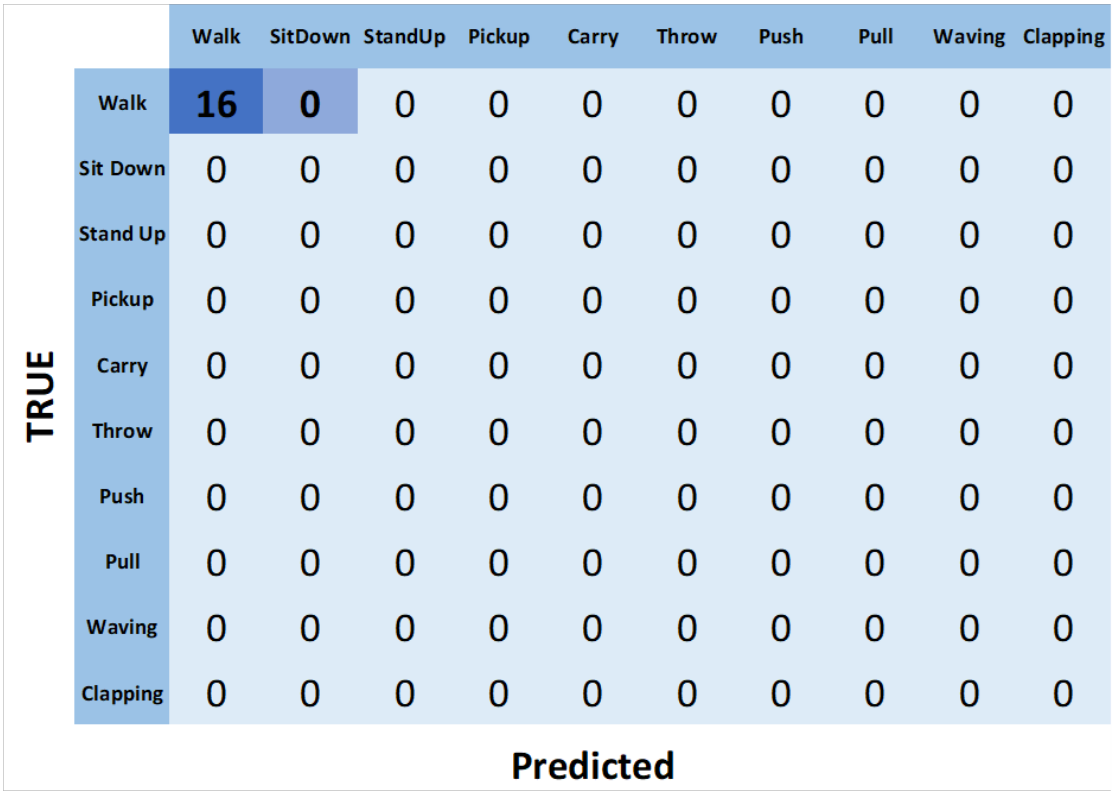

FIGURE 4.38: Transformer Right Arm and Leg

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this research, we proposed one novel generative adversarial network (GAN) model based on Transformer while one novel approach was already proposed in which GAN was based on CRNN generator model while we have improved this CRNN model by which we outperform the results as compare to the previous CRNN+LSTM based GAN model [1] for human skeleton reconstruction under occlusion to enhance human activity recognition (HAR). Our first model utilized a Convolutional Recurrent Neural Network with Bidirectional Long Short-Term Memory (CRNN+BiLSTM) as the generator, whereas the second model used transformer as generator, both paired with a transformer-based discriminator. Experimental results proofed that both models significantly outperformed baseline, reference, regression, and occluded scenarios, achieving weighted accuracies ranging from 0.7527 to 0.7903 across eight occlusion cases. As Compared to a recent benchmark using a GAN with CRNN+LSTM and pix2pix GAN function, our models exhibited notable improvements. The GAN-CRNN achieving up to 26.24% higher accuracy (e.g., 0.7624 vs. 0.5000 for both_arms) and the GAN-Transformer up to 23.89% higher accuracy (e.g., 0.7889 vs. 0.5500 for left_arm_leg). The average improvements were 15.32% for GAN-CRNN and 12.09% for GANTransformer which highlighting their superior performance.

