UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Developing Intelligent Sports Analysis System: Enhancing Keypoint Prediction, Object Tracking, and Computational Efficiency in Real-World Scenarios

by Zhencheng Fan

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Shiping Wen

University of Technology Sydney
Faculty of Engineering and Information Technology

February 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Zhencheng Fan*, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note:
Signature removed prior to publication.

[ZHENCHENG FAN]

DATE: 17th February, 2025

PLACE: Sydney, Australia

Abstract

In real-world sports examination scenarios, various challenges arise due to the complexity and uncontrollable factors in the testing environment. This thesis presents an intelligent sports analysis system designed to address key issues in objective sports examinations.

One major challenge we address is the problem of human keypoint prediction. During examinations, candidates' skeletons may be occluded, leading to inaccurate keypoint predictions. To address this, we propose a method inspired by techniques used in weather prediction. We employ Spatio-Temporal Graph Neural Processes (STGNP) for effective spatio-temporal extrapolation of skeleton data. STGNP learns deterministic spatio-temporal representations through cross-set graph neural networks and causal convolutions, then generates latent variables for target locations using Graph Bayesian Aggregation (GBA). GBA integrates contextual data with uncertainty estimates, allowing the system to accurately infer and complete occluded keypoints. Extensive experiments show that STGNP can effectively enhance the accuracy and stability of skeleton prediction.

Another significant issue in practical examination settings, such as basketball skill assessments, is the occlusions and errors that can lead to the loss of tracking for equipment like basketballs and cones. Traditional object tracking and Re-Identification (Re-ID) methods often struggle to cope in these dynamic scenarios. To overcome this, we propose an enhanced object tracking model that incorporates additional positional information

of the candidates. This integration not only improves the accuracy and robustness of tracking but also ensures that the system can maintain reliable tracking even when objects are momentarily occluded or lost from view, significantly enhancing the reliability of the analysis.

Furthermore, real-world sports examinations often demand substantial computational resources due to the high computational load of deep learning algorithms. We propose a novel method, Enhancing Skeleton-Based Human Motion Recognition with Lie Algebra and Memristor-Augmented Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs). This approach leverages Lie algebra for skeleton representation and employs a combination of LSTM and CNNs for motion recognition. By embedding the trained network weights into a memristor-based structure, we achieve faster inference and reduced computational requirements, ensuring efficient performance in resource-constrained environments. This method not only accelerates computation but also reduces energy consumption, making it suitable for real-time applications.

Additionally, we explore the implications of these methodologies specifically in the context of real-world basketball examination scenarios. The proposed solutions are validated through extensive experiments and trials in this specific setting, demonstrating their effectiveness and robustness.

In conclusion, this thesis provides innovative solutions to key challenges in sports examinations, including keypoint prediction, object tracking, and computational efficiency. These contributions improve the reliability and effectiveness of intelligent sports analysis systems, increasing their adaptability to real-world scenarios and opening new avenues for future advancements.

Dedication

To myself . . .

ACKNOWLEDGMENTS

My journey at the University of Technology Sydney (UTS) over the past three and a half years has been both memorable and exhilarating as I pursued my Ph.D. degree. I wish to extend my deepest gratitude to everyone who has offered their support and inspiration along the way. Your contributions have been invaluable, and I am sincerely thankful for all the encouragement and guidance I have received.

First and foremost, my deepest thanks go to my principal supervisor, Professor Shiping Wen, for his constant support and invaluable guidance throughout my Ph.D. journey. Professor Wen's patience, encouragement, and mentorship have been pivotal to my academic development and success. His belief in my potential and dedication to fostering my research interests have enabled me to explore new frontiers and innovate. His profound expertise and extensive knowledge have continually inspired me to delve deeper into my research pursuits. His guidance and insightful perspectives have provided invaluable clarity during moments of uncertainty, while his confidence and enthusiasm have motivated me to overcome challenges. I am profoundly honored to have been mentored by such a distinguished researcher and compassionate supervisor. The lessons learned and experiences gained under his guidance have significantly enriched my Ph.D. journey and will serve as invaluable assets throughout my life.

Additionally, I wish to convey my heartfelt thanks to my co-supervisors, Dr. Zheng Yan and Prof. Lu Qin. Their guidance has been pivotal in my transition from a novice to

a skilled researcher. They consistently offered precise direction and shared invaluable expertise, which have been instrumental in developing my research abilities. Our discussions have always provided fresh perspectives and deepened my understanding, greatly boosting my productivity and expanding my research scope. Their mentorship has not only bolstered my confidence but also provided much-needed support during challenging periods, both in my academic and personal life.

I would also like to extend my gratitude to the Australian Artificial Intelligence Institute (AAII) and the Faculty of Engineering and Information Technology (FEIT) at the University of Technology Sydney (UTS) for their invaluable support and funding throughout my Ph.D. journey. Their assistance has been crucial in providing the resources and environment necessary for my research. The opportunities and support provided by AAII and FEIT have significantly contributed to the successful completion of my studies.

A special thanks to Dr. Junfeng Hu, whose support and guidance have been instrumental from my undergraduate years through to my Ph.D. His mentorship, encouragement, and unwavering belief in my abilities have been a cornerstone of my academic and personal development. I am profoundly grateful for his significant and continued contributions to my journey.

Lastly, I would like to thank the members of my research group, Dr. Linhao Zhao, Dr. Ziyu Sheng, Dr. Boqian Li and Dr. Guangyang Tian. Their collaboration, insights, and support have been invaluable throughout my Ph.D. studies. I deeply appreciate their willingness to share their expertise and their constant encouragement.

TABLE OF CONTENTS

List of Figures

Figure **Page**

List of Tables

Table **Page**

List of Abbreviations

STGNP Spatio-Temporal Graph Neural Processes

GBA Graph Bayesian Aggregation

Re-ID Re-Identification

LSTM Long Short-Term Memory

CNNs Convolutional Neural Networks

HPE Human Pose Estimation

AR Augmented Reality

VR Virtual Reality

RNNs Recurrent Neural Networks

PCK Percentage of Correct Keypoints

AP Average Precision

MSE Mean Squared Error

GNNs Graph Neural Networks

IoU Intersection over Union

BC-IoU Balanced Corner-IoU

POM Point Offset Module

SAN Scale Adaptive Network

GIoU Generalized IoU

AIoU Adaptive IoU

CFIoU Corner-point and Foreground-area IoU

ICIoU Improved Loss based on Complete IoU

HOG Histogram of Oriented Gradients

SVM Support Vector Machines

DPM Deformable Parts Model

RPN Region Proposal Network

YOLO You Only Look Once

SSD Single Shot MultiBox Detector

FPN Feature Pyramid Networks

ViTs Vision Transformers

DETR Detection Transformers

BACNs Boundary-Aware Convolutional Networks

CIoU Complete Intersection over Union

LOMO Local Maximal Occurrence

GANs Generative Adversarial Networks

ANNs Artificial Neural Networks

DRL Deep Reinforcement Learning

STGNNs Spatio-Temporal Graph Neural Networks

NNs Neural Networks

GPs Gaussian Processes

NPs Neural Processes

MC Monte-Carlo

ELBO Evidence Lower Bound

STRL Spatio-Temporal Representation Learning

KNN K-Nearest Neighbors

RF Random Forest

ANP Attentive Neural Processes

SNP Sequential Neural Processes

VRNN Variational Recurrent Neural Network

CSGCN Cross-Set Graph Convolution

DCConv Dilated Causal Convolutions

AutoML Automated Machine Learning

FCN Fully-Convolutional Network

MHA Multi-head Attention Layers

FFN Feed-Forward Networks

AUC Area Under the Curve

ERD Energy-Relation Diagrams

LSTM-3LR Three Layers of Long Short-Term Memory

SRNN Stacked Recurrent Network

TPF Temporal-Domain Features

SPF Spatial-Domain Features

ML Machine Learning

LLMs Large Language Models

CHAPTER 1

INTRODUCTION

1.1 Background and Motivations

In real-world sports examination scenarios, various challenges arise due to the complexity and uncontrollable factors in the testing environment. The goal of this research is to design a comprehensive and fair intelligent sports examination system, taking basketball skill assessments as a case study. Traditional methods for human pose estimation, object tracking, and Re-ID face significant challenges in such dynamic and complex environments. Human pose estimation, crucial for understanding players' movements and biomechanics, often struggles with accuracy due to occlusions, varying lighting conditions, and rapid movements. On account of limited computational resources, using small models results in more accuracy issues, while large models cannot be practically applied. Advanced methods such as AlphaPose Fang et al. (2022), MMPose Contributors (2020), and OpenPose Cao et al. (2019) have made significant strides in addressing these issues by improving pose estimation accuracy even in challenging conditions Papandreou et al. (2017); Cao et al. (2017). In video-based sports examinations, object

detection and Re-ID techniques are used together primarily to detect auxiliary equipment used in the exams, such as basketballs, hoops, and cones. Modern object detection algorithms like YOLO, with YOLOv8 being one of the most commonly used versions, provide robust solutions for tracking in real-time, offering high accuracy and efficiency Jocher et al. (2022). Re-ID techniques, which help in distinguishing and re-identifying objects throughout the assessment, face issues with changes in appearance due to occlusions and overlap with identical targets. Recent advancements such as ByteTrack and BoT-SORT have shown remarkable improvements in Re-ID performance, enabling more reliable identification of equipment in dynamic and cluttered environments Zhang et al. (2022); Aharon et al. (2022). As shown in Figure 1.1, integrating these advanced methods into a cohesive system that performs reliably in real-world sports environments remains an ongoing research challenge, necessitating continuous development and refinement. Some details will be provided as follows:

One significant challenge in accurately predicting human keypoints during examinations is dealing with partial occlusion, which can lead to incorrect or incomplete keypoint data. To tackle this issue, we utilize the STGNP Hu et al. (2023). This approach employs spatio-temporal representations through cross-set graph neural networks and causal convolutions. It generates latent variables for target locations using GBA, effectively addressing occluded keypoints by robustly inferring and supplementing them. STGNP offers precise uncertainty estimates and robust learning capabilities. By learning these spatio-temporal representations and latent variables, STGNP can accurately predict missing keypoints even in challenging situations. This makes it particularly suitable for dynamic sports environments where occlusions are frequent, thus enhancing the reliability of keypoint predictions and improving the overall performance of the intelligent sports examination system.

Another major challenge in sports examinations is dealing with occlusion of equip-

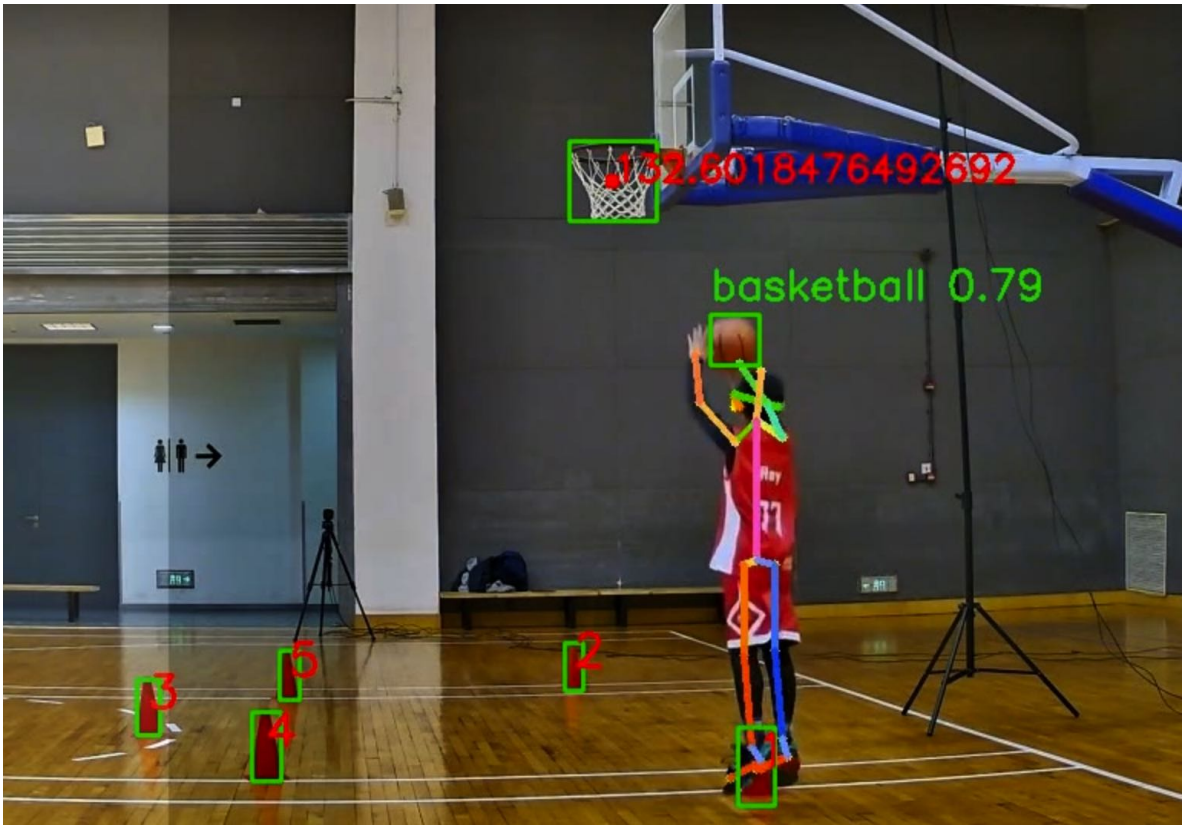

Figure 1.1: This figure illustrates a real-world basketball examination scenario where human pose estimation, cone detection, and ball tracking have been successfully performed. The image demonstrates the system's capability to accurately identify and track key elements such as the basketball, cones, player, and hoop, showcasing its effectiveness in dynamic sports environments.

ment like basketballs and cones. Traditional object tracking and Re-ID methods often falter in dynamic scenarios where objects are frequently occluded, lost from view, or overlap with identical items. This issue is particularly problematic in complex examination environments where multiple identical objects, such as basketballs, are used. The complexity of candidates' movements and overlaps caused by 2D imaging can lead to Re-ID failures, resulting in incorrect tracking and flawed evaluations in subsequent tasks. To address this, we enhance object detection algorithms by integrating positional information of the candidates. Our approach improves tracking accuracy and robustness by incorporating the spatial and temporal context of candidates, allowing the system to

correct Re-ID errors. As a result, the system maintains accurate object tracking even under challenging conditions. This positional integration significantly improves the reliability of examination analysis, enabling the system to better handle occlusions and complex interactions. This ensures fair and accurate skill assessments in dynamic sports environments.

Real-world sports examinations often require significant computational resources due to the demands of deep learning algorithms. To address this, we propose an innovative method that improves skeleton-based human motion recognition using Lie Algebra combined with Memristor-Augmented LSTM and CNN. Vision-based human action recognition is vital in many fields, including healthcare, video surveillance, autonomous driving, sports, and education Aggarwal and Ryoo (2011). Our method effectively represents human skeleton data by using Lie algebra and standard bone length measurements. We employ a multi-layer LSTM recurrent neural network and CNN to capture complex motion patterns with high accuracy Hu et al. (2019). To enhance performance, we embed the trained network weights into a memristor-based structure, achieving faster inference and lower computational requirements Wen et al. (2020). This approach not only speeds up computation but also reduces energy consumption, making it highly suitable for real-time applications. Although our current implementation is a software simulation, we aim to apply this technology practically in the future, improving the efficiency and reliability of human motion recognition in dynamic sports environments.

Many deep learning models have been applied to sports scenarios, but integrating these technologies into a single intelligent system remains unexplored. Hence, our work stands at the forefront of this field. There are three key issues that need to be addressed: 1) improving the accuracy of keypoint prediction, particularly under rapid and complex movements, varying lighting conditions, and occlusions; 2) enhancing object detection and recognition in intelligent examination scenarios, ensuring robustness amidst fre-

quent occlusions and complex interactions; and 3) developing efficient computational methods suitable for real-time applications in resource-constrained settings. Improving keypoint prediction involves addressing the accuracy of keypoints in dynamic environments, where advanced pose estimation methods have made progress, but there remains a need for techniques that can adapt to the unpredictability of real-world scenarios. Enhancing object detection and recognition requires not only robust algorithms capable of maintaining consistency but also the ability to handle sudden changes in movement and appearance, which are common in dynamic sports environments. Traditional object detection methods, despite advancements, often fail under these challenging conditions, necessitating more sophisticated approaches that incorporate contextual and temporal information. Developing efficient computational methods is critical, as real-time processing is essential for practical applications. The high computational demands of deep learning algorithms pose challenges, especially in resource-constrained environments like on-field sports assessments. Innovations in hardware and algorithmic efficiency, such as memristor-based networks, offer promising solutions by reducing energy consumption and accelerating computation. For example, basketball skill assessments require precise tracking and recognition of players and equipment in real-world conditions. Expanding these methodologies to practical scenarios and designing adaptive algorithms to address these challenges is imperative. Future research must focus on these areas to build more reliable, efficient, and adaptable intelligent sports examination systems that can operate effectively in diverse and dynamic environments.

This research aims to provide comprehensive solutions to these challenges. The integration of STGNP for keypoint prediction addresses occlusions; advanced object detection algorithms enhance tracking accuracy, and the use of memristor-based networks improves computational efficiency. The remainder of this paper is structured as follows: Chapter 3 focuses on challenge 1, Chapter 4 offers solutions to challenge 2,

and Chapter 5 tackles challenge 3. Chapter 6 further explores the practical application of these methods in basketball skill assessments, highlighting their effectiveness and robustness in real-world scenarios.

In conclusion, this study offers innovative approaches to dealing with key challenges in sports examinations, including keypoint prediction, object tracking, and computational efficiency. By addressing these issues, we improve the reliability and effectiveness of intelligent sports analysis systems, setting the stage for further advancements in the domain.

1.2 Research Questions and Objectives

1.2.1 Research Questions

The intelligent sports examination system for basketball skill assessments faces numerous challenges due to the complexity and unpredictability of the real testing environment. Traditional methods for human pose estimation, object tracking, and Re-ID often fall short in such dynamic settings. This study aims to address these challenges by proposing innovative approaches to improve the accuracy and efficiency of these tasks. Specifically, this research focuses on the following research questions:

RESEARCH QUESTION 1 (RQ1): *How to improve the robustness and accuracy of human pose estimation?*

Human pose estimation is crucial for understanding players' movements and biomechanics. However, traditional methods face significant challenges due to occlusions, varying lighting conditions, and rapid movements. Despite advancements made by methods such as AlphaPose, MMPose, and OpenPose in improving pose estimation accuracy, issues persist in real-world scenarios with limited computational resources. In dynamic

sports environments, occlusions are common and can significantly degrade the accuracy and reliability of keypoint predictions. These occlusions can result from players moving in close proximity, interactions with equipment, or rapid changes in body orientation. Additionally, while current deep learning models perform well, their application in resource-constrained settings is limited. Small models often fail to achieve the necessary accuracy, impacting downstream tasks, while large models consume excessive resources, making practical applications challenging. There is a need to develop methods that can robustly handle occlusions and provide precise keypoint predictions, even in limited resource settings, to ensure accurate analysis of players' movements. Such advancements are essential to enhance the overall understanding of biomechanics and improve the fairness and reliability of sports examinations.

RESEARCH QUESTION 2 (RQ2): *How to enhance object tracking accuracy amidst frequent occlusions and complex interactions in real basketball examinations?*

Object tracking is essential for monitoring the use of examination equipment, such as basketballs and cones, during basketball skill assessments. Traditional methods often struggle with occlusions and the presence of multiple identical objects, making it difficult to maintain consistency and precision amidst frequent and unpredictable movements. In real-world examination scenarios, occlusions can occur frequently as players move in close proximity to each other or interact with equipment. The presence of multiple identical props, such as basketballs, adds another layer of complexity to the tracking process. Traditional Re-ID techniques may fail to correctly track and identify these objects due to occlusions, brief disappearances from view, and the dynamic nature of sports activities. Rapid changes in position and interactions further complicate the tracking process. Addressing these challenges is crucial to improve the accuracy and robustness of object tracking in complex and dynamic sports environments, ensuring reliable data collection and analysis for performance evaluation.

RESEARCH QUESTION 3 (RQ3): *How to develop efficient computational methods suitable for real-time applications in resource-constrained settings?*

Real-world sports examinations require significant computational resources due to the intensive demands of deep learning algorithms. Although current methods provide accurate results, they often lack the efficiency needed for real-time applications, particularly in environments with limited computational capacity. The challenge is to maintain high performance and accuracy in human motion recognition while drastically reducing computational demands and ensuring rapid inference times. This requires developing innovative computational techniques that can operate effectively in resource-constrained settings without sacrificing the accuracy and reliability of the examination system. Balancing the computational load with the necessity for real-time processing is crucial for practical deployment, especially in field conditions where processing power and energy resources may be limited. Efficient algorithms must be designed to navigate these constraints while delivering dependable performance for accurate and timely sports assessments.

RESEARCH QUESTION 4 (RQ4): *How to ensure the practical applicability and scalability of proposed methods in real-world sports examination scenarios?*

While several methodologies address key challenges in sports examinations, their practical application and scalability in real-world conditions remain under-explored. Ensuring that these methods can be effectively applied in practical scenarios and adapt to various sports environments is crucial. This research aims to investigate the implementation and scalability of proposed methods, assessing their performance and robustness in real-world basketball skill assessments and their adaptability to different sports settings. Identifying potential barriers and requirements for practical deployment is essential for translating theoretical advancements into usable and reliable sports examination systems. This includes understanding the variability in examination conditions, the

diversity of sports disciplines, and the logistical aspects of deploying these systems at scale. Ensuring scalability involves not only technical robustness but also considerations of cost, ease of use, and integration with existing sports examination frameworks.

1.2.2 Research Objectives

To answer these research questions, we set up the corresponding Research Objectives (RO) as follows:

RESEARCH OBJECTIVE 1 (RO1): *To enhance the robustness of human pose estimation in dynamic sports environments using STGNP.* (Aims to answer RQ1)

Traditional methods for human pose estimation face significant challenges due to occlusions, varying lighting conditions, and rapid movements in dynamic sports environments. While many deep learning methods, such as AlphaPose and OpenPose, have shown strong performance, their effectiveness varies with model size, and real-world scenarios often have limited computational resources. Smaller models may lead to inaccurate keypoint predictions, affecting overall performance. This research aims to improve the robustness of human pose estimation by leveraging STGNP. Our approach builds upon existing methods like AlphaPose and OpenPose by first identifying and removing anomalous jittery keypoints. Subsequently, STGNP is employed to supplement the missing keypoints. Inspired by the application of STGNP in weather monitoring-where it predicts missing data from certain locations using information from surrounding weather stations-we apply a similar approach to human skeleton keypoint completion. STGNP effectively handles occluded keypoints by robustly inferring and supplementing them, providing precise uncertainty estimates. Additionally, it corrects anomalies in predicted keypoints, such as sudden coordinate changes, ensuring more accurate and reliable pose estimations. This approach will significantly enhance the reliability of pose estimations, which is crucial for understanding players' movements and biomechanics

in dynamic sports environments. By addressing the issues of occlusions and erroneous keypoint predictions, this research aims to provide a robust solution for accurate player movement analysis.

RESEARCH OBJECTIVE 2 (RO2): *To develop advanced object tracking algorithms that incorporate positional information to enhance object tracking accuracy amidst frequent occlusions and complex interactions.* (Aims to answer RQ2)

In sports examinations, traditional object tracking methods often struggle with occlusions and the presence of multiple identical objects, such as basketballs. In real-world examination scenarios, the use of examination equipment is necessary. For example, in basketball skill assessments, candidates need to use props like basketballs and cones. Traditional object tracking techniques face significant challenges in these complex examination environments, which may include multiple identical props or situations where the props are occluded or briefly disappear from view, leading to incorrect tracking and unreliable detection results. This research proposes integrating positional information of candidates into object tracking algorithms to improve accuracy and robustness. By incorporating the spatial context of candidates, the system can correct tracking errors, ensuring precise tracking even in challenging conditions. Building upon the method proposed in Yan et al. (2021), we enhance object tracking by including the candidate's positional information, enabling the system to correctly identify and track the relevant examination props throughout the entire examination process. This integration significantly improves the reliability of player and equipment tracking, ensuring accurate and robust analysis of sports examinations despite the complexities of real-world environments.

RESEARCH OBJECTIVE 3 (RO3): *To develop a novel method for efficient computational processing in real-time applications using Lie Algebra and Memristor-Augmented LSTM and CNN.* (Aims to answer RQ3)

Real-world sports examinations demand substantial computational resources due to the high load of deep learning algorithms. This research proposes a novel method that enhances skeleton-based human motion recognition using Lie Algebra and Memristor-Augmented LSTM and CNN. Lately, as a subset of human-centric studies, vision-oriented human action recognition has emerged as a pivotal research area, given its broad applicability in fields like healthcare, video surveillance, autonomous driving, sports, and education. This brief applies Lie algebra and standard bone length data to represent human skeleton data. A multi-layer LSTM recurrent neural network and CNN are applied for human motion recognition, capturing complex motion patterns with high accuracy. Finally, the trained network weights are converted into a crossbar-based memristor circuit, which can accelerate the network inference, reduce energy consumption, and obtain excellent computing performance. By embedding trained network weights into a memristor-based structure, this approach aims to achieve faster inference and reduced computational requirements, making it highly suitable for real-time applications. This innovation addresses the challenge of maintaining high performance and accuracy of human motion recognition while operating efficiently in resource-constrained settings, ensuring reliable and timely sports examinations.

RESEARCH OBJECTIVE 4 (RO4): *To develop a modular and scalable video-based intelligent sports examination system, ensuring practical applicability and adaptability across various sports and dynamic environments.* (Aims to answer RQ4)

While several methods address key challenges in sports examinations, their practical application and scalability in real-world conditions remain under-explored. This research aims to develop a comprehensive, modular, video-based intelligent sports examination system, using basketball skill assessments as a primary example. The proposed system will integrate all previously mentioned technologies, ensuring robust keypoint prediction, accurate object tracking, and efficient computational processing. Video is chosen as the

primary medium due to its ease of acquisition and deployment, making it a practical solution for diverse settings. Evaluating the system's performance in real-world basketball skill assessments will provide insights into its effectiveness and robustness. Additionally, the system will be designed for scalability and adaptability, facilitating its application to different sports and dynamic environments. This includes investigating the system's adaptability to various examination conditions, sports disciplines, and environmental challenges. By designing an adaptive framework, the research seeks to ensure that the intelligent sports examination system can handle similar challenges in different contexts, broadening its scope and impact. The objective encompasses both technical robustness and practical deployment aspects, considering cost, ease of use, and integration with existing sports examination frameworks to ensure the system's real-world applicability and scalability.

1.3 Research Contributions

This thesis is dedicated to providing a thorough examination of the obstacles encountered while devising an intelligent sports analysis system for evaluating basketball skills. It introduces novel methods aimed at enhancing the precision and speed of keypoint prediction, object tracking, and computational tasks within dynamic and multifaceted sports settings. The key contributions of this research are succinctly outlined as follows:

Enhanced Robustness of Human Pose Estimation

- Integration of STGNP with AlphaPose and OpenPose to improve the robustness of keypoint prediction.

- Effective handling of occluded keypoints by robustly inferring and supplementing them, providing precise uncertainty estimates.

- Correction of anomalies in predicted keypoints, such as sudden coordinate changes, ensuring more accurate and reliable pose estimations.
- Enhanced reliability of pose estimations, crucial for understanding players' movements and biomechanics in dynamic sports environments.

Advanced Object Tracking Algorithms

- Development of advanced object tracking algorithms that incorporate positional information of candidates.
- Integration of the spatial context of candidates to correct tracking errors and ensure precise tracking.
- Inclusion of candidate's positional information to accurately identify and track examination props, ensuring reliable object tracking amidst frequent occlusions and complex interactions.
- Significant improvement in the reliability of player and equipment tracking, ensuring robust analysis in complex and dynamic sports environments.

Efficient Computational Processing

- Proposal of a novel method using Lie Algebra and Memristor-Augmented LSTM and CNN for efficient computational processing.
- Reducing the computational load of representing human skeleton data through the use of Lie algebra and standard bone length data.
- Utilization of multi-layer LSTM recurrent neural networks and CNNs for human motion recognition, capturing complex motion patterns with high accuracy.

- Conversion of trained network weights into a crossbar-based memristor circuit to accelerate network inference and reduce energy consumption.
- Achievement of faster inference and reduced computational requirements, making the approach highly suitable for real-time applications.

Modular and Scalable Video-Based Intelligent Sports Examination System

- Development of a comprehensive modular intelligent sports examination system, using basketball skill assessments as a primary example.
- Integration of robust keypoint prediction, accurate object tracking, and efficient computational processing technologies.
- Evaluation of the system's performance in real-world basketball skill assessments to provide insights into its effectiveness and robustness.
- Design of the system for scalability and adaptability, facilitating its application to different sports and dynamic environments.
- Investigation of the system's adaptability to various examination conditions, sports disciplines, and environmental challenges.
- Consideration of cost, ease of use, and integration with existing sports examination frameworks to ensure real-world applicability and scalability.

Adaptability to Various Sports and Environments

- Investigation of the proposed methodologies for their potential adaptability to different sports and dynamic environments.
- Design of adaptive frameworks to ensure the intelligent sports examination system can handle similar challenges in different contexts.

- Demonstration of the versatility and robustness of the proposed solutions in diverse real-world scenarios.
- Ensuring accurate and reliable assessments across a wide range of sports and dynamic environments.

1.4 Research Significance

The theoretical and practical significance of this thesis is summarized as follows:

Theoretical Significance: This research provides a comprehensive and standardized definition of the challenges faced in intelligent sports examination systems, particularly in dynamic sports environments. It develops innovative methods to enhance human pose estimation, object tracking, and computational processing, addressing significant gaps in existing literature. By integrating STGNP with methods like AlphaPose and OpenPose, the study offers a robust solution for handling occluded keypoints, enriching the theoretical understanding of pose estimation. The incorporation of positional information into object tracking algorithms advances the understanding of tracking in complex scenarios, offering new insights into maintaining accuracy amidst frequent occlusions and identical objects. Additionally, the use of Lie Algebra and Memristor-Augmented LSTM and CNN for efficient computational processing introduces a novel approach to real-time applications in resource-constrained settings, contributing to the theoretical development of energy-efficient deep learning models. This research also sets a foundation for the scalability and adaptability of intelligent examination systems, ensuring their applicability across various sports disciplines and environmental conditions.

Practical Significance: The findings of this thesis hold significant practical implications for real-world sports examination scenarios. By developing a modular and scalable video-based intelligent sports examination system, the study provides a com-

prehensive solution that integrates robust keypoint prediction, accurate object tracking, and efficient computational processing. The proposed system is rigorously validated through real-world basketball skill assessments, demonstrating its effectiveness and robustness in practical settings. The advanced object tracking algorithms and efficient computational methods ensure precise tracking and rapid inference, addressing critical challenges in dynamic sports environments. The adaptability of the system to various sports disciplines and conditions underscores its practical relevance, offering a reliable tool for performance evaluation and skill assessment in diverse sports contexts. Moreover, this research lays the groundwork for the broader application of these methodologies, extending their benefits to other sports and dynamic environments. The study's contributions are pivotal in advancing the practical deployment of intelligent sports examination systems, ensuring accurate and reliable assessments that enhance the fairness and effectiveness of sports evaluations.

1.5 Thesis Structure

The structure of the thesis is shown in Figure 1.2 and the chapters are organized as follows:

- **CHAPTER 2**: This chapter presents a comprehensive literature review pertinent to this research. It introduces fundamental concepts and methodologies in human skeleton keypoint recognition, object detection, and Re-ID. It also covers memristor technology and its applications, along with a review of artificial intelligence in sports, highlighting relevant advancements and challenges. The review provides a foundation for understanding the current state of the art and identifies gaps that this research aims to address.

- **CHAPTER 3**: This chapter tackles the research objective of enhancing the robust-

ness of keypoint prediction in dynamic sports environments (RO1). It proposes integrating STGNP with AlphaPose to improve keypoint prediction accuracy by handling occlusions and correcting anomalies in predicted keypoints. Experimental results demonstrate the effectiveness of this approach in dynamic sports settings, addressing RQ1. The chapter details the methodology, implementation, and evaluation of STGNP, showcasing its impact on keypoint prediction robustness.

- **CHAPTER 4**: This chapter focuses on developing advanced object detection and tracking algorithms that incorporate positional information to enhance tracking accuracy amidst frequent occlusions and complex interactions (RO2). By integrating the spatial and temporal context of candidates and incorporating body information, the proposed algorithms ensure precise tracking of examination props. The chapter discusses challenges of traditional object detection and Re-ID methods and how the proposed approach overcomes these issues, addressing RQ2. It includes a detailed analysis of the algorithm's performance in real-world sports environments.

- **CHAPTER 5**: This chapter discusses the development of efficient computational methods suitable for applications using Lie Algebra and Memristor-Augmented LSTM and CNN (RO3). It presents a novel approach to enhance skeleton-based human motion recognition, achieving faster inference and reduced computational requirements, which is crucial for resource-constrained settings. The chapter elaborates on the theoretical underpinnings of Lie Algebra, the design of the memristor-based structure, and the integration with LSTM and CNN, addressing RQ3. Experimental validation and performance metrics are provided to demonstrate the efficacy of the proposed method.

- **CHAPTER 6**: This chapter aims to develop a modular and scalable intelligent sports examination system, ensuring practical applicability and adaptability across various sports and dynamic environments (RO4). It integrates robust human

pose estimation, accurate object tracking, and efficient computational processing into a comprehensive system. The system's performance is evaluated in real-world basketball skill assessments, providing insights into its effectiveness and robustness. The chapter discusses the design principles, modular architecture, and scalability features of the system, addressing RQ4. It also includes case studies and real-world application scenarios to highlight the system's versatility.

- **CHAPTER 7**: This chapter summarizes the findings of this thesis and suggests directions for future work. It consolidates the research contributions, discusses the implications of the results, and identifies potential areas for further exploration and development in the field of intelligent sports examination systems. The chapter emphasizes the significance of the research and outlines the steps needed to advance the current state of the art.

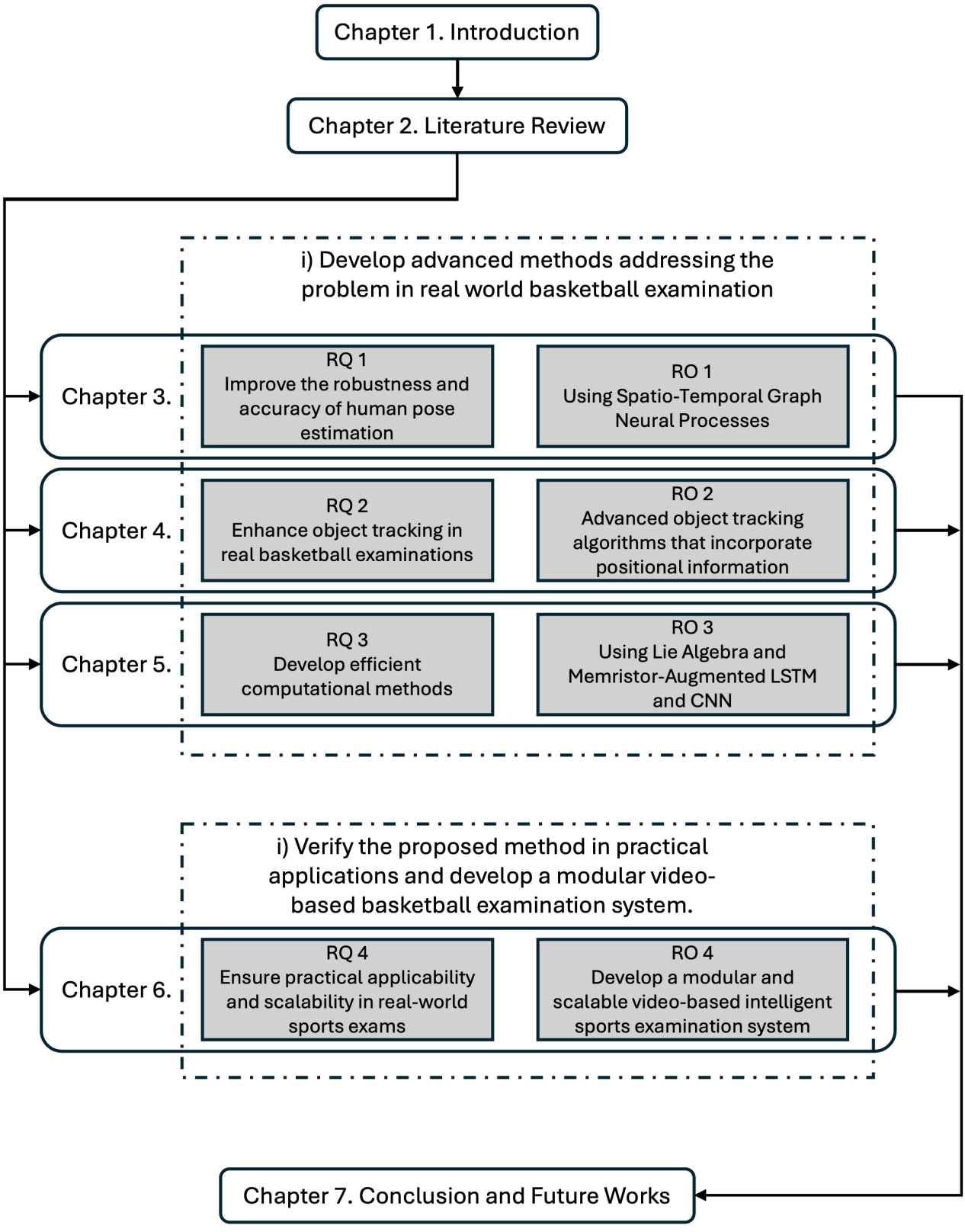

Figure 1.2: Thesis Structure.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a detailed survey of the literature relevant to this research. As shown in Figure 2.1 and Table 2.1, Section 2.1 examines human pose estimation, covering its core concepts, methodologies, and recent advancements. Section 2.2 investigates object detection techniques, addressing challenges like occlusions and the presence of multiple identical objects. In Section 2.3, Re-ID techniques are analyzed, focusing on their principles and applications. Section 2.4 discusses the potential of memristor-based neural networks for enhancing computational efficiency. Finally, Section 2.5 reviews related research, focusing on the application of these technologies in sports and considering their broader impact and future potential.

2.1 Human Pose Estimation

Human pose estimation (HPE) plays a vital role in the field of computer vision, vital for applications such as activity recognition and human-computer interaction. In the context of intelligent examination systems, particularly in sports, accurate HPE is essential for assessing and analyzing performance. This section provides a comprehensive exploration

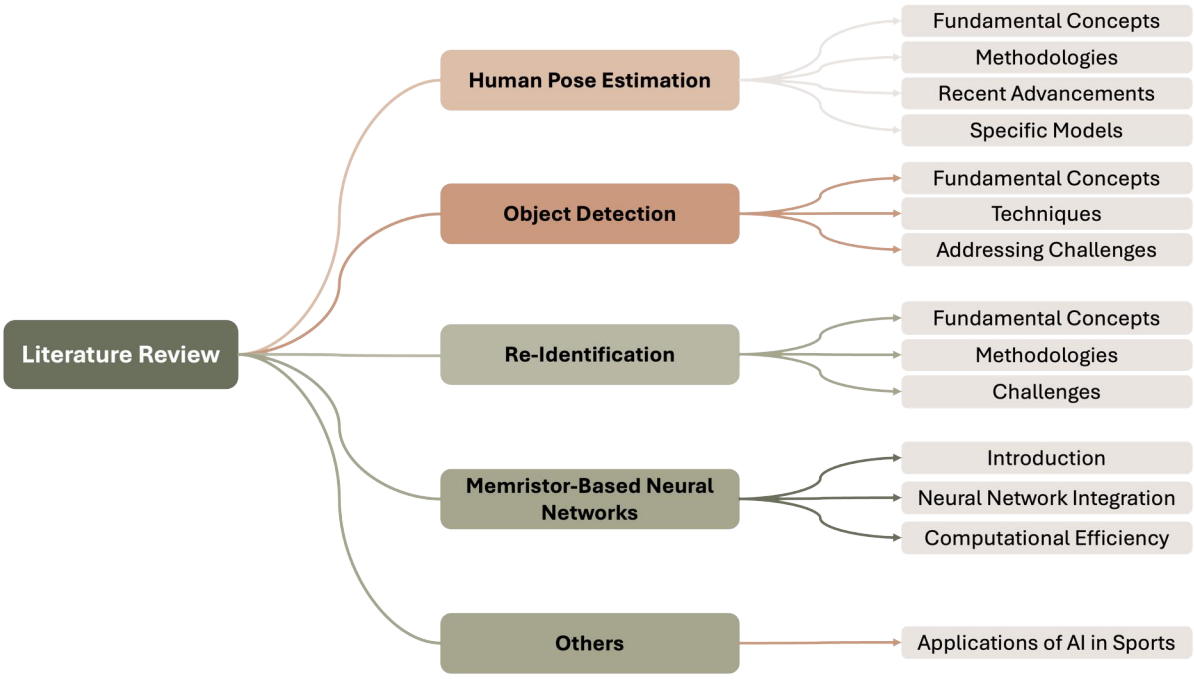

Figure 2.1: Overview of literature review.

of HPE, starting with fundamental concepts and terminologies, followed by an overview of the methodologies employed in this field, and concluding with recent advancements that have significantly propelled this research area forward.

2.1.1 Fundamental Concepts

HPE aims to locate and represent human body parts, constructing a detailed human body representation (e.g., body skeleton) from input data such as images and videos Cao et al. (2017). This task is foundational in computer vision, providing crucial geometric and motion information about the human body that is applied across various domains, including human-computer interaction, motion analysis, augmented reality (AR), and virtual reality (VR) Pavlakos et al. (2017). The primary objective of HPE is to accurately estimate the spatial configuration of human body parts from sensor data, particularly images and videos Newell et al. (2016). Recent advancements in HPE have been significantly driven by the development of deep learning-based solutions Sun et al. (2019).

Category	Description
Human Pose Estimation	Used for analyzing player movements and biomechanics in basketball skill assessments. Advanced models like STGNP, AlphaPose, and OpenPose enhance accuracy by handling occlusions and rapid motion, improving keypoint estimation reliability.
Object Detection	Identifies and tracks key elements such as basketballs, hoops, cones, and players. YOLOv8 ensures real-time, high-precision detection, with candidate position integration improving robustness against occlusions and interactions.
Re-Identification (Re-ID)	Ensures consistent tracking of identical objects (e.g., basketballs, cones) across frames and views. Techniques like ByteTrack and BoT-SORT reduce identity mismatches and enhance object continuity in dynamic scenarios.
Memristor-Based Neural Networks	Enhances computational efficiency by leveraging Lie Algebra and Memristor-Augmented LSTM/CNN. This reduces energy consumption and accelerates inference, making real-time sports analysis feasible in resource-limited settings.
Others (Applications of AI in Sports)	Highlights AI's role in modern sports analytics, optimizing skill assessments, training feedback, and decision-making. Advances in pose estimation, object detection, and computation improve fairness and efficiency in sports examinations.

Table 2.1: Hierarchical Relationships Among Literature Review Components in the Intelligent Sports Analysis System.

These approaches have demonstrated superior performance in both 2D and 3D pose estimation Xiao et al. (2018). Despite these advancements, challenges such as occlusion, depth ambiguities, and insufficient training data remain Mehta et al. (2017); Zhou et al. (2018). The adoption of deep learning methods has transitioned the emphasis from

manually engineered features and graphical models to approaches that are more reliant on data, utilizing extensive datasets and sophisticated neural network frameworks to enhance both precision and resilience.

Deep learning frameworks have introduced significant improvements in HPE, setting new benchmarks. Early methods, constrained by their limited ability to generalize, have given way to techniques utilizing CNNs and recurrent neural networks (RNNs), which have proven pivotal in enhancing performance. Zheng et al. provide a comprehensive review, systematically analyzing and comparing over 260 research papers, highlighting significant strides made in the field Zheng et al. (2023). Transformer-based architectures, with their self-attention mechanisms, have shown promise in capturing long-range dependencies and contextual information, thus improving robustness against occlusions and complex interactions Li et al. (2021); Xu et al. (2022). Researchers have also explored auxiliary information, such as depth data and multi-view imagery, to enhance accuracy. Multi-task learning, which optimizes for pose estimation and related tasks simultaneously, has been employed to leverage shared representations and improve overall performance Papandreou et al. (2017); Yang et al. (2020). Additionally, innovative data augmentation techniques and enhanced loss functions have been developed to address persistent challenges in the field. These advancements demonstrate the continuous evolution and potential of HPE technologies in addressing complex real-world scenarios.

2.1.2 Methodologies

Human pose estimation methodologies have undergone significant advancements in recent years, primarily driven by deep learning techniques. CNNs form the foundation of many HPE models, learning hierarchical features directly from large datasets to capture the complex patterns of human body parts in images effectively. Models like Simple Baselines Xiao et al. (2018) employ CNNs to refine pose estimates through upsampling

layers, while heatmap regression techniques, as seen in Stacked Hourglass Networks Newell et al. (2016), predict heatmaps for each keypoint, representing the likelihood of a keypoint's presence at each pixel location.

Graph Neural Networks (GNNs) have also contributed significantly by modeling spatial relationships between body parts. Representing the human body as a graph, with nodes corresponding to keypoints and edges representing spatial relationships, GNNs can capture dependencies and improve pose estimation accuracy. For example, methods like VNect Mehta et al. (2017) and MonoCap Zhou et al. (2018) extend 2D pose estimation to 3D by predicting the depth of each keypoint from monocular images or video sequences, providing a comprehensive understanding of human poses in three-dimensional space.

Recent methodologies encompass multi-person pose estimation, temporal models, and self-supervised learning techniques. Multi-person pose estimation, tackled by techniques such as OpenPose Cao et al. (2017) and HRNet Sun et al. (2019), involves estimating poses for multiple people in an image using bottom-up and top-down approaches. Temporal models, including RNNs and LSTM networks, capture motion information across frames in video-based pose estimation, enhancing the consistency and accuracy of pose estimates over time. Addressing the challenge of limited labeled data, self-supervised and semi-supervised learning methods leverage unlabeled data to pre-train models or use a combination of labeled and unlabeled data during training to improve performance.

Transformer-based architectures have also emerged as powerful tools in HPE, capturing long-range dependencies and contextual information. Models like ViTPose leverage the attention mechanisms in transformers to improve robustness against occlusions and complex interactions Xu et al. (2022). Additionally, multi-task learning approaches optimize for pose estimation and related tasks simultaneously, leveraging shared representations to enhance overall performance Yang et al. (2020).

Evaluation metrics for human pose estimation include Percentage of Correct Key-

points (PCK), Average Precision (AP), and Mean Squared Error (MSE), which assess the accuracy and performance of these methodologies in different contexts. These advancements collectively demonstrate the continuous evolution and potential of HPE technologies in addressing complex real-world scenarios.

2.1.3 Recent Advancements

Recent advancements in HPE have been significantly influenced by the development of deep learning techniques. These advancements have enabled substantial improvements in both 2D and 3D pose estimation accuracy and robustness. Deep learning methods have effectively addressed many challenges, such as occlusion, depth ambiguities, and insufficient training data, which have traditionally hindered HPE performance.

One of the significant advancements is the use of CNNs for learning hierarchical features from large datasets. Models like Simple Baselines Xiao et al. (2018) and Stacked Hourglass Networks Newell et al. (2016) utilize CNNs to refine pose estimates through upsampling layers and heatmap regression, respectively. These models predict heatmaps for each keypoint, representing the likelihood of a keypoint's presence at each pixel location. This approach has proven effective in capturing complex patterns of human body parts in images.

GNNs have also made a notable impact by modeling the spatial relationships between body parts. By representing the human body as a graph, with nodes corresponding to keypoints and edges representing spatial relationships, GNNs can capture dependencies and improve pose estimation accuracy. Methods like VNect Mehta et al. (2017) and MonoCap Zhou et al. (2018) extend 2D pose estimation to 3D by predicting the depth of each keypoint from monocular images or video sequences, providing a comprehensive understanding of human poses in three-dimensional space.

Transformer-based architectures have emerged as powerful tools in HPE, capturing

long-range dependencies and contextual information. Models like ViTPose Xu et al. (2022) leverage the attention mechanisms in transformers to improve robustness against occlusions and complex interactions. Additionally, multi-task learning approaches optimize for pose estimation and related tasks simultaneously, leveraging shared representations to enhance overall performance Yang et al. (2020).