A number of architectural benefits are responsible for our models' superior performance. Unlike the benchmark's unidirectional LSTM, which has trouble with complex occlusions, the GAN-CRNN+BiLSTM has bidirectional temporal processing, which enables it to use both past and future information to reconstruct occluded joints. By modeling long-range spatial and temporal dependencies utilizing self-attention, the GAN-Transformer further excels and allows for unified processing of joint interactions and motion dynamics. By assessing global coherence, the transformer-based discriminator improves the adversarial training procedure and raises the generator's accuracy in skeleton production. Although both models are resilient to occlusions, the transformer's ability to selectively focus on reliable data points allows it to function best in complex scenarios (e.g., left_arm_leg: 0.7889).

Not with standing these developments, real-world HAR still has many drawbacks. Considering the majority of datasets are gathered in controlled environments, occlusions brought on by self-occlusion, environmental barriers, or interactions between multiple people continue to be difficult to overcome. Real-world application is further constrained by sensor limits, computing complexity, and ethical issues with data collecting. The computing demands of the GAN-Transformer in particular make real-time deployment challenging.

5.2 Future Work

In order to improve skeleton reconstruction for HAR, future studies will concentrate on resolving these issues. In order to minimize computational cost and enable real-time applications, our first goal is to optimize the transformer-based model utilizing effective versions (such as Sparse Transformers). Second, the generalizability of the model will be enhanced by creating dynamic, varied datasets that represent actual occlusion situations. Third, adding multi-sensor systems (such depth, RGB, and inertial sensors) could increase resistance to changes in the environment. Fourth, the development of occlusion-handling techniques, like

transformers that consider uncertainty, will significantly improve reconstruction accuracy.

Finally, by establishing ethical frameworks for data collection in public spaces will address privacy concerns and reduce biases, paving the way for robust, real-world HAR systems.

Bibliography

[1] Vernikos, I., & Spyrou, E. (2025). "Skeleton Reconstruction Using Generative Adversarial Networks for Human Activity Recognition Under Occlusion." *Sensors*, **25**(5), 1567.

[2] Giannakos, I., Mathe, E., Spyrou, E., & Mylonas, P. (2021). "A study on the Effect of Occlusion in Human Activity Recognition". In *Proceedings of the 14th Pervasive Technologies Related to Assistive Environments Conference* (pp. 473–482). ACM.

[3] Vernikos, I., Spyrou, E., Kostis, I. A., Mathe, E., & Mylonas, P. (2023). "A deep regression approach for human activity recognition under partial occlusion". *International Journal of Neural Systems*, **33**(09), 2350047.

[4] Wang, P., Li, W., Ogunbona, P., Wan, J., & Escalera, S. (2018). "RGB-D-based human motion recognition with deep learning: A survey". *Computer Vision and Image Understanding*, **171**, 118–139.

[5] Wang, P., Li, Z., Hou, Y., & Li, W. (2016). "Action recognition based on joint trajectory maps using convolutional neural networks". In *Proceedings of the 24th ACM International Conference on Multimedia* (pp. 102–106). ACM.

[6] Hou, Y., Li, Z., Wang, P., & Li, W. (2016). "Skeleton optical spectra-based action recognition using convolutional neural networks". *IEEE Transactions on Circuits and Systems for Video Technology*, **28**(3), 807–811.

[7] Li, C., Hou, Y., Wang, P., & Li, W. (2017). "Joint distance maps based action recognition with convolutional neural networks". *IEEE Signal Processing Letters*, **24**(5), 624–628.

[8] Mathe, E., Maniatis, A., Spyrou, E., & Mylonas, P. (2020). "A deep learning approach for human action recognition using skeletal information". In *GeNeDis 2018: Computational Biology and Bioinformatics* (pp. 105–114). Springer, Cham.

[9] Du, Y., Fu, Y., & Wang, L. (2015). "Skeleton based action recognition with convolutional neural network". In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 579–583). IEEE.

[10] Kwon, T., Tekin, B., Tang, S., & Pollefeys, M. (2022). "Context-aware sequence alignment using 4D skeletal augmentation". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8172–8182). IEEE.