In recent years, methodologies for multi-person pose estimation, temporal models, and self-supervised learning techniques have also advanced. Multi-person pose estimation, tackled by techniques such as OpenPose Cao et al. (2017) and HRNet Sun et al. (2019), involves estimating poses for multiple people in an image using bottom-up and top-down approaches. Temporal models, including RNNs and LSTM networks, capture motion information across frames in video-based pose estimation, enhancing the consistency and accuracy of pose estimates over time. To address the challenge of limited labeled data, self-supervised and semi-supervised learning methods leverage unlabeled data to pre-train models or use a combination of labeled and unlabeled data during training to improve performance.

These recent advancements collectively demonstrate the continuous evolution and potential of HPE technologies in addressing complex real-world scenarios, paving the way for more accurate and reliable human pose estimation systems.

2.1.4 Specific Models in Human Pose Estimation

AlphaPose, **MMPose**, and **OpenPose** are three notable frameworks that have significantly advanced the field of HPE, especially in practical applications:

- **AlphaPose** is renowned for its high precision and efficiency in real-time human pose estimation. It employs a top-down approach, where a person detector first identifies human bounding boxes, followed by a single-person pose estimator that detects keypoints within each bounding box. AlphaPose is widely used in applications such as sports

analytics and video surveillance due to its accuracy and robustness.

- **MMPose**, part of the OpenMMLab project, provides a comprehensive toolbox for pose estimation. It provides a wide range of advanced models and comprehensive tools for training, evaluating, and deploying HPE models. MMPose is designed for flexibility and extensibility, making it suitable for both research and industrial applications. Its modular design allows easy integration and customization, catering to a wide range of pose estimation tasks.

- **OpenPose** is one of the most popular open-source frameworks for multi-person pose estimation. It utilizes a bottom-up approach, where keypoints for all individuals in an image are detected simultaneously, followed by a part affinity field to associate the detected keypoints with individual persons. OpenPose is extensively used in applications ranging from entertainment to healthcare due to its robustness, versatility, and ability to handle complex multi-person scenarios.

These models collectively push the boundaries of human pose estimation technology, enhancing its accuracy, efficiency, and applicability across various domains. Their development and widespread use demonstrate the practical potential of HPE in real-world applications.

2.2 Object Detection

2.2.1 Fundamental Concepts

Object detection, a fundamental task in computer vision, involves the concurrent classification and localization of objects within images or video streams. Essential components of object detection include bounding box regression, which delineates the object's extent, and confidence scores, reflecting the probability that a given bounding box contains an

object of interest. The Intersection over Union (IoU) metric is pivotal in evaluating detection performance, measuring the overlap between predicted and ground truth bounding boxes Zou et al. (2023). Recent advancements have focused on improving the accuracy and robustness of bounding box regression. For instance, Guo et al. Guo et al. (2023) proposed the Balanced Corner-IoU (BC-IoU) loss and Point Offset Module (POM) branch in the Scale Adaptive Network (SAN), enhancing small object detection performance. Another approach unifies classification and bounding box regression heads to achieve better overall precision, as demonstrated by Gao et al. Gao et al. (2022).

The limitations of traditional IoU metrics have also been addressed by introducing generalized versions. Rezatofighi et al. Rezatofighi et al. (2019) presented the Generalized IoU (GIoU) to optimize non-overlapping bounding boxes, while Wen et al. Wen et al. (2022) proposed the Adaptive IoU (AIoU) method, which improves localization performance. Moreover, specialized IoU-based loss functions have been developed to enhance detection in specific contexts. For example, Cai et al. Cai et al. (2023) introduced the Corner-point and Foreground-area IoU loss (CFIoU) for small object detection, and Wang et al. Wang and Song (2021) proposed the Improved Loss based on Complete IoU (ICIoU) to improve bounding box regression accuracy. The continual evolution of object detection methodologies reflects the dynamic nature of this field, marking significant milestones in both accuracy and efficiency.

2.2.2 Techniques

The evolution of object detection techniques has transitioned from traditional methodologies to advanced deep learning paradigms, each contributing significantly to the field's progression.

2.2.2.1 Traditional Methods

Initial methods primarily utilized exhaustive search strategies like sliding windows combined with handcrafted feature descriptors such as Histogram of Oriented Gradients (HOG) Dalal and Triggs (2005). These features were then classified using machine learning algorithms, most notably Support Vector Machines (SVM) Cortes and Vapnik (1995). Despite their innovation, these approaches encountered limitations in computational efficiency and robustness across varying object scales and poses. The Deformable Parts Model (DPM) Felzenszwalb et al. (2008) enhanced detection by modeling objects as a collection of parts, but still faced challenges with computational demands and real-time application viability.

2.2.2.2 Deep Learning-based Methods

The advent of deep learning has revolutionized object detection, primarily through CNNs. Significant advancements include:

- **R-CNN Family**: Starting with R-CNN Girshick et al. (2014), which employs selective search for region proposals followed by CNN-based feature extraction and classification. Fast R-CNN Girshick (2015) enhanced efficiency by combining region proposal generation and feature extraction within a unified network. Building on this, Faster R-CNN Ren et al. (2015) further optimized the process by incorporating a Region Proposal Network (RPN), making it more suitable for real-time tasks. Mask R-CNN He et al. (2017) extended Faster R-CNN by adding a parallel branch for predicting segmentation masks, facilitating instance segmentation.

- **Single-Shot Detectors**: Approaches like You Only Look Once (YOLO) Redmon et al. (2016) and Single Shot MultiBox Detector (SSD) Liu et al. (2016b) eliminate region proposals, opting for direct object localization and classification in a single

forward pass. YOLO's unified architecture provides real-time detection capabilities, while SSD improves this with multi-scale feature maps for enhanced detection of objects at various scales.

- **Feature Pyramid Networks (FPN)**: Proposed by Lin et al. (2017), FPNs employ a top-down architecture with lateral connections to create high-level semantic feature maps at multiple scales, significantly improving the detection of objects of various sizes.

- **Anchor-Free Methods**: Recent advancements, such as CenterNet Zhou et al. (2019) and FCOS Tian et al. (1904), have introduced anchor-free methods that forgo predefined anchor boxes, instead focusing on direct prediction of object centers and related attributes. This approach reduces computational complexity and enhances detection performance.

- **Transformers in Object Detection**: Vision Transformers (ViTs) Dosovitskiy et al. (2020) and Detection Transformers (DETR) Carion et al. (2020) have introduced a paradigm shift by leveraging self-attention mechanisms for object detection, achieving state-of-the-art performance without the need for region proposals or anchor boxes.

Overall, these deep learning-based methods have significantly advanced object detection by enhancing accuracy, robustness, and real-time performance. The continuous evolution in model architectures and training techniques promises even greater capabilities in the future.

2.2.3 Addressing Challenges

Despite significant advancements, object detection continues to face numerous challenges that require ongoing research and innovation.

2.2.3.1 Occlusions

Occlusions remain a substantial challenge, often resulting in partial visibility of objects and degrading detection performance. To address this, advanced strategies such as multi-scale feature fusion, which integrates information across different levels of the network, have been developed Saleh et al. (2021). These strategies enhance the ability to detect partially visible objects by improving the network's overall representation. Additionally, attention mechanisms dynamically focus on the most relevant parts of the image, selectively enhancing the representation of visible object parts. Techniques such as Boundary-Aware Convolutional Networks (BACNs), which emphasize boundary information through global feature fusion, have also been proposed to improve the accuracy of occluded object detection.

2.2.3.2 Small Object Detection

Detecting small objects is inherently difficult due to their minimal pixel representation and the high likelihood of being overshadowed by larger objects or background noise. Effective techniques to address this challenge include constructing multi-resolution feature pyramids that preserve fine-grained details and enhancing context information to provide additional cues about the presence of small objects Li et al. (2020b). For instance, models like YOLO-ACN, which incorporate attention mechanisms and advanced loss functions such as Complete Intersection over Union (CIoU), significantly improve the detection accuracy of small and occluded objects. Moreover, integrating super-resolution techniques with dynamic feature fusion has shown promise in enhancing detection capabilities for small objects by effectively increasing the resolution and clarity of the target objects Noh et al. (2019).

2.2.3.3 Multiscale Object Detection

Object detection across multiple scales presents a complex challenge due to the varying sizes of objects within a single image. Techniques such as FPNs utilize a top-down architecture with lateral connections to build high-level semantic feature maps at multiple scales Lin et al. (2017). This approach significantly improves the detection of objects of different sizes, ensuring that small objects are detected with high accuracy without compromising the detection of larger objects.

2.2.3.4 Illumination Variations and Background Interference

Variations in lighting and complex backgrounds introduce significant noise, complicating the detection process. Advanced data augmentation techniques, including random cropping, color jittering, and illumination adjustments, enhance the robustness of models against these variations. Furthermore, robust feature extraction methods that can differentiate between objects and background noise, coupled with models like BACNs that integrate global context, significantly improve the ability to accurately detect objects in challenging lighting and background conditions Fan et al. (2024).

2.2.3.5 Data Annotation and Model Generality

The acquisition and quality of labeled data are crucial for the performance of deep learning models, yet labeling data is typically expensive and time-consuming. To mitigate this, semi-supervised and unsupervised learning approaches, which utilize a combination of labeled and unlabeled data, are being developed to improve model training efficiency Li et al. (2024). These methods reduce the dependency on extensive labeled datasets. Additionally, achieving both generality and real-time capabilities in object detection systems is challenging. Domain adaptation techniques are essential for enhancing the versatility and applicability of object detection models across various application domains,

allowing models trained on one dataset to generalize to different domains.

In summary, ongoing research and innovation in multi-scale feature fusion, attention mechanisms, advanced loss functions, and robust data augmentation techniques continue to address these challenges, pushing the boundaries of object detection performance.

2.3 Re-ID Techniques

2.3.1 Fundamental Concepts

Re-ID involves several fundamental concepts that are crucial for identifying individuals across different views. One of the key concepts is feature extraction, where distinctive features are extracted to identify individuals. For example, Liao et al. Liao et al. (2015) emphasized the significance of local maximal occurrence (LOMO) representation in this context. Another essential concept is feature matching, which involves matching features between different views to re-identify individuals. Wang et al. Wang et al. (2018a) demonstrated how learning discriminative features with multiple granularities can enhance the effectiveness of feature matching. Additionally, metric learning is a critical component, focusing on measuring the similarity between extracted features. Li et al. Li et al. (2018) introduced a harmonious attention network to improve the accuracy of metric learning.

2.3.2 Methodologies

Re-ID methodologies can be broadly categorized into traditional handcrafted feature-based approaches and modern deep learning techniques. Traditional approaches rely on manually designed features such as color histograms and texture descriptors. Farenzena et al. Farenzena et al. (2010) discuss the symmetry-driven accumulation of local features as a significant approach in this category. On the other hand, modern approaches leverage

deep learning techniques, particularly CNNs and Generative Adversarial Networks (GANs). These methods have shown significant improvements in the robustness and scalability of Re-ID systems, as evidenced by Zheng et al. Zheng et al. (2019, 2017). Deep learning-based methods have revolutionized Re-ID by enabling the extraction of more discriminative and invariant features from images.

2.3.3 Challenges

Re-ID faces several challenges that impact its accuracy and robustness. One of the primary challenges is occlusion, which occurs when parts of the object of interest are blocked from view. Techniques to handle occlusions include local feature matching and attention mechanisms. For instance, Yang et al. Yang et al. (2019) propose a region attention network to effectively address occlusions. Another significant challenge is dealing with multiple identical objects, which requires fine-grained feature extraction and multi-task learning. Sun et al. Sun et al. (2018) emphasize the importance of refined part pooling for distinguishing between similar appearances. These challenges necessitate continuous advancements in Re-ID methodologies to ensure reliable performance in diverse real-world scenarios.

2.4 Memristor-based Neural Networks

2.4.1 Introduction to Memristor Technology

Memristors, short for memory resistors, are passive electronic components that retain a history of the voltage applied to them, embodying a non-volatile memory function. This characteristic allows them to emulate synaptic connections in neuromorphic computing systems, according to Wen et al. Wen et al. (2019). Unlike traditional transistors, which operate in a binary manner, memristors support analog computation through variable

resistance values. This capability enhances computational efficiency by enabling complex operations within memory units, minimizing the need for data transfer between processors and memory, as highlighted by Jo et al. Jo et al. (2009). Furthermore, their unique properties make memristors a promising component in the development of future computing architectures, offering potential solutions for the challenges faced in modern computing systems.

A significant feature of memristors is their ability to be tuned to specific resistance values by applying a voltage, which changes their conductivity. Remarkably, this change is retained even when power is turned off. By organizing memristors into a crossbar grid, numerous neural network computations can be performed in parallel, further leveraging their unique capabilities in neuromorphic computing architectures, as discussed by Taha et al. Taha et al. (2013). This parallelism significantly enhances the efficiency and scalability of neural network implementations. Memristors are also crucial for implementing synapses efficiently in neural networks, allowing for large-scale data processing with synaptic behaviors similar to human brain neurotransmitters, as highlighted by Secco et al. Secco et al. (2018). This efficient implementation of synaptic functions is vital for the development of advanced neuromorphic systems.

2.4.2 Neural Network Integration

The integration of memristor technology into neural networks has led to the development of various architectures, including LSTM networks and CNNs Wen et al. (2019). Memristor-based LSTM networks are particularly effective for temporal data analysis, capturing long-term dependencies crucial for tasks such as natural language processing and time series forecasting Liu et al. (2020b). These networks benefit from the inherent parallelism and low power consumption of memristor crossbars.

A typical LSTM network comprises multiple layers that process data in stages, with

each layer capturing different aspects of the input sequence, as described by Yakopcic et al. Yakopcic et al. (2016). Memristors can implement these layers efficiently, leveraging their ability to perform analog computations and retain state information without continuous power. Similarly, memristor-based CNNs are well-suited for spatial data analysis, such as image recognition, where they can perform convolutions and pooling operations in parallel, significantly accelerating the processing speed, as highlighted by Huang et al. Huang et al. (2018). Additionally, memristive devices are utilized in various neural network models, such as spiking, multilayer, and recurrent neural networks, showcasing their versatility in neuromorphic computing, as demonstrated by Yang et al. Yang (2014). These capabilities make memristors a key component in the advancement of efficient and powerful neural network architectures, providing a foundation for future innovations in the field.

2.4.3 Enhancing Computational Efficiency

One of the primary advantages of memristor-based neural networks is their enhanced computational efficiency, as noted by Yao et al. Yao et al. (2020). Memristors' ability to perform in-memory computing reduces the latency associated with data transfer between memory and processing units. This reduction in data movement not only speeds up computation but also lowers energy consumption, making memristor-based systems particularly attractive for edge computing applications, where power efficiency and quick data processing are crucial. The inherent parallelism in memristor arrays further boosts the computational throughput, enabling more complex and larger-scale neural network models to be implemented efficiently.

In practical implementations, memristor-based neural networks have demonstrated performance metrics close to those of traditional software-simulated networks. Hasan et al. Hasan et al. (2017) report that, for instance, in the context of human action recognition,

the integration of memristor circuits with LSTM and CNN architectures has shown minimal accuracy loss while significantly improving inference speed and reducing energy consumption. These improvements highlight the potential of memristor technology to revolutionize the deployment of deep learning models in real-time applications. The increased inference speed is especially advantageous for applications that demand quick decision-making, including autonomous vehicles, robotics, and real-time monitoring systems.

Furthermore, the use of fuzzy modeling to address device variation in multilevel memristors enhances the robustness of memristive neural networks. As discussed in the paper by Cui et al. Cui and Zhang (2019) on memristive synaptic circuits for deep convolutional neural networks, fuzzy modeling helps in compensating for the variations in the memristor devices, ensuring consistent performance across different manufacturing batches and operating conditions. This approach not only improves the reliability of memristor-based systems but also paves the way for their broader adoption in various practical applications. The advancements in addressing device variation are crucial for scaling up the production of memristor-based neural networks and integrating them into existing technological ecosystems.

2.5 Applications of AI in Sports

2.5.1 Overview of AI Applications in Sports

AI, particularly deep learning, has revolutionized sports technology through its ability to analyze and interpret vast amounts of data with high accuracy. One notable application is in human pose estimation, where deep learning models like CNNs and RNNs have been employed to enhance sports performance analysis despite challenges such as occlusion and crowded scenes Samkari et al. (2023). These models can capture the

intricate movements of athletes, providing valuable data for performance improvement, injury prevention, and tactical analysis. Furthermore, AI has significantly impacted video analysis programs like Coach's Eye and Dartfish, which assist in skill-based video capture, provide immediate feedback, and help game officials make informed decisions Gajendra (2023). These tools utilize AI to break down complex movements into understandable segments, offering coaches and athletes detailed insights into technique and form.

2.5.2 Specific Case Studies

TrackNet, a deep learning network, was developed to accurately track high-speed and tiny objects like a tennis ball in sports videos with impressive precision, recall, and F1-measure results Huang et al. (2019). This innovation enables more accurate analysis of player performance and game dynamics, offering deeper insights into aspects like shot accuracy and ball trajectory. Additionally, the BiGRU recognition model outperformed other deep learning networks in recognizing sport-related activities using multimodal wearable sensors, achieving a maximum accuracy of 99.62% Mekruksavanich and Jitpattanakul (2022). These sensors provide real-time data on athletes' physical conditions, helping to tailor training programs and prevent injuries. AI has also been employed in biomechanics for tasks such as evaluating faults in sports movements using Expert Systems and Artificial Neural Networks (ANNs), with applications in sports like javelin, discus throwing, shot putting, and football kicking Ratiu et al. (2010). These applications aid in refining athletes' techniques, reducing the risk of injury, and enhancing overall performance.

In another instance, deep learning-based approaches have shown enhancements in human activity recognition, including in sports technology, as demonstrated by the improved wolf swarm optimization with deep learning-based movement analysis and

self-regulated human activity recognition technique Thanarajan et al. (2023). This technique allows for more precise monitoring of athletes' movements, identifying areas for improvement and optimizing training routines. Moreover, AI applications in sports training have improved the effectiveness of training activities, providing detailed analysis and personalized guidance for athletes Xianguo and Cong (2021). By analyzing performance data, AI can create customized training plans that address individual athletes' strengths and weaknesses, leading to more effective and efficient training sessions.

2.5.3 Broader Impact and Future Potential

The potential of deep learning in sports technology extends to enhancing decision-making processes and performance analysis. For instance, deep reinforcement learning (DRL) has been applied in sports game design, specifically in the ball return decision of a table tennis robot, demonstrating higher accuracy rates in returning the ball and practical applications for IoT fitness and sports technology development Wang et al. (2022b). These advancements can lead to more intelligent sports equipment and training tools, providing athletes with immediate, actionable feedback. Furthermore, AI has been utilized in sports to monitor athletes' physical conditions, analyze sports data, and provide real-time event performance analysis, thus enhancing sports precision and maximizing athletes' physical function Huang (2022). Real-time analytics can help coaches make strategic decisions during games, improving the team's chances of success.

Moreover, the integration of AI in sports not only improves performance and analysis but also holds promise for future advancements in athlete training and game strategy optimization Zhao et al. (2023). Future applications may include more sophisticated predictive models that can simulate game scenarios and suggest optimal strategies. AI's application in sports also includes ethical considerations, ensuring fairness and addressing impacts on stakeholders like players, officials, and administrators Suman

(2022). As AI continues to integrate into various sectors, it is essential to confront challenges concerning data privacy, biases in algorithms, and the implications of AI on the fairness of sports.

2.5.4 Challenges and Future Directions

While the applications of AI in sports are promising, numerous challenges persist. These include identity mismatches due to similar appearances, motion blur from rapid movement, and occlusions by other players or objects, which pose considerable obstacles Zhao et al. (2023). The development of algorithms capable of real-time, accurate identification and tracking under such conditions remains an active field of research. Moreover, the task of standardizing datasets across different sports is complex, as each sport's distinct technical elements and rules complicate the creation of a uniform benchmark for specific tasks. Establishing standardized, accessible, open-source, high-quality, and extensive datasets is essential for furthering research and enabling accurate comparisons among various models and techniques in sports analytics Zhao et al. (2023).

Additionally, the sports sector produces a large volume of detailed data through sensors and IoT devices. Present data processing techniques mainly target computer vision and have yet to fully leverage the capabilities of comprehensive deep learning approaches. To optimize the utilization of these rich data sources, it is imperative to develop methods that integrate detailed sensor data with visual data Zhao et al. (2023). This convergence of diverse data streams could foster more thorough and perceptive analyses, significantly advancing sports performance research. Integrating these data types offers a holistic perspective on an athlete's performance, encompassing biomechanics, physiological aspects, and environmental factors.

Future research should focus on integrating multi-modal data and multi-task learning, developing foundational models, and generating high-quality synthetic data. The

success of models like ChatGPT and recent breakthroughs in large models for image segmentation indicate that merging these technologies for sports applications could be highly beneficial Zhao et al. (2023). The practical deployment of these technologies can enhance athletic performance, support real-time decision-making, and improve the experience for sports professionals and enthusiasts. By employing these sophisticated AI models, sports organizations can obtain deeper insights into performance metrics and make more informed decisions.

CHAPTER 3

STGNP FOR HUMAN SKETLON KEYPOINT PREDICTION

To address the challenges identified in RQ1 regarding the robustness of keypoint prediction in dynamic sports environments, this chapter aims to achieve RO1 by proposing the integration of STGNP with existing human pose estimation methods. Traditional methods for keypoint prediction often face significant obstacles due to occlusions, varying lighting conditions, and rapid movements. Furthermore, existing human pose estimation method like AlphaPose offers multiple model sizes, but in practical applications, the larger, more accurate models are often not feasible due to resource constraints. Opting for the smaller, lightweight models results in a loss of accuracy, leading to prediction inaccuracies. To mitigate these issues, we introduce a novel approach that leverages STGNP to enhance keypoint prediction accuracy by effectively handling occluded keypoints and correcting anomalies such as sudden coordinate changes.

Section 3.1 provides an introduction to our motivations and the challenges associated with keypoint prediction in dynamic sports environments. It discusses the limitations of current methodologies and the need for robust solutions. Section 3.2 introduces the definitions, notations, and the proposed integration of STGNP with human pose

estimation methods. This section explains how the proposed method addresses the identified challenges, particularly the trade-off between model size and accuracy, and improves keypoint prediction robustness. In Section 3.3, the proposed method is validated through experimental evaluations, demonstrating its effectiveness in real-world sports settings. Finally, Section 3.4 concludes this chapter, summarizing the key findings and their implications for improving keypoint prediction in dynamic sports environments.

3.1 Introduction

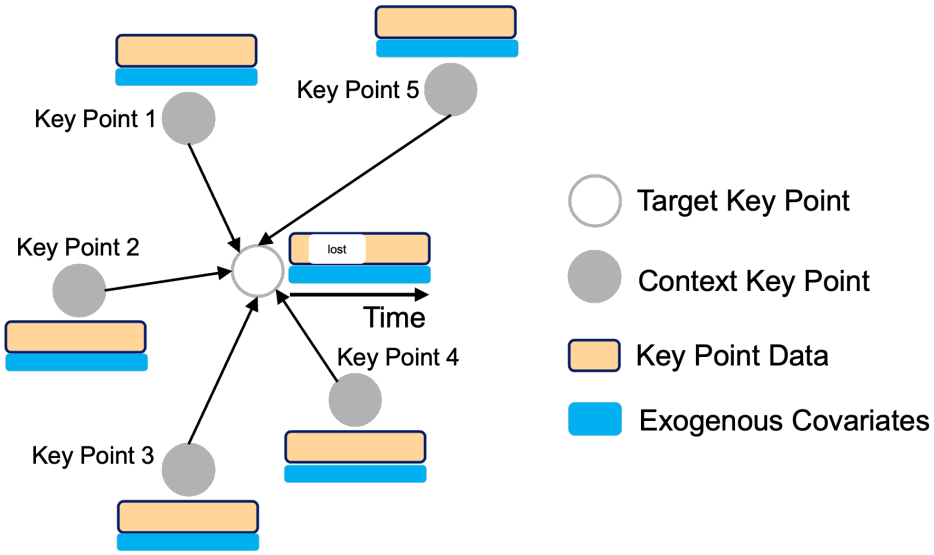

Figure 3.1: Data for the target keypoint is generated using context keypoints 1-5, considering both the graph structure and exogenous covariates.

In Human Pose Estimation, numerous excellent results have been achieved by models such as Alphapose Fang et al. (2022), MMPose Contributors (2020), and Openpose Cao et al. (2019), which have performed well on various human keypoint datasets. However, these models exhibit a common issue: keypoint drift. In practical applications, we have observed that these models tend to suffer from varying degrees of keypoint drift. Keypoint drift occurs when the prediction confidence of keypoints is relatively low during complex rotations or occlusions of human movements, leading to inaccurate

predictions and significant differences compared to adjacent frames. Human skeleton data exhibit both spatial features, due to their graph structure, and temporal features, as skeleton sequences form a time series. Therefore, human skeleton data can be considered spatio-temporal data. Inspired by the application of Spatio-Temporal Graph Neural Networks (STGNNs) in weather forecasting Lin et al. (2022), we apply this approach to enhance human skeleton prediction.

In this paper, we tackle the challenge of *spatio-temporal extrapolation for human pose estimation*. This process involves forecasting spatio-temporal data at target keypoints using the surrounding context nodes and related external covariates, all within a fixed graph structure composed of human skeleton keypoints, as depicted in Figure 3.1. For example, we use multiple human pose estimation models to predict pose data. We then extrapolate target keypoint data from context keypoints, considering covariates like confidence levels that can affect keypoint prediction.

To fulfill our objectives, it is crucial to address spatio-temporal correlations, which represent spatial interdependencies in a graph coupled with temporal dynamics over time. STGNNs are increasingly favored in this domain due to their robust learning capabilities Han et al. (2021); Wu et al. (2019). Nevertheless, Neural Networks (NNs), including STGNNs, face significant shortcomings: *(i) They do not inherently estimate uncertainties.* Incorporating such estimations is essential for dependable decision-making Wang et al. (2019); Wen et al. (2023), yet most NNs operate deterministically, failing to handle uncertainties. *(ii) Their generalization to novel scenarios is constrained.* While NNs demand extensive data to learn effectively, their parametric nature can restrict their flexibility in new or changing conditions without retraining. Furthermore, they are particularly sensitive to hyperparameter settings, requiring thorough tuning for best results.

The constraints of NNs have encouraged scholars to explore probabilistic models, notably Gaussian Processes (GPs) Seeger (2004). GPs establish a probabilistic process

where the spatio-temporal interactions are characterized using diverse kernels Patel et al. (2022). Their foundation in Bayesian statistics and their non-parametric approach allow effective handling of uncertainties and excellent generalization across various functions Luttinen and Ilin (2012). However, the expressivity of GPs' kernels may be restrictive, presenting certain drawbacks. To mitigate these challenges, Neural Processes (NPs) Garnelo et al. (2018) have been developed. NPs employ neural networks to formulate stochastic processes and introduce an aggregator for context integration, effectively merging the advantages of NNs and GPs. This integration makes NPs an attractive option for modeling complex spatio-temporal dynamics.

Regrettably, NPs are not readily adaptable to spatio-temporal graph data for several reasons: *(i) Their inefficiency in learning temporal dynamics.* Current implementations of NPs Singh et al. (2019); Qin et al. (2019) employ latent state transitions to recurrently grasp temporal patterns. Nonetheless, these transitions often focus predominantly on latent variables from earlier steps, overlooking essential context in subsequent sequences. This issue, termed transition collapse, can hinder effective learning across extended sequences Singh et al. (2019). *(ii) Their inability to effectively represent spatial connections within graphs.* The aggregation methods used in existing NPs Gordon et al. (2020); Kim et al. (2019); Volpp et al. (2020) do not sufficiently capture the intricate spatial relationships inherent in graph structures. Moreover, the deterministic nature of these processes proves less effective in handling data ambiguities, such as noise or missing entries, as illustrated in Figure 3.1.

To address these limitations, we introduce the STGNP, designed for spatio-temporal extrapolation across graphs. STGNP operates in two phases: initially, a deterministic network captures the spatio-temporal node representations. This is achieved not through recurrent architectures but by sequentially stacking convolution layers for temporal dynamics Aksan and Hilliges (2019), and employing cross-set graph neural networks

for spatial interactions. In the subsequent phase, state transitions are utilized to amalgamate the latent variables of target nodes in a top-down approach. These transitions maintain horizontal time independence and incorporate extensive temporal evolution from higher layers. Given that the number of transitions corresponds directly to the number of layers rather than the sequence length, our model inherently avoids the pitfalls of transition collapse.

In the transition's aggregator, we recognize that various context nodes exhibit differing significance levels. Inspired by Volpp et al. (2020), we introduce the GBA approach, which directly aggregates distributions over latent variables influenced by the graph's topology. This method posits that a context node's impact on the latent distribution diminishes if it is geographically distant from the target or if it shows significant ambiguity, as identified by the system. By incorporating the graph structure into the NP's aggregator, this strategy not only enhances the model,Äôs ability to handle node uncertainties but also improves its overall efficacy.

In conclusion, our principal contributions are as follows:

- We introduce the STGNP, a pioneering approach in extending Neural Processes to the realm of spatio-temporal graph analytics. STGNP uniquely excels in explicitly handling uncertainties and demonstrates robust generalization across various functions, presenting a significant advancement over traditional NN-based methods. It also adeptly learns temporal dynamics and understands graph-structured data, distinguishing it from conventional NP models.

- We develop the Graph Bayesian Aggregation technique, a Bayesian strategy for contextual node aggregation. This method effectively incorporates the graph topology and node uncertainties into the aggregation process, enhancing the model's accuracy and predictive quality.

- We validate STGNP by conducting rigorous experiments across multiple human skeletal datasets, and compare its performance with various foundational models. The results from these experiments confirm that STGNP not only significantly outperforms these models but also provides reliable uncertainty estimates, showcases exceptional generalizability, and maintains resilience against noisy inputs.

3.2 Preliminaries

Initially, we establish the definitions and notations for spatio-temporal graph data. Subsequently, we delve into the fundamental principles of neural processes.

3.2.1 Definitions and Notations

Definition 1 (Graph) A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is composed of a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$, which establish the connections and their weights between vertices. For any vertex $v \in \mathcal{V}$, its K-hop neighborhood, denoted by $\mathcal{N}_k(v)$, comprises vertices that can be reached from v within K steps. Using $\mathcal{E}$ and the specified K, a K-hop adjacency matrix A^K is constructed to quantify the non-Euclidean distances among connected neighbors.

Definition 2 (Spatio-Temporal Data) In a graph, signals are gathered from each node. Representing the data for node i, we define $Y_i = (y_{i,1}, .., y_{i,t}, .., y_{i,T}) \in \mathbb{R}^{T \times d_y}$, capturing measurements across a time window T with d_y as the feature dimension. The collective data for all nodes is denoted by $Y = (Y_1, .., Y_n, .., Y_N) \in \mathbb{R}^{N \times T \times d_y}$, where N represents the total number of nodes observed in the graph during the time window.

Definition 3 (Exogenous Covariates) Exogenous covariates, which often exhibit strong correlations with node data, enhance the learning process. These covariates can be sourced from various origins. For example, the confidence level can significantly

influence the outcome of keypoint predictions. We represent these covariates as a tensor $X \in \mathbb{R}^{N \times T \times d_x}$, taking them into explicit consideration in our analysis.

3.2.2 Neural Processes

NPs, as introduced by Garnelo et al. (2018), create stochastic mappings from inputs $x \in \mathbb{R}^{d_x}$ to outputs $y \in \mathbb{R}^{d_y}$, relying on a context set $\mathscr{C} = \{(x_n, y_n)\}_{n=1}^{N}$ of observed pairs. Functionally similar to Gaussian Processes, NPs differ in that the stochastic mappings are learned via neural networks rather than explicitly defined. NPs utilize a conditional latent variable model, where the latent variable z's distribution is determined by a conditional prior $p(z|\mathscr{C})$ that is learned from the context set. With target inputs $X_{\mathscr{D}} = \{x_m\}_{m=1}^{M}$ from a target set $\mathscr{D}$, a likelihood module $p(Y_{\mathscr{D}}|X_{\mathscr{D}}, z)$ is trained to predict the output $Y_{\mathscr{D}}$. The generative process of NPs is given by:

$$p(Y_{\mathscr{D}}|X_{\mathscr{D}}, \mathscr{C}) = \int p(Y_{\mathscr{D}}|X_{\mathscr{D}}, z) p(z|\mathscr{C}) dz. \tag{3.1}$$

Practically, NPs treat target variables independently, leading to a decomposed likelihood $p(Y_{\mathscr{D}}|X_{\mathscr{D}}, z)$ factored as $\prod_{m=1}^{M} p(y_m|x_m, z)$. This meta-learning structure, where each context-target pair $\{\mathscr{C}, \mathscr{D}\}$ forms a distinct stochastic process, enhances the model's generality with minimal parameter dependence. The stochastic processes are ensured by aggregating conditions $\mathscr{C}$ using a permutation-invariant function (e.g., mean, attention), as required by the Kolmogorov Extension Theorem Øksendal (2003). As direct computation of the latent variable z's marginalization is often impractical, the model typically employs Monte-Carlo (MC) sampling to approximate Equation 3.1 Foong et al. (2020) or uses a variational approach to maximize the evidence lower bound (ELBO) Garnelo et al. (2018):

$$\log p(Y_{\mathscr{D}}|X_{\mathscr{D}}, \mathscr{C}) \geq \mathbb{E}_{q(z|\mathscr{C} \cup \mathscr{D})} \left[\sum_{m=1}^{m} \log \frac{p(y_m|x_m, z) p(z|\mathscr{C})}{q(z|\mathscr{C} \cup \mathscr{D})} \right], \tag{3.2}$$

where $q(z|\mathscr{C} \cup \mathscr{D})$ is the approximated posterior and $p(y_m|x_m, z)$ is the likelihood, both learned through neural networks. To address the intractability of the true conditional prior $p(z|\mathscr{C})$, the same approximation function $q(\cdot)$ is utilized to estimate $p(z|\mathscr{C}) \approx q(z|\mathscr{C})$.

3.3 Methodology

In this section, we introduce STGNP, a neural latent variable model designed to improve spatio-temporal extrapolation. As depicted in the graphical model in Figure 3.2, STGNP operates in two phases: the first involves learning deterministic representations (STRL), and the second focuses on stochastic latent variables (GBA). We commence by outlining the challenge of spatio-temporal extrapolation. This is followed by an explanation of the deterministic phase where spatio-temporal representations are developed, and the derivation of Graph Bayesian Aggregation for context aggregation in the stochastic phase. We conclude with a discussion on the generative process and the optimization methods employed. For clarity and conciseness, we limit our discussion to a single target node m in subsequent sections.

3.3.1 Problem Statement

In this work, we adapt the Neural Processes framework to address spatio-temporal extrapolation. Initially, we define the context set $\mathscr{C}$, which includes nodes characterized by exogenous covariates and observed data $\{(X_n, Y_n)\}_{n=1}^{N} \in \mathbb{R}^{N \times T \times (d_x + d_y)}$. Our objective is to construct a posterior predictive distribution $p(Y_{\mathscr{D}}|X_{\mathscr{D}}, \mathscr{C}, A)$ that can predict outcomes $Y_{\mathscr{D}} \in \mathbb{R}^{M \times T \times d_y}$ for the target set D over an identical time interval. Here, M represents the number of target nodes. This predictive task utilizes the covariates $X_{\mathscr{D}}$, the information from the context set $\mathscr{C}$, and the adjacency matrix A. Throughout this paper, we use the subscript m for target nodes and n for context nodes, and the terms location, node, and sensor are used synonymously.

3.3.2 Spatio-Temporal Extrapolation

Spatio-temporal extrapolation aims to predict environmental states using available data. Historically, methods like K-Nearest Neighbors (KNN) and Random Forest (RF) have been employed to address this challenge, with KNN focusing on linear relationships and RF on non-linear dependencies Fawagreh et al. (2014). Despite their effectiveness in modeling spatial relationships, these techniques often fail to capture more dynamic, complex correlations. GPs, which formulate stochastic processes with versatile kernels, attempt to address this by adapting to diverse data features Seeger (2004); Li et al. (2020a). For example, Patel et al. Patel et al. (2022) implement periodic and Hamming distance kernels for different feature types. Nonetheless, the specificity of these kernels and their computational demands limit broader application. Other methods treat spatio-temporal extrapolation as a tensor completion problem, leveraging low-rank matrix assumptions to efficiently identify patterns Yu et al. (2016). While efficient, these methods are transductive and cannot generalize beyond the training dataset's nodes. In contrast, NNs have emerged as a predominant approach, exemplified by Cheng et al. Cheng et al. (2018) who use attention models to infer air quality from dynamic and static data through RNNs and MLPs, and by Han et al. Han et al. (2021) who enhance GCNs with multi-channel attention modules. However, NNs often face challenges with uncertainty and may overfit in data-scarce scenarios. Some NNs address extrapolation akin to kriging, with methods that include or exclude temporal dynamics and exogenous covariates, as shown by Appleby et al. Appleby et al. (2020) and Wu et al. Wu et al. (2021). Unlike these methods, our approach successfully integrates both spatial relationships and temporal dynamics.

3.3.3 Neural Processes Family

NPs integrate the strengths of NNs and GPs by combining potent learning capabilities with reliable uncertainty assessments Garnelo et al. (2018). NPs introduce latent variables across the context set, creating a conditional latent variable model, and employ a likelihood function to produce target predictions. Le et al. Le et al. (2018) highlighted that Neural Processes often struggled with underfitting, which was linked to the limitations of the aggregation functions they employed, such as using a simple mean or sum. Addressing this, Kim et al. Kim et al. (2019) developed Attentive Neural Processes (ANP), enhancing the model's ability to discern critical elements within and between the context and target sets. Progressing further, Kim et al. Kim et al. (2022) introduced a stochastic attention mechanism that uses Bayesian inference for weight determination, and Volpp et al. Volpp et al. (2020) designed a stochastic aggregator for direct context variable integration. Nonetheless, these advancements primarily focus on spatial considerations and do not extend to graph-structured data. Singh et al. Singh et al. (2019) then shifted the focus to sequential data, proposing Sequential Neural Processes (SNP) which incorporate state transitions through a variational recurrent neural network (VRNN) Chung et al. (2015) to model temporal sequences stochastically. Yoon et al. Yoon et al. (2020) later added Recurrent Memory Reconstruction to address distribution shifts in sequences. Despite these innovations, the issue of transition collapse remains a significant challenge in learning temporal relationships over extended periods. Our approach mitigates this by employing causal convolutions to better handle temporal dynamics.

3.3.4 Spatio-Temporal Representation Learning

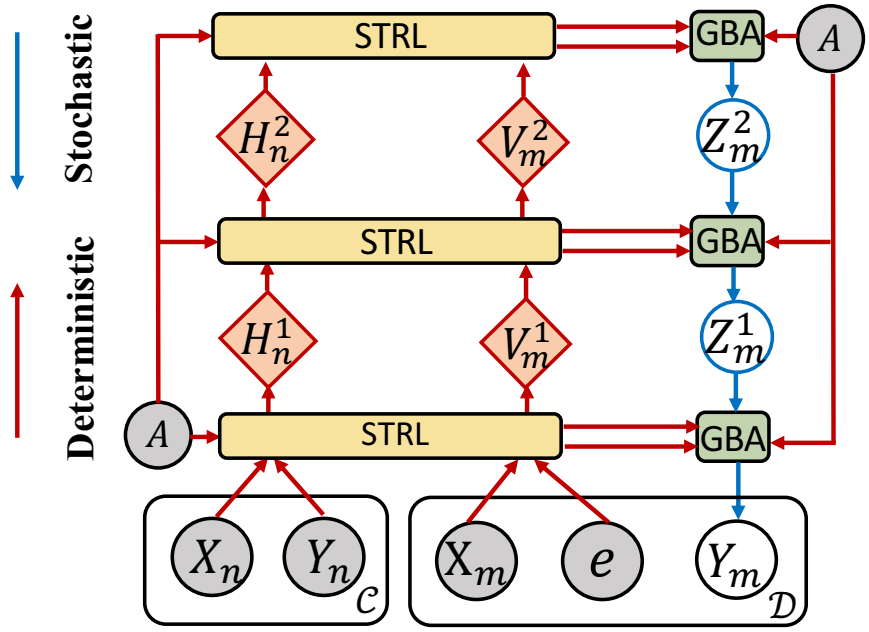

Figure 3.2: The graphical model depicted consists of three layers for clarity. The variables V_m^l, Z_m^l, and H_n^l in $\mathbb{R}^{T \times d_l}$ represent the deterministic representations and latent variables for a target node m, and the representations for a context node, respectively. The variable e in $\mathbb{R}^{d_0}$ is a learnable target token. The diagram also includes a shadowed circle indicating an observed variable, while the labels STRL and GBA refer to the processes of Spatio-Temporal Representation Learning and Graph Bayesian Aggregation, respectively.

The deterministic phase of the model consists of three core components designed to encapsulate both spatial and temporal correlations: a learnable target token, dilated causal convolution, and cross-set graph convolution. We detail each component separately before presenting an overview of the entire learning architecture.

Learnable Target Token Our model processes inputs of human pose estimation keypoints and associated covariates; however, the data Y_m for a specific target keypoint remains undisclosed. Conventional techniques often preprocess this by either substituting zeros Appleby et al. (2020); Wu et al. (2021) or using linear interpolation to approximate the values Hu et al. (2021). The zero-filling approach is unsuitable for

human pose estimation as it requires complete keypoint data for subsequent analyses. Additionally, interpolation tends to introduce substantial errors, adversely affecting the model's efficacy. Drawing inspiration from the Masked AutoEncoder He et al. (2022), our approach employs a shared learnable token $e \in \mathbb{R}^{d_0}$ to represent target node embeddings and utilizes an embedding layer with parameter $W \in \mathbb{R}^{d_y \times d_0}$ for context nodes. This token is dynamically refined by the network to accurately represent the target node's position in the feature space, thereby circumventing the drawbacks of traditional preprocessing methods.

Cross-Set Graph Convolution Layer Graph convolution is a fundamental technique for capturing spatial relationships within graph structures. Traditional GCN approaches often assume equal dependency across all nodes Wu et al. (2021); Hu et al. (2021), yet in our scenario, the interactions between the target and context sets are critical due to their impact on target keypoints. With this understanding, we suggest that neglecting intra-set relations does not detrimentally impact our model's effectiveness. Thus, we introduce the cross-set graph convolution (CSGCN), which focuses exclusively on interactions between the sets $\mathscr{C}$ and $\mathscr{D}$. In detail, the update process for the target keypoint representation $V_m^{l-1} \in \mathbb{R}^{T \times d_{l-1}}$ at layer $l-1$ involves incorporating influences from its K-hop neighbors H_n^{l-1} in the context set, weighted by adjacency weights $A_{m,n}^k$:

$$V_m^l = \sum_{k=0}^{K} \frac{V_m^{l-1} + \sum_{n \in \mathcal{N}_k^c(m)} A_{m,n}^k H_n^{l-1}}{1 + \sum_{n \in \mathcal{N}_k^c(m)} A_{m,n}^k} W_k^l, \tag{3.3}$$

where $W_k^l \in \mathbb{R}^{d_{l-1} \times d_l}$ represents learnable weights and $\mathcal{N}_k^c(m)$ denotes the k-hop neighbors of target node m identified from A^k. At the initial layer $l = 0$, V_m^0 denotes the broadcasted target token and H_n^0 the context embeddings. Compared to conventional GCNs, CSGCN not only reduces computational complexity from $\mathcal{O}((N+M)^2)$ to $\mathcal{O}(N \times M)$, but also retains robust learning performance, as verified in our experimental results.

Dilated Causal Convolution Layer We employ dilated causal convolutions (DCConv) Yu and Koltun (2016) to capture temporal dependencies. Unlike the recurrent structure, it learns temporal relations over long sequences by stacking causal layers. This approach proves advantageous as the number of layers is considerably smaller than the length of the sequence, mitigating the issue of transition collapse in the later stage. Specifically, at time t, a 1D causal convolution learns a temporal representation $h^l_{i,t} \in \mathbb{R}^{d_l}$ for node i:

$$h^l_{i,t} = H^{l-1}_i \star \mathcal{K}^l(t) = \sum_{s=0}^{k-1} \mathcal{K}^l(s) \odot H^{l-1}_i(t - \eta \times s), \tag{3.4}$$

where $H^{l-1}_i \in \mathbb{R}^{T \times d_{l-1}}$ is a node representation at the previous layer, $\star\mathcal{K}^l$ means a DCConv with the kernel size $c \times d_{l-1} \times d_l$, and $\odot$ is the Hadamard product. The dilation factor η is initialized to 1 with an exponentially increasing rate of 2 van den Oord et al. (2016) and zero-padding is used to ensure the inputs and outputs have the same time length T.

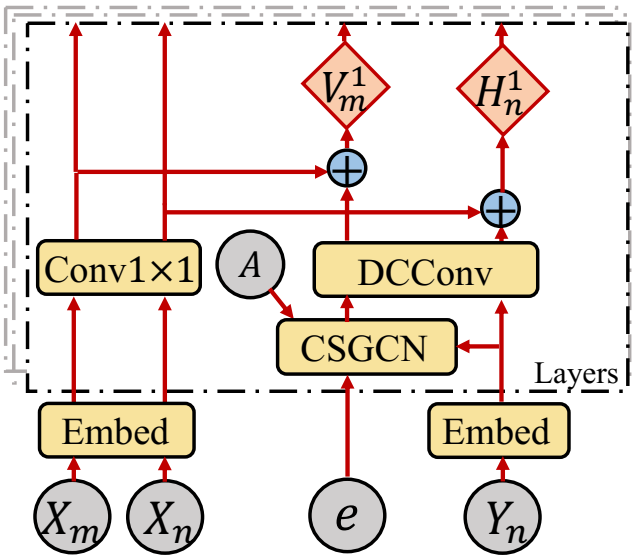

Figure 3.3: The pipeline of the spatio-temporal representation learning network, where we first capture temporal dependencies using the DCConv and then learn spatial relations by CSGCN. Embed denotes an embedding layer.

Learning Framework As illustrated in Figure 3.3, each network layer initiates with a CSGCN to delineate spatial relationships, subsequently integrating a DCConv to address temporal dependencies within node representations. Moreover, covariate features are embedded into node representations using a 1×1 convolution, though these covariates are not directly engaged in CSGCNs and DCConvs due to potential variations in spatio-temporal dynamics or non-existent relationships in certain contexts Tashiro et al. (2021). The architecture employs stacked layers, connected by skip links, to formulate the target node's representations, where each layer preserves temporal relationships at different scales. This setup ensures that upper layers capture broader, long-range interactions while lower layers focus on detailed, granular details. Consequently, the stochastic stage benefits from a hierarchical structure that provides access to multiple scales of dependencies.

3.3.5 Graph Bayesian Aggregation

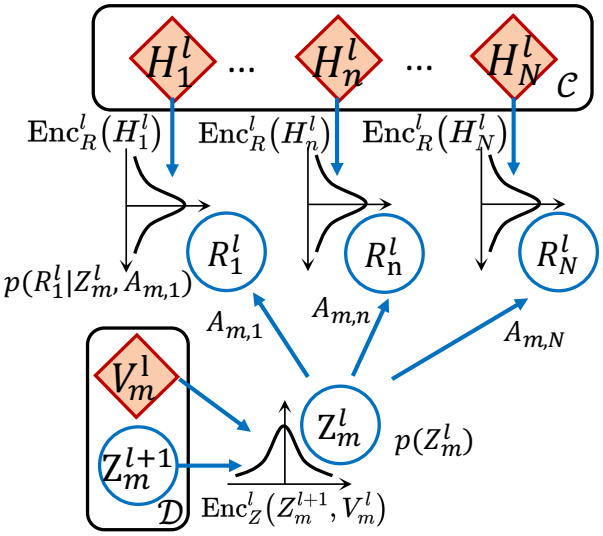

Figure 3.4: Graph Bayesian Aggregation involves two neural networks, $\text{Enc}_Z^l(\cdot)$ and $\text{Enc}_R^l(\cdot)$, which are tasked with learning the mean and variance of the prior and latent observation distributions respectively.

The core component for the stochastic stage is our proposed Graph Bayesian Aggregation, which aggregates information from context nodes and derives latent variables $Z_m^l \in \mathbb{R}^{T \times d_l}$ describing stochastic processes over target nodes. Figure 3.4 illustrates the aggregation process. Based on Bayes' theorem Bishop and Nasrabadi (2006), we assume a prior $p(Z_m^l)$ over the target node. Then for each context node n, a latent observation model $p(R_n^l|Z_m^l, A_{m,n})$ is derived in which its mean conditions on a linear transformation of Z_m and $A_{m,n}$. Thus once observe R_n^l, the latent variable Z^l is updated through its posterior:

$$p(Z_m^l|\{(R_n^l, A_{m,n})\}_{n=1}^N) = \frac{\prod_{n \in \mathcal{N}_1^c(m)} p(R_n^l|Z_m^l, A_{m,n}) p(Z_m^l)}{\prod_{n \in \mathcal{N}_1^c(m)} p(R_n^l)}, \tag{3.5}$$

where we suppose the latent observations are independent and only consider the 1-hop neighbor to simplify the computation. The prior $p(Z_m^l)$ follows a factorized Gaussian:

$$\begin{aligned} p(Z_m^l) &= \mathcal{N}(Z_m^l|\mu_{Z_m^l}, \mathrm{diag}(\sigma^2_{Z_m^l})), \\ (\mu_{Z_m^l}, \sigma_{Z_m^l}) &= \mathrm{Enc}_Z^l(Z_m^{l+1}, V_m^l), \end{aligned} \tag{3.6}$$

where $\mu_{Z_m^l}$ and $\sigma^2_{Z_m^l}$ are mean and variance learned by $\mathrm{Enc}_Z^l(\cdot)$ that will be discussed in the following section. For the latent observation model, we also impose a factorized Gaussian. Note that instead of learning its mean, we learn the observation R_n^l directly together with its variance $\sigma^2_{R_n^l}$, which guarantees valid Gaussian conditioning during inference Volpp et al. (2020):

$$\begin{aligned} p(R_n^l|Z_m^l, A_{m,n}) &= \mathcal{N}(R_n^l|A_{m,n}Z_m^l, \mathrm{diag}(\sigma^2_{R_n^l})), \\ (R_n^l, \sigma_{R_n^l}) &= \mathrm{Enc}_R^l(H_n^l), \end{aligned} \tag{3.7}$$

where R_n and $\sigma^2_{R_n}$ are parameterized by $\mathrm{Enc}_R^l(\cdot)$. The Gaussian assumption avoids an intractable computation of the marginal likelihood of the posterior's denominator. In fact, we can calculate it easily by Gaussian conditioning in a closed-form solution:

$$\bar{\sigma}^2_{Z_m^l} = \left[\left(\sigma_{Z_m^l}\right)^{-2} + \sum_{n \in \mathcal{N}_1^c(m)} \left(\sigma_{R_n^l}/A_{m,n}\right)^{-2}\right]^{-1}, \tag{3.8}$$

$$\bar{\mu}_{Z_m^l} = \bar{\sigma}^2_{Z_m^l}\left(\mu_{Z_m^l}/\sigma^2_{Z_m^l} + \sum_{n \in \mathcal{N}_1^c(m)} A_{m,n}R_n/\sigma^2_{R_n^l}\right), \tag{3.9}$$

where $\bar{\sigma}^2_{Z^l_m}$ and $\bar{\mu}_{Z^l_m}$ are updated parameters and the operations are conducted in an element-wise manner. With factorization, the conditioning is efficient to compute, avoiding costly matrix inversion. In addition, all the calculations are differentiable so that GBA can be optimized in an end-to-end way by stochastic gradient descent.

The aggregation mechanism has profound implications. Primarily, it integrates the graph structure by employing a linear transformation via the adjacency matrix, aligning it functionally with GCNs, albeit without considering uncertainty terms. This alignment suggests that GBA possesses learning capabilities comparable to those of GCNs. Additionally, the aggregation accounts for node uncertainties, enhancing the model's effectiveness compared to previous methods. Analytically, the influence of a context node is governed by its observed value R^l_n, the variance $\sigma_{R^l_n}$, and the weight assigned by the adjacency matrix $A_{m,n}$. Equation 3.8 posits that a context node's contribution to the target is inversely proportional to its distance, implying decreased confidence from distant nodes. Moreover, Equation 3.9 indicates that a node's contribution is reduced when its variance $\sigma_{R^l_n}$ is high, reflecting increased uncertainty. This feature theoretically enhances the model's resilience against noisy data. Furthermore, the independence of latent observations assumed in Equation 3.5 ensures a robust posterior, independent of the context node count, reinforcing GBA's inherent inductive capabilities without the need for external sampling techniques Wu et al. (2021); Hamilton et al. (2017).

3.3.6 Generative Process

The target latent variable Z^l_m depends on its representation V^l_m and those of the context nodes H^l. The longer-range temporal dependencies are transited by conditioning Z^l_m on Z^{l+1}_m, forming a vertical time hierarchy. In practice, given V^l_m and a sample from $p(Z^{l+1}_m)$, the network $\text{Enc}^l_Z(Z^{l+1}_m, V^l_m)$ first learns a prior $p(Z^l_m)$ over the target node in Equation 3.6. Then, the deterministic representations of context nodes are adopted

to learn their latent observations by $\text{Enc}_R^l(H_n^l)$ in Equation 3.7. Next, parameters of $p(Z_m^l)$ are updated according to Equation 3.8 and 3.9. After the bottom layer $l = 1$, a likelihood model concatenates samples $Z_m = (Z_m^1, ...Z_m^L)$ from all layers and the target node's exogenous covariates X_m to predict its extrapolations Y_m. Formally, the generative process of STGNP is summarized as:

$$p(Y_m, Z_m | X_m, \mathscr{C}, A) = p(Y_m | X_m, Z_m) \prod_{l=1}^{L} p(Z_m^l | Z_m^{l+1}, V_m^l, H^l, A), \tag{3.10}$$

where the first term is a likelihood; the second term denotes a conditional prior aggregated through the GBA. Note that at the top layer L, $Z_m^{L+1} = \mathbf{0}$ and the likelihood is assumed to be a factorized Gaussian distribution.

3.3.7 Inference and Optimization

Typically, closed-form solutions for non-linear transitions and likelihood do not exist; thus we train the model through variational approximation. The approximated posterior $q(Z_m | \mathscr{C} \cup \mathscr{D}, A)$ has the same structure as the conditional prior but takes target node data Y_m as inputs. Then the deterministic and stochastic modules can be optimized together by the ELBO:

$$\begin{aligned} \log p(Y_m | X_m, \mathscr{C}, A) \geq \mathbb{E}_{q(Z_m)}[\log p(Y_m | X_m, Z_m)] \\ - \mathbb{KL}(q(Z_m | \mathscr{C} \cup \mathscr{D}, A) || p(Z_m | X_m, \mathscr{C}, A)). \end{aligned} \tag{3.11}$$

Given the hierarchical structure of Equation 3.10, the Kullback-Leibler divergence term $\mathbb{KL}$ can be further decomposed as:

$$\mathbb{KL}(\cdot||\cdot) = \sum_{l=1}^{L} \mathbb{E}_{q(Z_m^{l+1})}\left[\mathbb{KL}(q(Z_m^l | Z_m^{l+1}, V_m'^{\,l}, H^l, A) || p(Z_m^l | Z_m^{l+1}, V_m^l, H^l, A))\right] \tag{3.12}$$

where unlike using the learned token, $V_m'^{\,0}$ is the feature embeddings of Y_m. Following Garnelo et al. (2018), we use the same variational module to approximate the conditional prior so that $p(\cdot) = q(\cdot)$ in Equation 3.12 During optimization, ELBO can be

minimized using stochastic gradient descent with the reparameterization trick Kingma and Welling (2014).

3.4 Datasets and Evaluations

3.4.1 Datasets

Halpe-FullBody Similar to Alphapose Fang et al. (2022), we used the Halpe-FullBody dataset. However, unlike them, we do not focus on the accuracy of the face and hands here because, through observation, the accuracy of the face and hands is very low during dynamic movements, far from being practically usable. Therefore, we do not consider this situation for now. For each person, they annotated 136 keypoints, including 20 for body, 6 for feet. The keypoint format is illustrated in Fig. 3.5.

COCO-WholeBody Concurrently, Jin et al. have annotated 133 whole body keypoints using the COCO framework Jin et al. (2020). This dataset's definition of keypoints is largely in line with the *Halpe-FullBody* dataset, but it lacks annotations for the head, neck, and hip. The training dataset encompasses 118K images featuring 250K instances, while the test set includes 5K images. Our algorithm has also been tested using this dataset.

3.4.2 Experimental Setup

3.4.2.1 Baselines.

We considered three baseline models, namely *Openpose*, *HRNet*, and *FastPose50(Alphapose)*.

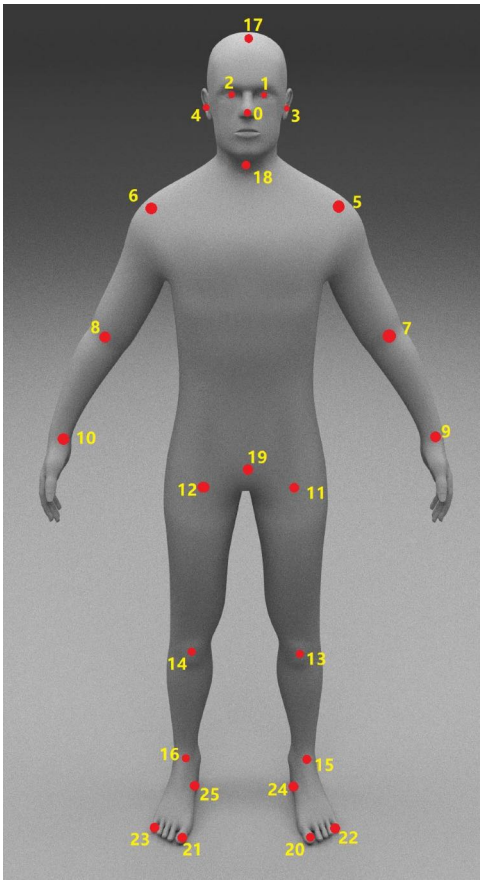

Figure 3.5: Keypoint configuration for body and foot in the Halpe-FullBody dataset.

3.4.3 Overall Performance

Here, in order to simulate the drift of skeletal key points in a real scenario, we set the confidence threshold to 0.3. Points below this value are considered missing key points. The results are shown in Tables 3.1 and 3.2. It can be seen that after integrating our model, the accuracy of most cases has improved, proving the effectiveness of our method.

Method	Input Size	full-body						foot		body	
		AP	AP^{50}	AP^{75}	AP^{L}	AP^{M}	AR	AP	AR	AP	AR
OpenPose Cao et al. (2019)	N/A	0.281	0.569	0.233	0.331	0.297	0.343	0.409	0.644	0.542	0.589
OpenPose-STGNP	N/A	**0.414**	**0.674**	**0.396**	**0.425**	**0.319**	**0.370**	**0.529**	**0.685**	**0.603**	**0.633**
HRNet Sun et al. (2019)	256×192	0.391	0.803	0.362	0.389	0.447	0.559	0.595	0.747	0.610	0.721
HRNet-STGNP	256×192	**0.460**	**0.834**	**0.411**	**0.464**	**0.497**	**0.592**	**0.671**	**0.749**	**0.634**	**0.743**
FastPose50 Fang et al. (2022)	256×192	0.454	0.794	0.466	0.482	0.498	0.544	0.719	0.787	0.649	0.711
FastPose50-STGNP	256×192	**0.519**	**0.854**	**0.555**	**0.521**	**0.548**	**0.605**	**0.776**	**0.823**	**0.682**	**0.862**

Table 3.1: Human pose estimation results on Halpe-FullBody dataset. Results are obtained using single-scale testing for fair comparisons Fang et al. (2022). "STGNP" represents our model. Here we define key points with a confidence level less than 0.3 as missing points.

Method	Input Size	whole-body		body		foot	
		AP	AR	AP	AR	AP	AR
OpenPose Cao et al. (2019)	N/A	0.338	0.449	0.563	0.612	0.532	0.645
OpenPose-STGNP	N/A	**0.429**	**0.498**	**0.644**	**0.689**	**0.613**	**0.726**
HRNet Sun et al. (2019)	256×192	0.432	0.520	0.659	0.709	**0.314**	0.424
HRNet-STGNP	256×192	**0.521**	**0.595**	**0.668**	**0.749**	0.302	**0.449**
FastPose50 Fang et al. (2022)	256×192	0.554	0.625	0.673	0.717	0.636	0.718
FastPose50-STGNP	256×192	**0.602**	**0.714**	**0.766**	**0.817**	**0.682**	**0.812**

Table 3.2: Human pose estimation results on COCO-WholeBody dataset. Results are obtained using single-scale testing for fair comparisons Fang et al. (2022). "STGNP" represents our model. Here we define key points with a confidence level less than 0.3 as missing points.

3.5 Conclusion

We present the STGNP, marking the inaugural application of spatio-temporal extrapolation in the Neural Processes family. This model adeptly handles temporal relationships and mitigates transition collapse through the use of causal convolutions, while also proficiently learning spatial dependencies with a cross-set graph network.

We applied our method to enhance human skeleton prediction results, demonstrating its effectiveness in improving the accuracy and reliability of these predictions. The Graph Bayesian Aggregation mechanism aggregates context nodes by considering their uncertainties, thereby enhancing the learning capability of Neural Processes on graph data.

In future work, we plan to further refine the STGNP model and explore its applications in other domains requiring spatio-temporal data analysis. Additionally, we aim to investigate more sophisticated aggregation techniques and extend the model's capabilities to handle even more complex and dynamic scenarios.

CHAPTER 4

APPLYING SPATIO-TEMPORAL TRANSFORMERS TO BASKETBALL TRACKING IN SPORTS EXAMINATIONS

To address RQ2, this chapter proposes a comprehensive framework for integrating positional information into Re-ID algorithms. Section 4.1 details the motivations and challenges of enhancing object tracking amidst frequent occlusions and complex interactions. In Section 4.2, the definitions, notations, and the proposed integration method are listed and explained. Section 4.3 introduces the basketball exam dataset we collected and organized. Finally, Section 4.4 validates the proposed method through experimental evaluations.

4.1 Introduction

In recent years, the field of object tracking has seen substantial advancements, particularly in the context of automated systems and intelligent sports examinations. Traditional tracking algorithms such as ByteTrack and BoT-SORT Zhang et al. (2022); Aharon et al. (2022) have been widely applied across various domains, including sports analytics. However, these methods often fall short in real-world sports examination scenarios where

occlusions and multiple identical objects, such as basketballs, are prevalent. This results in significant challenges such as target loss and ID confusion Zhang et al. (2020); Ivasic-Kos et al. (2021). To address these challenges, researchers have been exploring novel approaches that integrate advanced Re-ID techniques and enhanced tracking methodologies. These approaches aim to improve the robustness and accuracy of object tracking under complex conditions Zhang et al. (2012). Additionally, the incorporation of machine learning models, particularly those leveraging deep learning, has shown promise in overcoming the limitations of traditional methods Morimitsu et al. (2017). By continuously adapting to changes in object appearance and leveraging contextual information, these advanced models are better equipped to handle the dynamic and unpredictable nature of sports environments Ivasic-Kos et al. (2021). As a result, the development of more sophisticated tracking systems is crucial for enhancing the performance and reliability of intelligent sports examinations.

Advanced tracking methods aim to tackle the challenges of object tracking in complex sports environments by continuously adapting to dynamic conditions and occlusions. These methods involve discarding outdated models and training new ones with updated data to maintain accuracy in object identification and tracking. In this way, the tracking system can always fit the latest data distribution, ensuring precise performance. Recent studies have introduced state-of-the-art ideas, such as integrating automated machine learning (AutoML) techniques to optimize tracking pipelines and adjust them in real-time as conditions change Jiang and Zhang (2021). Online incremental learning-based methods are also employed to accumulate knowledge and enhance the system's ability to adapt to evolving sports scenarios Cheng et al. (2015); Cao et al. (2024). Moreover, lifelong learning approaches are developed to expand the system's knowledge base, making it adaptable to new and unforeseen situations. For example, advanced tracking algorithms like the streaming decision tree utilize innovative techniques to address

occlusions and maintain accurate tracking without knowledge forgetting Cheng et al. (2015). These algorithms leverage continuous learning and adjustment mechanisms, ensuring that the system remains responsive to new data and changing conditions Huang et al. (2024). Additionally, the integration of Re-ID technologies has significantly enhanced the ability to distinguish between multiple identical objects, reducing ID confusion and improving tracking reliability Xalabarder (2021). The incorporation of deep learning models, particularly CNNs and RNNs, has further improved the system's performance by enabling more sophisticated feature extraction and temporal modeling Jiang and Zhang (2021); Cao et al. (2024). These advancements demonstrate that modern tracking methods can effectively handle the dynamic and challenging environments of sports examinations, providing reliable and robust performance Xuan and Xu (2022).

However, real-world sports examinations, such as basketball skill assessments, often involve complex scenarios where tracking targets may experience occlusions or overlap with identical objects, leading to issues in target tracking or Re-ID. These challenges significantly impact subsequent computational processes, such as performance evaluation and data analysis. For instance, during a basketball examination, multiple players and identical basketballs may occlude each other, causing difficulties in maintaining accurate tracking and identification of individual players and balls. Traditional tracking methods, which are typically designed to handle single-stream tasks, struggle to manage these multi-object and occlusion-rich environments effectively Zhang et al. (2022); Aharon et al. (2022). Furthermore, the need to differentiate between multiple identical objects, such as basketballs used in the same exam, adds an additional layer of complexity to the tracking system.

Advanced tracking systems must, therefore, be capable of handling both labeled and unlabeled data streams simultaneously, adapting to the dynamic and unpredictable nature of sports environments Singh and Srivastava (2022). By leveraging sophisticated

Re-ID techniques and deep learning models, these systems can improve their robustness against occlusions and overlapping objects, ensuring more reliable and accurate tracking Ko et al. (2021). Additionally, hybrid models that combine different tracking and Re-ID approaches offer a promising solution to address these challenges, enhancing the overall performance and efficiency of intelligent sports examinations Paik and Kim (2022). The continuous adaptation and learning capabilities of these advanced systems are crucial for managing the complexities of real-world sports scenarios, providing reliable data for subsequent analysis and decision-making Hsu et al. (2019).

We introduce a sophisticated tracking framework that integrates examinee data into the target tracking process, thereby enhancing the system's ability to preserve consistent IDs through occlusions or when objects overlap. Drawing inspiration from the Learning Spatio-Temporal Transformer for Visual Tracking, our methodology employs an encoder-decoder transformer structure at its core. The encoder captures the global spatio-temporal feature interactions between target objects and their search areas, while the decoder develops a query embedding to accurately locate the target objects in space. Departing from conventional methods that depend on either proposals or predefined anchors, our approach redefines object tracking by directly predicting bounding boxes, utilizing a straightforward fully-convolutional network to directly determine the positions of object corners Yan et al. (2021). The inclusion of examinee information not only enriches the contextual data but also bolsters the model's ability to reliably maintain target identities under difficult circumstances.

This approach not only enhances the robustness of the tracking system but also ensures more reliable and stable target recognition results for subsequent processes. The encoder-decoder transformer allows our method to operate end-to-end without requiring post-processing steps such as cosine window or bounding box smoothing, thereby simplifying existing tracking pipelines. Our improved tracker effectively addresses the challenges

faced by other methods in real-world examination scenarios by uniquely identifying and tracking each player and object. This ensures that the equipment used by examinees, in this case, basketballs, is consistently tracked, providing stable recognition results for subsequent processes. This capability is particularly crucial in basketball examinations where players and balls frequently occlude each other, and multiple identical balls are used simultaneously.

Our system's ability to maintain accurate and stable IDs under these conditions significantly improves the reliability of the data collected for performance evaluation and analysis. Moreover, this robust tracking capability supports advanced analytics, such as player movement patterns and ball handling efficiency, providing deeper insights into examinee performance. The continuous adaptation and learning capabilities of our system are crucial for managing the dynamic and unpredictable nature of sports environments, ensuring reliable data for performance evaluation and analysis. The ability to adapt to new data and evolving conditions in real-time not only enhances the system's performance but also its applicability to various sports and examination scenarios, making it a versatile tool for intelligent sports assessments.

4.2 Proposed Method

In this section, we formalize an enhanced spatio-temporal visual tracking framework that incorporates contextual dynamic information and describe our proposed model in detail. In a real basketball test, the appearance presented by the examinee as the detection target may change significantly due to movement changes and object occlusion, etc. Therefore, in order to achieve accurate target tracking, the model must dynamically capture the examinee's condition in real time. However, existing traditional baseline methods tend to use only the first and current frames as inputs to the model, and this input mode only considers spatial information and almost completely ignores the rich

temporal correlation between multiple frames, and thus it tends to suffer from the problems of target loss and target confusion. In view of the aforementioned inherent flaws prevalent in target tracking models, we thoroughly redesign the proposed enhanced spatio-temporal visual tracking framework. In terms of model inputs, we not only employ an initial spatial template but also introduce a dynamically updated template sampled from intermediate frames, which contains important temporal information about the primary target-the basketball-as it changes over time, as well as a contextual template that incorporates information related to the examinee's posture. This enables the model to include dynamic information from both spatial and temporal dimensions, maintaining tracking integrity even under challenging conditions. Architecturally, our model employs a hybrid approach combining ResNet-50 and Transformer components to fully extract spatio-temporal features from the inputs. In addition, there may be cases where the examinee is occluded or out of the field of view during the examination, so the image of a specific frame may not be able to be input into the model as a dynamic template. To address this situation, we also add two score prediction heads consisting of a three-layer perceptron to determine whether the current frame is reliable as a dynamic/context template. To ensure that the model performs optimally on both localization and classification subtasks, we adopt a two-stage training scheme to improve the model's performance on the two different tasks by decoupling localization and classification and performing targeted end-to-end training separately.

In the experimental part, we collect and organize a dataset containing test videos of nine basketball test events and compare our proposed model with several different mainstream target tracking methods in specific application scenarios, and the experimental results show that the proposed enhanced spatio-temporal visual tracking framework significantly outperforms similar algorithms in several metrics, demonstrating a stunning superiority. In addition, we also visualize the tracking effect of the model in real-world

application scenarios, which fully demonstrates that it can effectively handle complex tracking scenarios with higher accuracy and robustness.

4.2.1 Spatio-Temporal Transformer Tracking with Examinee Info

In practical sports education examinations, examinees' appearances may evolve over time, leading to complex scenarios involving obstructions, target loss, or the detection of multiple identical targets. Consequently, dynamic detection and tracking of targets become crucial. In the enhanced spatio-temporal visual tracking framework we propose, we not only harness information from both spatial and temporal dimensions, but also integrate contextual information about examinees' postures to address issues of ID inconsistency, thereby achieving more stable tracking. The proposed framework makes significant innovations in three aspects: network input diversification, introduction of an additional score prediction head to compute the reliability of the dynamic templates, and adoption of a two-stage training strategy that decouples localization and classification, which are described point by point in the following subsections. Fig. 4 illustrates the framework diagram of the proposed enhanced spatio-temporal visual tracking model.

Overall Architecture. Demonstrated in Fig 4.1, our proposed enhanced spatio-temporal visual tracking framework combines advantages from *STARK* and *MixFormer*. It mainly consists of four components: a convolutional backbone based on ResNet-50, a transformer encoder and decoder, and finally a bounding box prediction head, and a score head.

Input. Target tracking models widely utilized today typically extract features from the spatial information within a single frame image. However, in complex application scenarios such as basketball tests, the appearance of a tracking target does not remain constant throughout the entire motion cycle. Furthermore, obstructions from surrounding

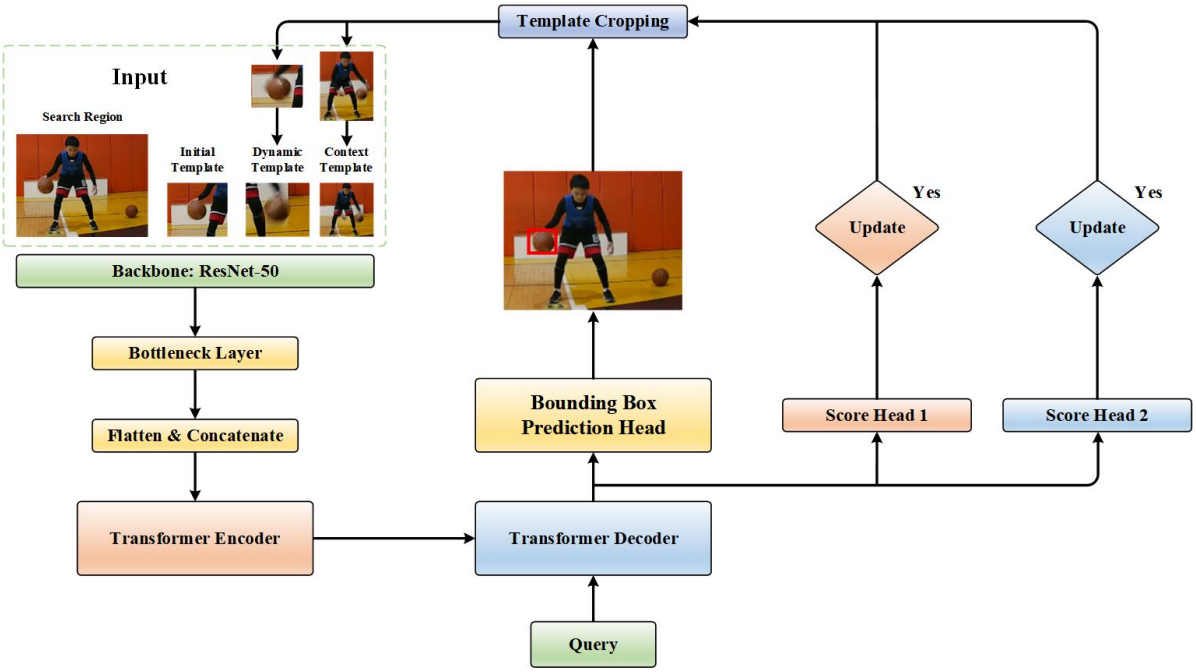

Figure 4.1: Framework for our network

objects or temporary exits from the field of view during motion can severely limit the performance of these models that only consider a single spatial dimension. In fact, the multiple frames involved in the tracking process also have a strong temporal correlation. Therefore, using diversified inputs that include both temporal and spatial dimensional information for dynamic identification of tracking targets can significantly enhance model robustness in complex environments. Inspired by the reference method (STARK), our proposed enhanced spatio-temporal visual tracking framework employs three types of templates at the input stage as illustrated in Figure 4.1: the traditional spatial template, a dynamic template that updates with the basketball target as it changes over time, and a context template that includes information related to the examinee's posture. The adoption of these three templates provides the model with an extensive array of multi-dimensional spatio-temporal information, facilitating the simultaneous modeling and capture of highly abstract global relationships between multiple levels of features. This enables continuous tracking and target locking even in noisy environments or in

the presence of obstructions.

Our Based Backbone. The backbone employed in our proposed framework is similar to that used in STARK, where any network suitable for image-related tasks, such as convolutional networks or ViTs, can serve as the primary architecture for initial feature extraction. This flexibility allows for the integration of mainstream models that are adept at handling various aspects of visual data processing, enhancing the framework's adaptability and performance in feature extraction from complex visual inputs. Considering that in real world, the compute resource is limited, so here, we adopt the vanilla ResNet He et al. (2016) as the backbone due to its proven effectiveness in similar tasks, ease of deployment, and structural extensibility. Concretely, we remove the last stage and fully-connected layers from the original ResNet He et al. (2016), making no other changes. The input to the backbone includes a pair of images: a template image of the initial target object $z \in \mathbb{R}^{3\times H_z\times W_z}$ and a search region of the current frame $x \in \mathbb{R}^{3\times H_x\times W_x}$. After passing through the backbone, the template z and the search image x are mapped to two feature maps $f_z \in \mathbb{R}^{\frac{C}{s^2}\times H_z\times W_z}$ and $f_x \in \mathbb{R}^{\frac{C}{s^2}\times H_x\times W_x}$.

Our approach diverges from this method by incorporating two additional dynamically updated images: a dynamically updated template and a dynamically updated associated region. In the context of our basketball skill assessment system, the associated region corresponds to the examinee's region. This addition allows our system to adapt more effectively to the dynamic nature of the examination environment.

Encoder & Decoder. After initial feature extraction through the Backbone ResNet-50, the resulting feature maps are passed through a bottleneck layer to reduce the dimensionality in the channel dimension, compressing the original C feature channels into d channels. Subsequently, the compressed feature maps are fed into an Encoder and Decoder composed of Transformers. To accommodate the sinusoidal positional encoding required by the Transformer input, the feature maps are then flattened and

concatenated along the spatial dimension, producing a feature sequence with a length of $\frac{H_z}{s} \times \frac{W_z}{s} + \frac{H_x}{s} \times \frac{W_x}{s} + \frac{H_{dt}}{s} \times \frac{W_{dt}}{s} + \frac{H_{dr}}{s} \times \frac{W_{dr}}{s}$ and a dimension of d, where H_{dt} and W_{dt} represent the height and width of the dynamically updated template, and H_{dr} and W_{dr} represent the height and width of the dynamically updated associated region. The encoder consists of stacked multi-head attention layers, with each layer capturing the hidden relationships between any two elements in the input sequence-post positional embedding-from a global perspective and in parallel. This structure allows the encoder to thoroughly learn the dependencies of features across both temporal and spatial dimensions, thereby enabling more precise localization and tracking of targets.

The decoder concurrently receives target queries and the output sequences passed from the encoder, and based on this, it generates the tracking target object-specifically, the bounding box of the basketball held by the examinee. The structure of the decoder is essentially identical to that of the encoder, consisting of layers stacked with multi-head attention mechanisms. However, the decoder employs masked multi-head attention, which effectively screens out future data to prevent information leakage. By further recognizing the spatio-temporal feature sequences learned by the encoder, the decoder can robustly and accurately generate the final tracking bounding box. This process ensures precise alignment and prediction in the context of visual tracking tasks.

Head. During target tracking predictions, the dynamic and context templates provide the model with rich temporal information, yet these templates are not always available in every frame. Complex scenarios, such as when the target briefly moves out of view, is disrupted by similar targets, or becomes occluded, may render the current dynamic and contextu templates unavailable for the model input. Therefore, in our proposed enhanced spatio-temporal visual tracking framework, we incorporate two score prediction heads that assess the reliability of the current dynamic and context templates. These mechanisms ensure the model's robustness by dynamically adjusting the input based

on the contextual and temporal relevance and reliability of the information provided by these templates. Inspired by the prediction head designs in both STARK Yan et al. (2021) and MixFormer Cui et al. (2022), our method aims to integrate the strengths of both approaches. Like STARK, we use a probability distribution to estimate the box corners, thereby improving the robustness and accuracy of bounding box predictions. However, similar to MixFormer, we simplify the design by adopting a fully-convolutional corner-based localization head.

Initially, the search region features are derived from the output sequence of the encoder, and their similarity to the decoder's output embedding is calculated. These similarity scores are then multiplied element-wise with the search region features to highlight significant areas while diminishing the less distinctive ones. The resultant feature sequence is transformed into a feature map $f \in \mathbb{R}^{\frac{d}{s^2} \times H_s \times W_s}$, which is input into a streamlined fully-convolutional network (FCN). This FCN, composed of multiple Conv-BN-ReLU layers, produces two probability maps, $P_{tl}(x, y)$ for the top-left and $P_{br}(x, y)$ for the bottom-right corners of the bounding boxes.

Subsequently, the bounding box coordinates $(\hat{x}_{tl}, \hat{y}_{tl})$ and $(\hat{x}_{br}, \hat{y}_{br})$ are determined by calculating the expectation values of the corners' probability distributions. Our approach enhances accuracy and robustness in object tracking by precisely modeling the uncertainty in the coordinate estimates. This technique maintains the resilience seen in STARK's methodology while embracing the simplicity and efficiency inherent in MixFormer's fully-convolutional architecture.

Training and Inference. In our proposed target tracking framework, the detection process necessitates completing both localization and classification steps. However, traditional training methods optimize these subtasks simultaneously, leading to a trained model that compromises between the two to achieve balance, thereby failing to reach optimal performance in either localization or classification. Therefore, we employ a two-

stage training method that decouples these subtasks. The first stage focuses on training the model's localization capabilities, with the loss function defined as follows:

$$L = \lambda_{iou} L_{iou}(b_i, \hat{b}_i) + \lambda_{L_1} L_1(b_i, \hat{b}_i). \tag{4.1}$$

where b_i represents the label and $\hat{b}_i$ is the predicted bounding box. λ_{iou} and L_{iou} are predefined hyperparameters. This loss function primarily targets losses related to the localization phase and does not optimize the prediction score heads used for classification. After the completion of the first stage, all parameters unrelated to classification within the model are fixed, and only the score heads are optimized. The loss function used for this phase is accordingly modified as follows:

$$L_{ce} = y_i log(P_i) + (1 - y_i) log(1 - P_i) \tag{4.2}$$

where y_i is the label and P_i represents the the confidence. Through the novel two-stage training process, our proposed tracking framework is able to achieve optimal capabilities in both localization and classification. During the inference process, the initial template, dynamic template, and context template, along with their corresponding features, are initialized at the beginning. As illustrated in Figure 4.1, the search region is appropriately cropped and fed into the model, which then generates the corresponding bounding box and confidence score. The presence of confidence scores ensures that the model does not update the dynamic and context templates in scenarios where the tracking target is occluded or under other non-ideal conditions. This mechanism guarantees the model's precision and stability during the inference stage.

4.3 Dataset

To better address the issue, we collected a real youth backetball level examination dataset from China. This dataset contains 810 videos from 30 youth testers aged between 8-14

years old, containing 9 testing items: Triangle slide defense, dribbling layup, five-point spot shooting, passing and catching the ball, front and back spin dribble, stationary in-front dribble, stationary two-hand dribble, stationary behind-the-back dribble, stationary between-the-legs dribble. There are different examination tools used in these actions, such as cones, basket and basketball used in five-point spot shooting, different action requires different tools.

4.3.1 Actions Definition

- **Triangle Slide Defense.** As shown in 4.2, in this action, the examinee should move between the cones. From left to right are cone 2, cone 0, cone 1. The examinee starts at cone 0 and completes the preparation action. Then, they slide to cone 1 and touch it, followed by sliding to cone 2 and touching it. Next, they slide back to cone 0 and touch it, then slide to cone 2 and touch it. After that, they slide to cone 1 and touch it, and finally slide back to cone 0 and touch it. The sequence of touches should be: 0-1-2-0-2-1-0. In this action, the props used are cones.

Figure 4.2: Triangle Slide Defense.

- **Dribbling Layup.** As shown in 4.3. The examinee starts dribbling from the preparation spot and dribbles around the cones from the outside, proceeding to

make a layup.

When dribbling around the nearby cone: dribble with the right hand, and the takeoff foot for the layup should be the left foot. At the moment of release, the right knee should be higher than the left knee, and the right hand should be higher than the left hand. If the layup misses, do not dribble around the cone again; perform a follow-up shot.

When dribbling around the far cone: dribble with the left hand, and the takeoff foot for the layup should be the right foot. At the moment of release, the left knee should be higher than the right knee, and the left hand should be higher than the right hand. If the layup misses, do not dribble around the cone again; perform a follow-up shot. In this action, the props used include cones, baskets, and basketballs.

Figure 4.3: Dribbling Layup.

- **Five-point Spot Shooting.** As shown in 4.4, in this action, the examinee should shoot twice at each of the five positions. In this action, the props used include cones, baskets, and basketballs. In this item, it is necessary not only to detect the number of hits and misses but also to judge whether the action during shooting is standard and whether the position during shooting is correct.

Figure 4.4: Five-point Spot Shooting.

- **Passing and Catching the Ball.** As shown in 4.5. The examinee is positioned on the left. A completed air pass is defined as: one air pass from the right to the left, followed by a return pass from the left to the right, constituting one air pass. A completed bounce pass is defined as: one bounce pass from the right to the left, followed by a return pass from the left to the right, constituting one bounce pass. Perform 2 air passes forward and 2 bounce passes backward. In this action, the test equipment used is a basketball.

Figure 4.5: Passing and Catching the Ball.

- **Front and Back Spin Dribble.** As shown in 4.6. For a right turn dribble: dribble with the right hand, pivot on the left foot, and turn clockwise to the right. For

a left turn dribble: dribble with the left hand, pivot on the right foot, and turn counterclockwise to the left. In this exam, the prop that appeared was a basketball.

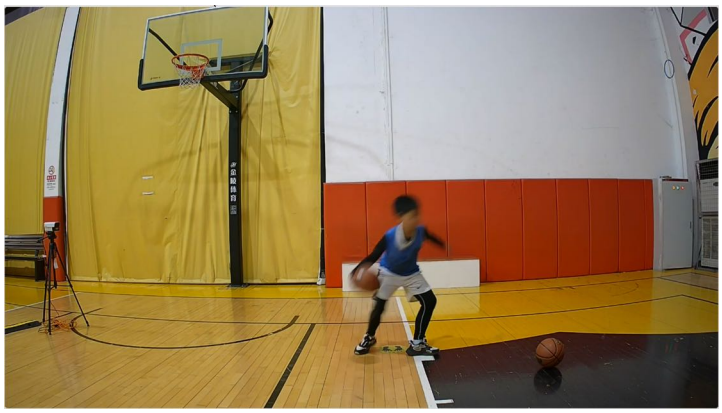

Figure 4.6: Front and Back Spin Dribble.

- **Stationary In-front Dribble.** As shown in 4.7. The examinee holds the ball with one hand and completes the preparation action. Then, they perform 10 front dribbles. A valid dribble is defined as dribbling the ball from one hand to the other and back again. When transferring the ball between hands, it should bounce on the ground only once. In this exam project, the prop used is a basketball.

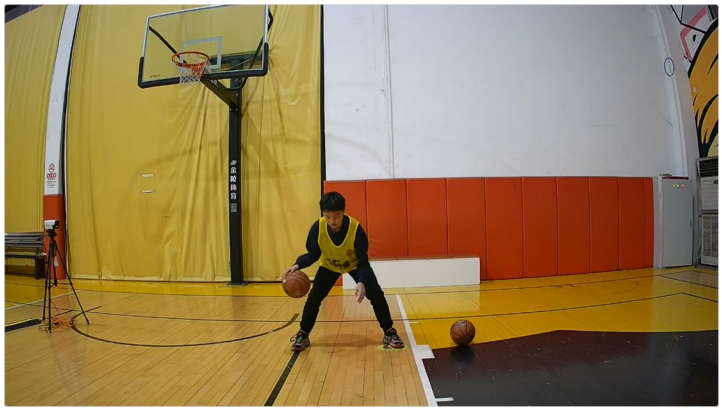

Figure 4.7: Stationary In-front Dribble.

- **Stationary Two-hand Dribble.** As shown in 4.8. The examinee holds the ball with both hands and completes the preparation action. Then, they perform 10

simultaneous dribbles and 10 alternating dribbles. For simultaneous dribbles, both balls must hit the ground at the same time. For alternating dribbles, the two basketballs should hit the ground alternately. In this exam project, the prop used is a basketball.

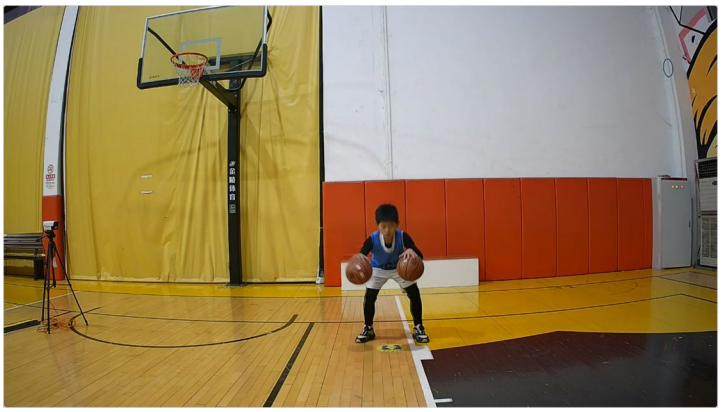

Figure 4.8: Stationary Two-hand Dribble.

- **Stationary Behind-the-back Dribble.** As shown in 4.9. The examinee holds the ball with one hand and completes the preparation action. Then, they perform 10 behind-the-back dribbles. A valid behind-the-back dribble is defined as dribbling the ball at least once beside the body, then dribbling it behind the back to the other side, where it is caught by the opposite hand. The equipment used is a basketball.

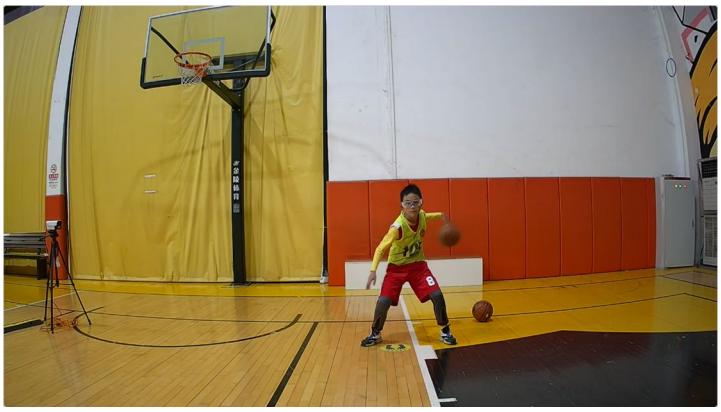

Figure 4.9: Stationary Behind-the-back Dribble.

- **Stationary Between-the-legs Dribble.** As shown in 4.10. The examinee holds the ball with one hand and completes the preparation action. Then, they perform 10 behind-the-back dribbles. A valid between-the-legs dribble is defined as dribbling the ball at least once beside the body, then dribbling it between the legs, where it is caught by the opposite hand. The equipment used in this examination is a basketball.

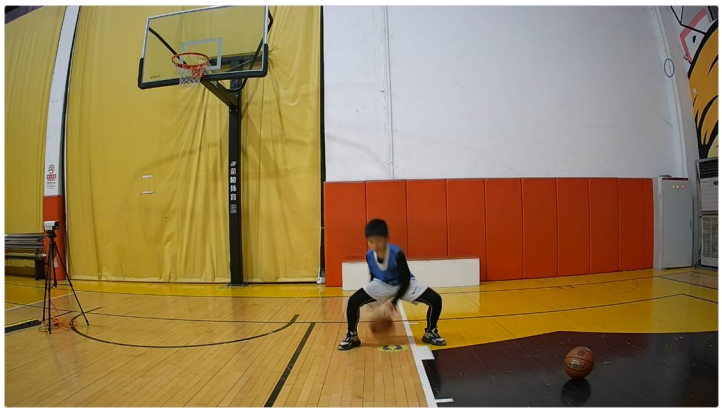

Figure 4.10: Stationary Between-the-legs Dribble.

In such real-world scenarios, there are many target tracking situations that include occlusion, overlapping of the same targets, lighting conditions, and more. We first trained our own YOLOv8 tracking model on nearly 80,000 images, including target recognition for basketballs, hoops, and cones. Then we applied this model to the dataset and finally manually corrected the recognition and tracking results for each video. This provided a highly challenging target tracking recognition dataset.

4.4 Experiment

To evaluate our proposed method, we conduct comprehensive experiments to show the performance of our model when dealing with the real examination sceneries. For our model, the first consideration is the actual application scenario. Therefore, when

choosing comparison methods, we need to consider memory usage and inference speed. Hence, we carefully select 8 mainstream baseline methods to compare with our proposed model. These methods encompass both deep learning and traditional machine learning approaches. Experiments show that our target recognition method can not only accurately identify the correct targets in this specific scenario but also provide better data for the re-id model, making re-id more stable.

This section first introduces the implementation details of the model and specific parameters. Then, it presents the evaluation metrics used in this part of the experiment and displays the performance of other baseline methods, comparing them with our method to demonstrate the superiority of the proposed method. Finally, the ablation studies are presented to provide a detailed explanation of how the key components of the proposed network impact the model's performance.

4.4.1 Implenmentation Details

We implement our trackers with Python 3.9 and PyTorch 2.2.0, and run experiments on a server equipped with two 48GB Nvidia A40 GPUs.

Model. For our experiments, we utilize ResNet-50 He et al. (2016) as the backbone, initialized with parameters pre-trained on ImageNet. Throughout training, the Batch-Norm Ioffe and Szegedy (2015) layers remain static. We extract backbone features from the fourth stage at a stride of 16. Our transformer's design mirrors that of DETR Carion et al. (2020), featuring 6 encoder and 6 decoder layers, which include multi-head attention layers (MHA) and feed-forward networks (FFN). The MHAs are configured with 8 heads and a width of 256, while the FFNs possess 2048 hidden units. A dropout ratio of 0.1 is applied. The bounding box prediction head, influenced by MixFormer, utilizes a fully convolutional corner-based localization approach. It directly computes the bounding box of the tracked object by applying multiple Conv-BN-ReLU layers to predict the

top-left and bottom-right corners. The final bounding box coordinates are determined by calculating the expectation of the corner probability distributions. Unlike STARK, which relies extensively on both the encoder and decoder within a more intricate framework, our model simplifies the approach with a fully convolutional head.

Training. The training data comes from our collected dataset, which includes 10,000 frames sampled from 500 videos. In these frames, we only labeled the position of the basketball. This focus is due to the high likelihood of the basketball being occluded, overlapped, or repeated in exam scenarios.

Inference. By default, the dynamic template update interval Tu is set at 10, and the confidence threshold tau at 0.5. The inference process involves merely a forward pass followed by a coordinate transformation that maps the search region coordinates back to the original image coordinates, omitting any additional post-processing steps.

4.4.2 Evaluation Metrics

In this study, we employ three key evaluation metrics to measure the performance of our basketball tracking algorithm: Area Under the Curve (AUC), Normalized Precision (P_{Norm}), and Precision (P). These metrics provide a comprehensive assessment of the tracking algorithm's effectiveness in various aspects. Below, we detail the significance and calculation of each metric.

The AUC metric evaluates the overall performance of the tracking algorithm by calculating the area under the Success Plot curve. The Success Plot is a graph that plots the success rate against different overlap thresholds, where the success rate is defined as the proportion of frames in which the overlap between the predicted bounding box and the ground truth exceeds a certain threshold. A higher AUC value indicates better tracking performance, reflecting a higher average success rate across all thresholds. In this study, AUC is expressed as a percentage to facilitate comparative analysis.

The P_{Norm} measures the accuracy of the tracking algorithm by evaluating the average distance between the predicted object position and the ground truth, normalized by the size of the object. This normalization allows for fair comparison across objects and scenes of different scales. P_{Norm} is particularly useful for understanding how precisely the tracker follows the object, regardless of its size. It is expressed as a percentage, with higher values indicating higher precision and better tracking accuracy. By using P_{Norm}, we can assess the algorithm's capability to maintain accurate tracking over time and varying conditions.

The P evaluates the accuracy of the tracking algorithm by calculating the proportion of frames where the center of the predicted bounding box is within a specified distance from the center of the ground truth bounding box. Precision is expressed as a percentage, representing the fraction of frames that meet this criterion. A higher Precision value indicates better tracking accuracy. This metric provides a straightforward and intuitive way to assess how closely the predicted positions match the actual positions of the tracked object, making it particularly relevant for practical applications where exact positioning is critical.

By utilizing AUC, P_{Norm}, and P, we can obtain a holistic understanding of our basketball tracking algorithm's performance. AUC provides insight into the overall success rate across various overlap thresholds, P_{Norm} evaluates the normalized tracking precision across different scales, and P measures the exact positional accuracy of the tracked object. These metrics together ensure a thorough evaluation, highlighting both the strengths and areas for improvement in our tracking approach.

4.4.3 Results and Analysis

Due to the special nature of the application scenario, we did not make comparisons on publicly available major benchmarks. Instead, we compared with the latest object

tracking methods using real-world datasets that we collected. Table 4.1 shows the experimental results of different methods on real basketball dataset.

Metrics	Methods								
	STARK	KeepTrack	DTT	SAOT	AutoMatch	TREG	DualTFR	TransT	Our method
AUC(%)	67.0	67.0	59.9	61.5	58.1	63.8	63.4	64.8	**68.2**
P_{Norm}(%)	76.5	76.6	66.8	70.8	68.4	74.0	71.9	73.7	**77.6**
P(%)	70.1	70.2	61.1	63.4	59.9	66.2	66.5	69.0	**73.7**

Table 4.1: Comparison of various methods on the dataset we collected, including STARK Yan et al. (2021), KeepTrack Mayer et al. (2021), DTT Yu et al. (2021), SAOT Zhou et al. (2021), AutoMatch Zhang et al. (2021), TREG Cui et al. (2021), DualTFR Xie et al. (2021), and TransT Chen et al. (2021). Black bold text indicates the best results.

The comparison results in Table 4.1 demonstrate that our method achieves the best performance across all three evaluation metrics: AUC, P_{Norm}, and P, with values of 68.2, 77.6, and 73.7, respectively. We analyze the key structural components of our approach and explain the reasons behind its superior performance compared to other methods.

Our method employs a template cropping mechanism, which significantly enhances the adaptability of the tracking algorithm. By continuously updating the template based on the latest visual context, our model can effectively handle variations in the appearance of the basketball players, such as changes in posture and occlusions. This dynamic updating helps maintain high tracking accuracy, contributing to superior performance in AUC and P_{Norm}.

We utilize ResNet-50, a robust convolutional neural network, for feature extraction. ResNet-50 is known for its deep architecture and residual learning capabilities, which allow it to capture rich and discriminative features from input images. These high-quality features are crucial for precise object tracking, providing a solid foundation for subsequent processing stages. The effectiveness of ResNet-50 is reflected in our method's overall tracking success metric AUC.

Our method leverages a Transformer Encoder and Decoder framework, which is

highly effective for capturing both spatial and temporal dependencies. The self-attention mechanism in the Transformer architecture enables the model to concentrate on key input features, enhancing the accuracy of both bounding box predictions and confidence scores. This architecture's capability to model long-range dependencies contributes significantly to the high P_{Norm} score.

The Bounding Box Prediction Head is designed to predict the precise location of the target. By using the rich features generated by the Transformer Decoder, this module can accurately estimate the target's bounding box. The precise bounding box predictions lead to high performance in AUC and P, as the model consistently locates the target accurately.

The Score Head evaluates the confidence of the predicted bounding boxes, ensuring that only the most reliable predictions are considered. This scoring mechanism helps in filtering out low-confidence predictions, thereby improving the overall tracking precision and robustness. The effectiveness of the Score Head is evident in the high P and P_{Norm} scores, indicating reliable and accurate tracking.

4.4.4 Ablation Study

Ablation studies help understand the contribution of each component in a model by systematically removing parts and observing the impact on performance. For the proposed basketball tracking algorithm, we will perform ablation studies on three key components: Dynamic Template (DT), Context Template (CT), and Score Head (SH). The ablation study results presented in Table 4.2 provide insight into the contributions of different modules to the overall performance of our proposed tracking method.

Removing the dynamic template (DT) mechanism results in a noticeable decrease across all performance metrics. The AUC drops by 1.9 points from 68.2 to 66.3, indicating a reduction in overall tracking performance. The P_{Norm} decreases by 3.4 points from

Dataset	Method	AUC(%)	P_{Norm}(%)	P(%)
Basketball Dataset	Method without the DT	66.3	74.2	71.6
	Method without the CT	67.4	76.2	72.3
	Method without the SH	67.6	76.8	72.9
	Proposed method	**68.2**	**77.6**	**73.7**

Table 4.2: The performance of our method after removing various modules.

77.6 to 74.2, suggesting a decline in the precision of tracking over different scales. The precision (P) also falls by 2.1 points from 73.7 to 71.6, showing that the ability to maintain accurate target positions is compromised. This highlights the importance of dynamic template cropping in adapting to changes in the target's appearance and improving tracking robustness.

Eliminating the context template (CT) causes a moderate decrease in performance metrics. The AUC drops by 0.8 points from 68.2 to 67.4, indicating a slight reduction in overall tracking performance. The P_{Norm} decreases by 1.4 points from 77.6 to 76.2, reflecting a reduction in normalized precision. The precision (P) falls by 1.4 points from 73.7 to 72.3, indicating a decrease in the accuracy of the bounding box predictions. These results underscore the role of the context template in providing additional contextual information that aids in accurate target localization.

Removing the score head (SH) leads to a slight drop in performance. The AUC decreases by 0.6 points from 68.2 to 67.6, showing a minor reduction in overall tracking performance. The P_{Norm} falls by 0.8 points from 77.6 to 76.8, indicating a slight decline in normalized precision. The precision (P) decreases by 0.8 points from 73.7 to 72.9, reflecting a reduction in the reliability of the tracking results. The score head is crucial for evaluating the confidence of the predicted bounding boxes and filtering out unreliable predictions.

4.5 Conclusion

This study presents a new tracking framework designed to effectively tackle the challenges of object tracking in complex sports environments, with a particular focus on basketball examinations. Our method integrates examinee information into the tracking model, significantly enhancing the robustness and accuracy of target identification and tracking, even under conditions of occlusions and overlapping identical objects. By leveraging an encoder-decoder transformer architecture, our approach captures global spatio-temporal feature dependencies and predicts object positions directly, providing a more reliable and stable tracking solution compared to traditional methods. Furthermore, we proposed a new dataset collected from real-world sports scenarios. This dataset is specifically designed to support research in human skeleton prediction, object tracking, and Re-ID. The dataset offers rich annotations and diverse scenarios, making it a valuable resource for advancing the state-of-the-art in these fields. Our experimental results demonstrate that our tracking framework not only improves the performance and reliability of intelligent sports examinations but also provides a robust foundation for subsequent analytical processes such as performance evaluation and data analysis. The continuous adaptation and learning capabilities of our system ensure that it remains effective in dynamic and unpredictable environments, thereby supporting advanced analytics and deeper insights into examinee performance.

CHAPTER 5

ENHANCING SKELETON-BASED HUMAN MOTION RECOGNITION WITH LIE ALGEBRA AND MEMRISTOR-AUGMENTED LSTM AND CNN

5.1 Introduction

As artificial intelligence evolves, particularly with advancements in deep learning neural networks, human action recognition has found extensive applications in healthcare Rafferty et al. (2017); Sun et al. (2022); Guo et al. (2021). Due to its wide application in video surveillance, autonomous driving, physical education, and other fields, it can greatly improve people's quality of life and simplify work processes Li et al. (2022). In the realm of human action recognition, visual approaches to represent human actions can be broadly categorized into three groups: RGB-oriented Finn et al. (2016), skeleton-driven, and those rooted in depth maps Liu et al. (2016a). Among them, bone-based (skeleton-based) representations have received extensive attention due to their viewpoint independence and ease of describing motion.

At present, the data collection technology of human body posture is becoming more

and more mature. There are 3D bone data acquisition devices such as Kinect on the hardware and Openpose Cao et al. (2017), which uses deep neural networks for bone recognition and synthesis on the software level. Human skeletal data can be seamlessly extracted from either videos or images. In the context of human action recognition, the manner in which actions are represented is pivotal. An optimal representation should precisely encapsulate the spatial dynamics associated with joints and bones. Predominantly, unit quaternions and Euler angles serve as the go-to methodologies to encapsulate human movement. However, the unit quaternion may cause numerical and analytical difficulties, and the Euler angle will have the problem of a Vientiane lock. Representation methods based on Lie groups and Lie algebras Vemulapalli et al. (2014) solve these problems well and provide a more reasonable representation method for the representation of human behavior. At the same time, using Lie algebra to represent bone data can not only gain computational advantages, but can also be combined with standard bone data, ignoring the effect of bone length. At the same time, the representation method based on Lie algebra divides the human skeleton data into five parts; we can calculate these five parts separately, and further realize the accurate judgment of an action.

At the same time, many model methods and target detection algorithms have also been proposed for human action recognition in deep learning, including energy-relation diagrams (ERD), 3 layers of long short-term memory (LSTM-3LR) Fragkiadaki et al. (2015), stacked recurrent network (SRNN) Jain et al. (2016), YOLO Redmon and Farhadi (2018), etc. These methods have made important contributions to human action recognition and target detection. However, an effective model often needs to combine data of multiple dimensions for calculations, and contains a large number of parameters, which requires a lot of computing power, and may consume a lot of time when processing skeleton data, which makes the model less applicable. Therefore, the model cannot achieve a good performance when dealing with real-time streaming information.

To address these issues, the advent of memristors has significantly contributed to accelerating model computation as one of the branches of neuromorphic computing. The emergence of memristors provides options for the hardware implementation of neuromorphic computing. Memristor crossbar-based networks can achieve extremely high parallel speeds and consumes very little energy during computation. This has comprehensive practical value Smagulova and James (2019).

This brief applies Lie algebra and standard bone length data to represent human skeleton data. A multi-layer LSTM recurrent neural network and CNNs are applied for human motion recognition. Finally, the trained network weights are converted into the crossbar-based memristor circuit, which can accelerate the network inference, reduce energy consumption, and obtain an excellent computing performance.

Our work advances human action recognition and neuromorphic computing with key contributions: (1) Implemented network structures with memristors, demonstrating minimal accuracy loss, and showcasing the efficiency of memristor technology in deep learning; (2) adapted the use of Lie algebra for skeletal representation within a memristor-based network structure for the first time, enhancing the integration of advanced motion capture techniques with neuromorphic computing; (3) and explored potential applications of memristors in neuromorphic computing, setting a foundation for future low-power, high-speed computing solutions.

5.2 Proposed Method

5.2.1 Skeletal Human Motion Representation

As a fundamental problem in human action-related tasks, there are currently three popular methods: RGB-based, depth-image-based, and bone-based methods. Considering generality and easy availability, this paper chooses a representation method based on

skeletal data Hu et al. (2019).

For a more refined representation of human action data, while mitigating the effect of skeletal length variation, we've adopted the methodologies of Lie algebra and the standard skeleton data representation. Lie algebras are mathematical structures essential in studying continuous symmetry and differential equations. They serve as the algebraic counterparts to Lie groups, which represent continuous transformations. A Lie algebra is defined as a vector space equipped with a binary operation known as the Lie bracket, adhering to bilinearity and the Jacobi identity. These properties ensure that the algebra captures the concept of infinitesimal symmetries. Lie algebras play a crucial role in mathematics and physics, especially in differential equations, geometry, and quantum mechanics. As visualized in Figure 5.1, we configure the local coordinate framework for e_m by imposing minimal rotation and translation adjustments to the global coordinate setup. This results in e_m serving as the definitive reference for the x-axis's position and orientation, taking its starting joints as the coordinate's inception point. Post this transformation, as demonstrated in Equation (5.1), the relative positioning of e_n within the localized system of e_m is discerned, signified as e_n^m. Subsequent to this, we engineer a 3D rigid transformation articulated as $\begin{pmatrix} R_{n,m} & d_{n,m} \\ 0 & 1 \end{pmatrix}$, where $R_{n,m}$ constitutes a 3×3 rotational matrix, and $d_{n,m}$ is the corresponding 3D translational vector, facilitating the shift of e_m to align with the position and orientation of e_n.

$$\begin{bmatrix} e^m_{n,end} \\ 0 \end{bmatrix} = \begin{bmatrix} R_{n,m} & d_{n,m} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} l_n \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{5.1}$$

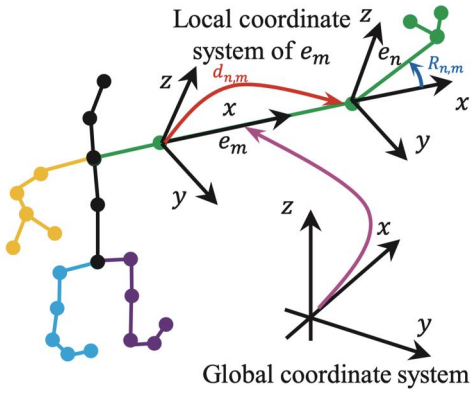

Figure 5.1: Lie Group Depiction of Skeletal Translation and Rotation.

In the context of this representation, $e^m_{n,end}$ signifies the terminal joint of e^m_n, and l_n denotes the length of e_n. Analogously, by employing a distinct transformation matrix, we ascertain e_m's position within the local system of e_n. Consequently, given M as the total count of bones, we derive $M \times (M-1)$ transformation matrices. From a computational standpoint, a 3D rigid transformation can be characterized within the framework of the special Euclidean group, denoted as $SE(3)$. Ultimately, a skeleton can be characterized as a trajectory in the multi-dimensional space $SE(3) \times ... \times SE(3)$.

To negate the impact of varying bone lengths, which essentially eliminates the influence of diverse body types on identical posture evaluations, we adopt a standard-length bone data for classification. This implies that, only the rotation matrix becomes essential for a human pose representation. Moreover, given that the human form can be depicted as a linkage structure with five primary segments - the spine, a pair of legs, and a pair of arms, as illustrated in Figure 2 - our focus is on computing the rotation matrix specifically between two contiguous bones sharing a joint, rather than between any arbitrary bones within a segment. This approach retains the inherent structure of the skeletal framework by honoring the anatomical constraints between chains. A subsequent advantage of this methodology is the reduced count of rotation matrices, thus

offering potential computational efficiencies.

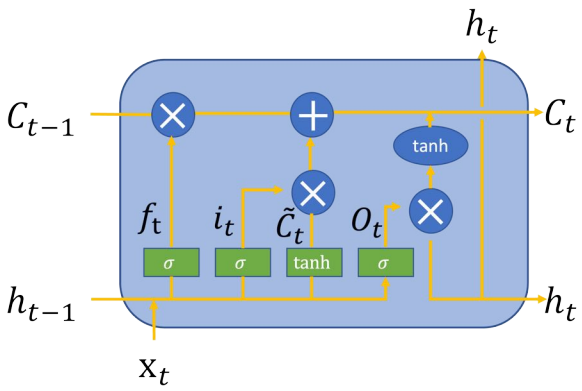

Figure 5.2: Schema of LSTM Unit.

Operationally, the axis-angle representation $(\mathbf{n}, \theta)$ is initially derived as follows:

$$\mathbf{n} = \frac{cross(e_n, e_m)}{\|cross(e_n, e_m)\|}, \tag{5.2}$$

$$\theta = arccos(e_n \cdot e_m). \tag{5.3}$$

Here, *cross* signifies the outer product, and · represents the inner product. Following this, the rotation matrix $R_{n,m}$ is inferred via the Rodriguez formula:

$$R_{n,m} = I + sin(\theta)\mathbf{n}^{\wedge} + (1 - cos(\theta))\mathbf{n}^{\wedge\mathbf{2}}. \tag{5.4}$$

In our discussion, $I \in \mathbb{R}^{3\times3}$ represents the identity matrix, while $\mathbf{n}^{\wedge}$ signifies the skew-symmetric matrix associated with $\mathbf{n}$. It's essential to recognize that this collection of rotation matrices is a member of the special orthogonal group $SO(3)$. Consequently, the skeleton can be envisioned as navigating a path in $SO(3) \times ... \times SO(3)$. Given the intricate nature of regression within the curved domain $SO(3) \times ... \times SO(3)$, we aim to convert this domain to its tangent space, which is seen as the Lie algebra $SO(3) \times ... \times SO(3)$. To

achieve this, we employ an approximate logarithm map method:

$$\omega(R_{n,m}) = \frac{1}{2sin(\theta(R_{n,m}))} \begin{bmatrix} R_{n,m}(3,2) - R_{n,m}(2,3) \\ R_{n,m}(1,3) - R_{n,m}(3,1) \\ R_{n,m}(2,1) - R_{n,m}(1,2) \end{bmatrix}, \tag{5.5}$$

$$\theta(R_{n,m}) = arccos(\frac{Trace(R_{n,m}) - 1}{2}). \tag{5.6}$$

In essence, the skeleton is projected onto a set of Lie algebra vectors, denoted as follows: $\omega = [\omega_1^{1^T}, ..., \omega_{K_1}^{1^T}, ..., \omega_1^{C^T}, ..., \omega_{K_C}^{C^T}]^T$, where C indicates the total chains (for our setup, C = 5, which mirrors human movement) and $K_c (c \in 1, ..., C)$ signifies the number of bones in the c-th chain reduced by one.

The bone representation method we've employed offers dual benefits: first, it negates the effect of bone length on the final outcomes; and second, it curtails the computational parameter count needed for the subsequent neural network processing.

5.2.1.1 Utilizing LSTM and CNN Architectures

In pursuit of an enhanced accuracy, this study integrates both LSTM and CNN architectures, as described by Li et al. (2017b), to handle data presented in the Lie algebra form. Recognizing that human skeletal action data encompasses both temporal and spatial dimensions, an amalgamation of LSTM and CNN networks is deemed optimal. This is because the LSTM structure excels at amalgamating temporal context features, while CNN thrives at spatial feature extraction Huang et al. (2023); Liu et al. (2020a); Wang et al. (2022a).

As an advanced version of the traditional RNN, LSTM proficiently captures long-range temporal characteristics. Importantly, it addresses the notorious gradient explosion or vanishing challenges encountered in conventional RNNs, marking it particularly adept for a time-series data analysis. A classic LSTM model is employed here. The concept of

gating–encompassing the input, forget, and output gates lies at the core of the LSTM's functionality. The mathematical computations within the LSTM unit are articulated as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_t), \tag{5.7}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{5.8}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \tag{5.9}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{5.10}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t, \tag{5.11}$$

$$h_t = o_t \otimes tanh(C_t). \tag{5.12}$$

Here, $\otimes$ represents the Hadamard product, C_t and h_t indicate the cell and hidden states, respectively, and f_t, i_t, and o_t distinguish between the forget, input, and output gates, respectively. Within the scope of this research, a tri-layered LSTM architecture is leveraged to mine temporal features from the dataset.

As previously highlighted, the LSTM network excels at extracting features in the temporal domain. To further amplify our model's classification efficacy, we've incorporated auxiliary network structures as delineated in Li et al. (2017b); Krizhevsky et al. (2012). AlexNet, which is a seminal deep CNN, has demonstrated a robust performance across various applied tasks. By integrating the capabilities of both the LSTM and AlexNet architectures, our approach is adept at capturing the nuanced interplay between the temporal and spatial features within the dataset. This synergy ensures that the strengths of one network offset any limitations of the other.

5.2.2 Memristor-based LSTM and CNN

Both LSTM and CNN networks have a very large amount of parameters, which makes practical applications difficult. In many edge computing devices, the computing power that meets the conditions cannot be provided. At the same time, for the future computing systems, power consumption and speed are two goals currently pursued. Our ultimate goal is to want low-power, fast computing devices. While neuromorphic computing has significant advantages, it can solve complex problems while consuming very little power and area Taha et al. (2013). This feature gives neuromorphic computing the ability to be widely used.

In recent years, the memristor Chua (1971) has received great attention as one of the directions of neuromorphic computing. Memristors, which are defined as memory resistors, are passive electronic components capable of retaining a voltage history, thus embodying a non-volatile memory function. This feature facilitates their application in neuromorphic computing systems by emulating synaptic connections. Unlike binary-operating transistors, memristors support analog computations through variable resistance values, enhancing the computing efficiency by enabling complex computations within memory units, thereby minimizing the data transfer between the processors and memory. This physical property, which can be tuned to a specific resistance value by applying a voltage to change its conductivity, is crucial for its functionality. Remarkably, this characteristic can be retained in the memristor even after a power down. By organizing memristors into a grid of crossbars Wen et al. (2018a); Jo et al. (2009), many neural network computations can be performed in parallel, further leveraging the unique capabilities of memristors in neuromorphic computing architectures.

5.2.2.1 Memristor Crossbar

As shown in Figure 5.3, a single layer feed forward neural network is implemented by using a 5×6 crossbar with four inputs and three outputs. Memristors are placed at the intersections of the bar structure and represent the weights of the network. Thanks to this special structure, the input can be processed in parallel, resulting in a faster speed Hasan et al. (2017). Similarly, by leveraging the output from the prior crossbar layer as the input for the subsequent one, we can construct a multi-tiered feedforward neural network.

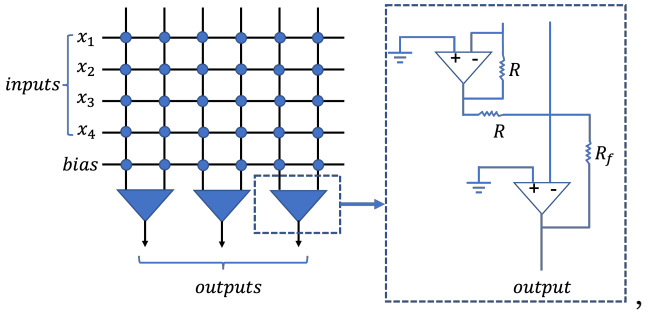

Figure 5.3: Schema of Memristor Crossbar.

5.2.2.2 Memristor-based LSTM

ANNs have become a cornerstone in the field of machine learning, mimicking the structure and function of the human brain's neural networks. These computational models consist of nodes or neurons, organized in layers, that process input data through a series of transformations and connections. The most basic form of these networks includes fully connected layers, where each neuron in one layer connects to every neuron in the subsequent layer, thus facilitating the learning of complex patterns in the data.

The transition from theoretical neural network models to practical applications within computer systems has been marked by significant advancements in the computational power and algorithms. This evolution has enabled the implementation of complex neural

network architectures, such as CNNs for image processing and RNNs for sequential data analyses. Among these, LSTM networks, a specialized form of RNNs, have been particularly effective in capturing long-term dependencies in the sequence data, which is a critical aspect in fields such as natural language processing and time series forecasting.

There are already many LSTM circuits implemented with memristors. A crossbar based LSTM architecture was proposed Wen et al. (2019), and the effectiveness of the structure was demonstrated by a textual sentiment analysis. Then, an on-chip trained LSTM, namely the MbLSTM, was proposed in Liu et al. (2020b) . Similar to Wen et al. (2019), the activation functions sigmoid and tanh were approximately implemented through intentionally designing circuit parameters. In this paper, in order to realize the LSTM network structure, we adopt the scheme in Liu et al. (2020b). Instead, we ended up using ex-situ training to write the trained weights into the LSTM architecture.

According to the architecture in Liu et al. (2020b), the general structure of LSTM cell is shown in Figure 5.4. Thus we can get the following:

$$c^t(k) = -i^t(k) \cdot [-a^t(k)] - [-f^t(k) \cdot c^{t-1}(k)] \tag{5.13}$$

and

$$h^t(k) = -o^t(k) \cdot tanh(c^t(k)), \tag{5.14}$$

where tanh in (5.14) is the approximate activation function implemented by a circuit. Moreover, the multiplication is perfermed by existing analog multipliers. $h^t(k)$ is converted to $[-V_r, V_r]$ for the next step

$$V_h^t(k) = \frac{R_4}{R_3} h^t(k). \tag{5.15}$$

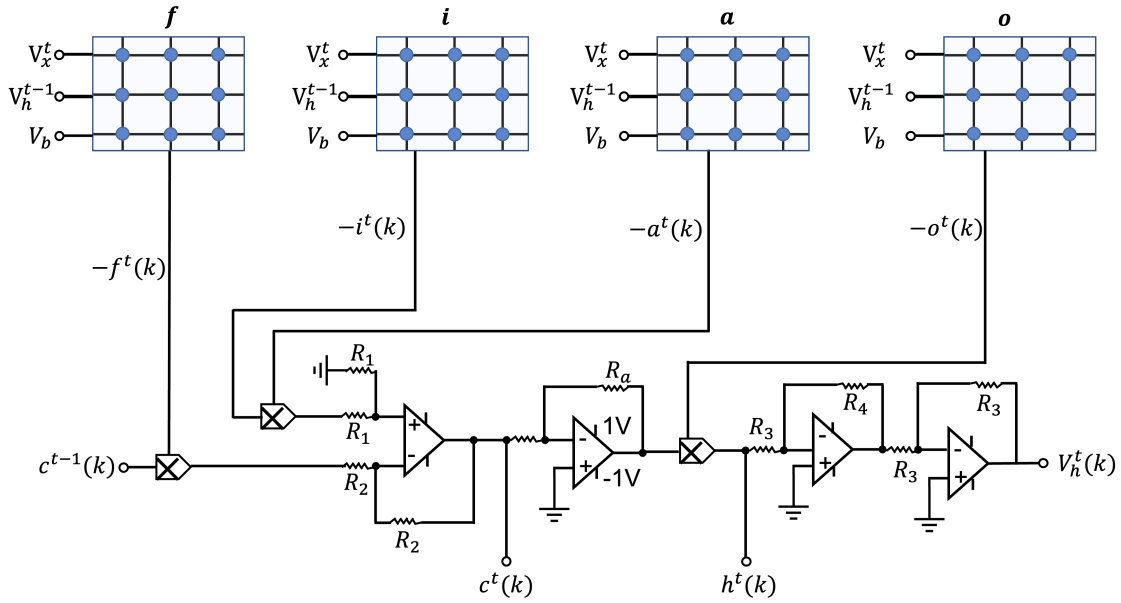

Figure 5.4: Memristor crossbar based LSTM cell, where f, i, a, o are four memristor based LSTM units.

5.2.2.3 Memristor-based CNN

As another auxiliary network of the overall network, CNNs can capture spatial features well. In Yakopcic et al. (2016), the authors proposed a simulated memristor crossbar implementation of the CNN. In this structure, the convolution of the image is not done once, but divided into multiple iterations. Thus, considering the size of the memristor crossbar, an image is divided into multiple inputs, and the final convolution results are spliced to obtain the final result. Certain arrangements were made through the convolution kernel to realize the CNN structure that processed the entire picture at one time in Yakopcic et al. (2017). However in this structure, if the size of the picture increases, the number of memristors significantly increases, which is one of the drawbacks of this method. Meanwhile, a new convolution method was proposed to reduce the parameters by about 75% and reduce the number of multiplication computations for the convolutional layers by 30% within an acceptable accuracy loss Wen et al. (2020). A fully hardware-implemented memristor convolutional neural network was proposed in Yao et al. (2020).

In this brief, we consider the possibility of a practical application and reduce the number of the memristor. We adopt the structure in Yakopcic et al. (2016) to implement the CNN. Figure 5.5 shows a single column of the memristor crossbar for performing convolution. Same as in Yakopcic et al. (2016), we set $V_{S1} = -1V, V_{D1} = 0V, V_{S2} = 0V, V_{D2} = 1V$, M_g is a memristor used to control the feedback gain, σ_β is the bias, and R_f is the unity gain. In this structure, the convolution kernels are determined in advance during the network training process. Since the convolution kernel may have negative values, in order to allow the convolution operation to process both positive and negative values, the convolution kernel and input values are divided into two column vectors:

$$\begin{bmatrix} 0.1 & -0.2 & 0.3 \\ -0.4 & 0.5 & -0.6 \\ 0.7 & -0.8 & 0.9 \end{bmatrix} \rightarrow \begin{bmatrix} 0.9 \\ -0.6 \\ 0.3 \\ -0.8 \\ 0.5 \\ -0.2 \\ 0.7 \\ -0.4 \\ 0.1 \end{bmatrix} \rightarrow \begin{bmatrix} 0.9 \\ 0 \\ 0.3 \\ 0 \\ 0.5 \\ 0 \\ 0.7 \\ 0 \\ 0.1 \end{bmatrix} \begin{bmatrix} 0 \\ 0.6 \\ 0 \\ 0.8 \\ 0 \\ 0.2 \\ 0 \\ 0.4 \\ 0 \end{bmatrix}. \tag{5.16}$$

As shown in (5.16), a convolution kernel will be rearranged into two column vectors, each storing the absolute value of the original value of the convolution kernel. One column

represents positive values and the other column represents negative values:

$$\begin{bmatrix} 0 & 0.5 & 0.3 \\ 0.5 & 0.8 & 0.5 \\ 0 & 0.5 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 0.5 \\ 0.3 \\ 0.5 \\ 0.8 \\ 0.5 \\ 0 \\ 0.5 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 0.5 \\ 0.3 \\ 0.5 \\ 0.8 \\ 0.5 \\ 0 \\ 0.5 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ -0.5 \\ -0.3 \\ -0.5 \\ -0.8 \\ -0.5 \\ 0 \\ -0.5 \\ 0 \end{bmatrix}. \tag{5.17}$$

As shown in (5.17), for input x, the permutation method is different. x is divided into two columns, each containing all the values in x, one positive and another negative.

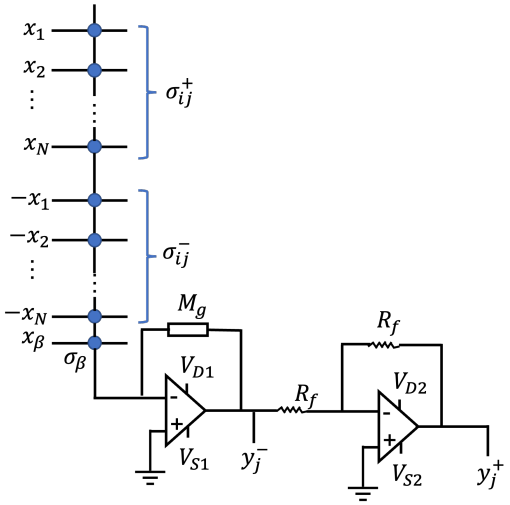

Figure 5.5: A Single Column of Memristor Crossbar for Performing Convolution.

Finally, through (5.18), the convolution kernels are converted to conductivity values:

$$\sigma^{\pm} = \frac{(\sigma_{max} - \sigma_{min})}{max(|W|)} W^{\pm} + \sigma_{min}. \tag{5.18}$$

For the final output activation function sigmoid, a circuit simulation is also used here to approximate the sigmoid function Wen et al. (2019); Liu et al. (2020b); Yakopcic

et al. (2016). At this point, we implement a single convolution operation. Based on the convolution operation under this structure, it is impossible to process all input values at one time; therefore, it is necessary to divide a feature map into multiple inputs, then ultimately splicing to obtain the result of the convolution operation. To some extent, this approach reduces the space of the memristor, sacrificing a certain amount of time.

5.2.2.4 Dataset

In this brief, we use H3.6M dataset Ionescu et al. (2013), which contains 3.6 million 3D skeleton data of human action sequences, and the NTU RGB+D Dataset Shahroudy et al. (2016), which is a comprehensive collection encompassing 56,578 samples of 60 distinct action categories. These actions are captured from multiple perspectives, including a frontal view, two lateral views, and oblique views at 45 degrees to the left and right. The dataset features performances by 40 participants, whose ages range from 10 to 35 years, providing a diverse basis for action recognition research. According to the method from Hu et al. (2019); Du et al. (2015), we transformed the 3D data into Lie algebras; in order to exclude the effect of bone length on the classification, we used a uniform standard bone length.

5.2.3 System Structure Overview

The architecture of our system is depicted in Figure 5.6. Initially, the skeleton data from the dataset is transformed into the Lie algebra representation. This approach diverges from traditional methods by utilizing skeletal data encoded in the Lie algebra, as opposed to the direct use of the skeleton data. Inspired by the methodologies in Li et al. (2017b,a); Wang et al. (2018b); Hu et al. (2019), we compute temporal-domain features (TPF) from the transformed data. A key modification in our process is reshaping the Lie algebra-encoded skeletal data to align with the TPF extraction techniques described in

these references, ensuring our method remains consistent with established practices. Unlike, Li et al. (2017b) where the LSTM network inputs were spatial-domain features (SPF), our model's inputs are action sequences transformed into a Lie algebra.

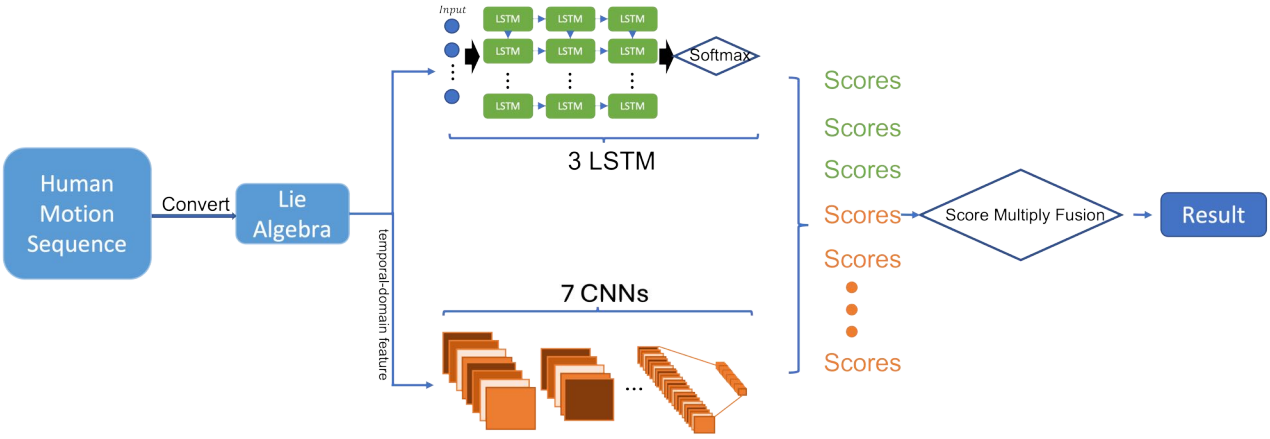

Figure 5.6: Overview of the Proposed System.

In our model, frame indices are denoted by $i \in (1,...,T)$ and elements within the Lie algebra vector ω by $j \in (1,...,K)$, where $K = \sum_{c=1}^{C} K_c$. For simplicity, we refer to elements in ω as bones. Our three-layer LSTM architecture processes this data in stages: the first layer captures the overall motion information from the bones represented in Lie algebra; the second layer employs a dedicated LSTM to model the spine; and the final layer uses another set of LSTMs to analyze the remaining skeletal parts.

For computing the final output score of each network, we adopt a multiply-score fusion method as described in (5.19):

$$label = Fin(max(v_1 \circ v_2 \cdots v_9 \circ v_{10})). \tag{5.19}$$

In this context, v represents the score vector, with $\circ$ signifying element-wise multiplication. Meanwhile, Fin identifies the index corresponding to the maximum element.

First, we train the weights of the network via software network, and then map the weights to the memristor circuit through a transformation. The resulting circuit

achieves a significant improvement in the inference speed over the software-implemented network.

5.2.4 Experiment and Result

The Human 3.6M dataset contains 3.6 million 3D human pose data, including 17 scenes: discussion, smoking, taking pictures, talking on the phone, and so on. First, we convert the dataset to a Lie algebra representation. Compared with the original representation, the human pose data represented by the Lie algebra is more conducive to a calculation, and we use the standard bone length to replace the original bone length, excluding the influence of bone length on classification.

Then, we implement and train networks by Pytorch Paszke et al. (2019). The result is shown in Table 5.1. We adopt the network architecture in Li et al. (2017b), which is combined with three LSTM networks and seven multi-layer CNNs. We make a small change in the front part of the network structure. For the input to the LSTM network, our structure contains the transformed Lie algebra. At the same time, we also adopt the method of calculating a TPF for the input of the CNN network. We train the weights of the network on the software and obtain an accuracy rate close to Ref. Li et al. (2017b).

In this section of our study, we employ a simulated memristor architecture using MemTorch Lammie et al. (2022), a simulation platform for memristive deep learning systems that seamlessly integrates with the PyTorch machine learning (ML) library. MemTorch facilitates the emulation of memristor crossbars and enables direct interaction with PyTorch, allowing for the straightforward mapping of network structures such as LSTM and CNN onto the crossbar architecture. We utilize MemTorch's capability to map both the network structure and the weights for simulated inference, opting to utilize perfect-state memristor structures despite MemTorch's support for modeling imperfect memristor properties.

Dataset	Method	Cross Subject	Cross View	Accuracy
H3.6M	All-Mul-Score fusion (LSTM+CNN) (Software Implementation)	81.78%	88.97%	86.31%
	All-Mul-Score fusion (LSTM+CNN) (Memristor Simulation)	**80.67%**	**88.44%**	**85.98%**
NTU RGB+D	All-Mul-Score fusion (LSTM+CNN) (Software Implementation)	82.78%	91.13%	86.53%
	All-Mul-Score fusion (LSTM+CNN) (Memristor Simulation)	**79.80%**	**87.97%**	**83.31%**

Table 5.1: Experimental results on H3.6M and NTU RGB+D Datasets. A large number of experiments have shown that the structure based on Memristors consumes much less energy than traditional software-simulated neural networks Wen et al. (2019); Liu et al. (2015); Wen et al. (2018b).

To further validate the acceleration capabilities of memristors, we conduct tests on individual neurons using a simulated circuit setup to calculate power consumption. In these tests, we focus on the maximum values of the input and internal weights derived from our trained model. For a fair comparison, we use these maximum weights throughout the simulations. The results, summarized in Table 5.2, underscore the significant reduction in hardware complexity and power consumption achieved with memristor-based synapses compared to traditional CMOS-based designs.

	CMOS-based	**Memristor-based**
Hardware Units per Synapse	16	1
Max Power of A Synapse (μW)	≈76.0	9.7

Table 5.2: Comparison of Power Consumption for Single Neuron: CMOS-based vs. Memristor-based Systems.

Drawing on insights from existing literature Budiman et al. (2018); Sarwar et al. (2013); Wen et al. (2019, 2018a), it is essential to note that employing actual memristor structures can further reduce power consumption and accelerate inference speeds com-

pared to the simulated memristor structures used in our study. However, within the scope of this research, we choose to focus on the simulation aspects of memristor-based systems, given the constraints of our experimental setup, rather than empirically demonstrating these potential enhancements.

5.3 Conclusions

This study explores the application of memristor-based circuits to simulate neural networks in the context of human action recognition using skeletal data. By leveraging Lie algebra and standardized bone length data for an efficient representation of human skeletons, we demonstrate the feasibility of using memristor technology to approximate the functionality of multi-layer LSTM recurrent neural networks combined with CNNs. Our work contributes to the field by showcasing a novel use of memristor circuits for network inference, which offers a promising avenue for reducing energy consumption and accelerating inference times in deep learning models.

A pivotal aspect of our research focuses on the construction of networks using memristor circuits, which are capable of achieving performance metrics closely approximating those of software-simulated networks. Although our memristor network implementation remains within the realm of the simulation, the inherent efficiency and low power consumption of memristor structures are well-documented. This approach not only addresses critical challenges in deploying deep learning models for real-time applications, but also highlights the potential of the memristor technology as a sustainable and efficient computing alternative to traditional, power-intensive computational methods.

Furthermore, we illustrate that it is possible to maintain a balance between computational efficiency and model accuracy, which is often a significant challenge in optimizing deep learning models. The ability to achieve near-original performance metrics with

memristor-based simulations underscores the potential of our method for broad applications in various sectors, including healthcare, autonomous driving, surveillance, and sports analytics.

In conclusion, our research highlights the viability of memristor-based deep learning systems for human action recognition, marking a step towards the practical implementation of energy-efficient and fast neural network simulations. The implications of our work are far-reaching, suggesting a future where memristor technologies play crucial roles in enabling real-time, energy-efficient, and accurate computational tasks across diverse applications.

CHAPTER 6

VIDEO BASED INTELLIGENT SPORTS ANALYSIS SYSTEM FOR OBJECTIVE SPORTS EXAMINATIONS

Ensuring the practical applicability and scalability of proposed methods in real-world sports examination scenarios is a significant challenge. Despite advancements in methodologies addressing key issues in sports examinations, their real-world application and adaptability remain under-explored. This chapter aims to investigate the practical implementation and scalability of these methods, with a focus on basketball skill assessments. By evaluating the performance and robustness of the proposed intelligent sports examination system in real-world scenarios, we aim to identify potential barriers and requirements for practical deployment. This includes understanding the variability in examination conditions, the diversity of sports disciplines, and the logistical aspects of scaling these systems.

In this chapter, we first discuss the motivations and challenges of deploying intelligent sports examination systems in Section 6.1. Section 6.2 provides a review of related work, highlighting existing solutions and their limitations. Section 6.3 introduces the design of our modular and scalable intelligent sports examination system, detailing the integration

of robust keypoint prediction, accurate object tracking, and efficient computational processing. Section 6.4 describes the datasets used for evaluating the system, covering both basketball skill assessments and other sports scenarios to test adaptability. In Section 6.5, we present the experiments and evaluations conducted to assess the system's performance and scalability in real-world conditions. Section 6.6 offers a discussion on the findings, addressing the practical challenges and proposing solutions for effective deployment. Finally, Section 6.7 concludes the chapter, summarizing the key insights and future directions for research.

This chapter aims to develop and test a comprehensive, modular intelligent sports examination system that ensures practical applicability and adaptability across diverse sports and dynamic environments, ultimately facilitating the transition from theoretical advancements to real-world implementation.

6.1 Introduction

Sports exams are widely used for evaluating the performance of athletes in various sports disciplines, including basketball, soccer, and track and field. These exams are typically conducted by human examiners who evaluate the performance of athletes based on specific criteria. However, such evaluations can be subjective, leading to inaccurate results and unfair outcomes. In recent years, computer vision and machine learning techniques have emerged as promising solutions to address the subjectivity and inaccuracy issues associated with sports exams. These techniques can help in tracking the movements of athletes, identifying errors, and providing objective and accurate scores.

One critical aspect of computer vision and machine learning techniques in sports exams is the ability to track the movements of athletes and objects. Object tracking is a technique that involves detecting and tracking the position of objects in a video sequence.

It is widely used in sports exams to track the movements of balls, hurdles, and other objects. Several studies have proposed different object tracking algorithms, such as the kernelized correlation filter-based approach Henriques et al. (2014), deep learning-based approaches Nam and Han (2016), and multi-object tracking approaches Berclaz et al. (2011).

Another crucial aspect of computer vision and machine learning techniques in sports exams is the ability to predict human body movements accurately. Human body prediction involves identifying the position and orientation of body parts in a video sequence. One widely used technique for human body prediction is skeleton prediction. Skeleton prediction involves detecting the position and orientation of human joints in a video sequence. Several studies have proposed different skeleton prediction algorithms, such as the graph-based approach Shotton et al. (2011), the physics-based approach Poppe (2007), and the deep learning-based approach Cao et al. (2017).

Moreover, action recognition is another important aspect of computer vision and machine learning techniques in sports exams. Action recognition involves identifying the type of movement performed by an athlete in a video sequence. Several studies have proposed different action recognition algorithms, such as the deep learning-based approach Carreira and Zisserman (2017), the 3D convolutional neural network-based approach Tran et al. (2015), and the spatiotemporal attention-based approach Girdhar et al. (2017).

However, the integration of these techniques into an intelligent sports analysis system for objective sports exams remains a challenging task. One of the primary challenges is the interference-resistant person and object tracking in the exam scenarios. Another challenge is the accurate classification of bone motion sequences and error point identification. Therefore, this study aims to develop an intelligent sports analysis system for objective sports exams using computer vision and machine learning techniques. The

proposed system will address challenges such as interference-resistant person and object tracking, accurate classification of bone motion sequences, and error point identification. The system will be tested using basketball exams as an example, and its potential applications will be evaluated.

6.2 System Design

This section provides an overview of our system design, explaining the rationale and methodology behind our architectural choices. Our goal is to create a highly available, modular, scalable, and portable system for various sports examination scenarios. By adhering to these principles, we aim to develop a robust platform that meets current needs and anticipates future advancements. The architecture emphasizes real-world applicability and operational efficiency, focusing on high availability to minimize downtime, modularity for easy upgrades and maintenance, and scalability to handle varying loads. Portability ensures the system can be adapted to different sports examinations with minimal reconfiguration. Our design leverages cutting-edge technologies and best practices in software engineering. As illustrated in Figure 6.1, the system's structure begins with the front end collecting video data and exam information, while the backend interface processes requests from the front end. The prediction module includes human key point prediction and object tracking. Results are sent to the Rule module for scoring, then visualized and returned to the front end. The following sections detail each system component, their roles, interactions, and the underlying technologies driving their functionality, demonstrating how they collectively achieve reliability, efficiency, and adaptability.

Front-End System. As illustrated in Figure 6.2, the front-end system is a sophisticated web application developed using Java, designed to provide an interactive and user-friendly interface. It is primarily responsible for facilitating seamless interaction

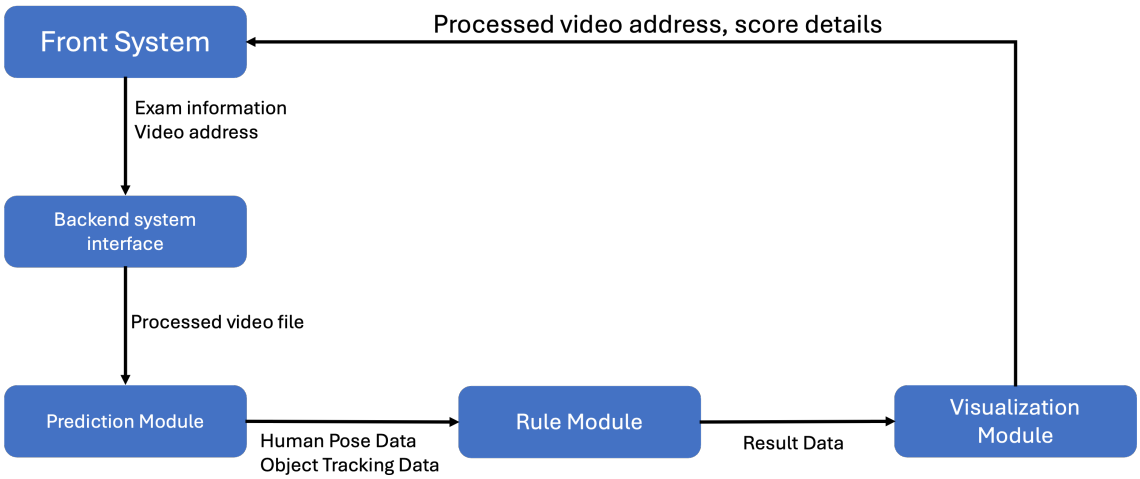

Figure 6.1: This figure illustrates the system's structure. The front end collects video data and exam information, while the backend interface processes requests from the front end. The prediction module includes human key point prediction and object tracking. Results are sent to the Rule module for scoring, then visualized and returned to the front end.

between the users and the underlying hardware and software components. The system's functionalities can be categorized into several key tasks:

- **Camera Interaction and Video Recording:** The front-end system interfaces with multiple cameras, leveraging their APIs to initiate and control video recording sessions. This involves capturing real-time video streams, handling synchronization issues, and ensuring that the video data is correctly encoded and stored for subsequent processing.

- **Video Playback and Feedback:** Users can review recorded videos through the front-end interface. This feature includes capabilities for playing back videos, navigating through video timelines, and cropping videos as needed. The system provides intuitive controls and feedback mechanisms to ensure a smooth user experience.

- **Submission of Examination Data:** The system allows users to submit examination-

related information, including the recorded video footage, to the back-end APIs. This process involves packaging the video data along with relevant metadata, securely transmitting the data over the network, and handling any potential transmission errors or interruptions.

- **Score Calculation Requests:** Once the examination data is submitted, the front-end system sends requests to the back-end for score calculation. This involves invoking RESTful APIs, handling asynchronous responses, and ensuring that the requests are processed efficiently.

- **Result Video Playback and Score Display:** After the back-end processes the examination data and computes the scores, the front-end retrieves and displays the results. This includes playing the result videos, if applicable, and presenting detailed score statistics through an interactive and visually appealing interface. The system ensures that the data is presented in a clear and comprehensible manner, facilitating easy interpretation by the users.

The architecture of the front-end system is designed to be modular and extensible, allowing for easy integration of new features and components. It employs modern web development frameworks and follows best practices in terms of security, performance, and usability. The use of Java provides robustness and portability, ensuring that the system can be deployed across various platforms with minimal modifications.

Overall, the front-end system plays a crucial role in bridging the gap between the users and the complex back-end processing, providing an efficient, reliable, and user-centric interface for the examination process.

Backend System Interface. The Backend System Interface, shown in Figure 6.3, is crucial for seamless interaction with the front-end system. It is responsible for handling requests from the front end, which include examination data, video information, and

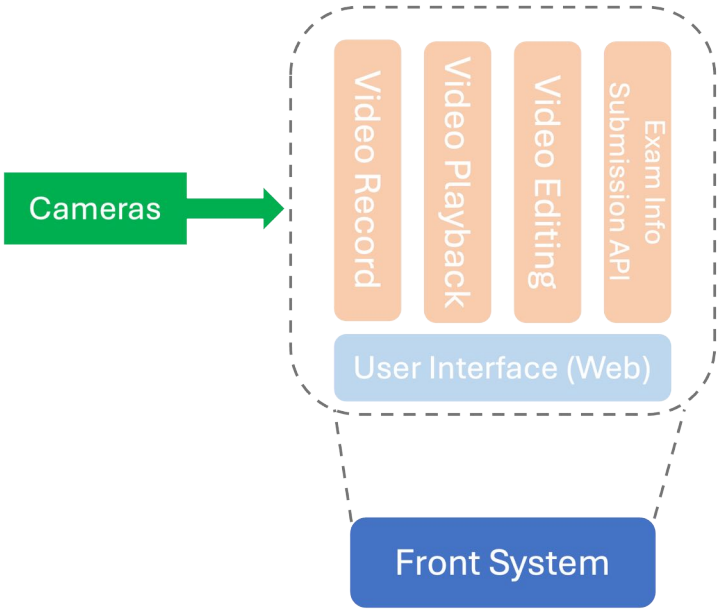

Figure 6.2: This diagram depicts the overall structure of the front-end system. This system is responsible for interacting with cameras, recording videos, playing back videos, and submitting exam information functions, while also interacting with users through a web page.

other relevant statistics. This section details the backend's role in processing these requests and interfacing with subsequent deep learning models.

Upon receiving examination requests from the front end, the backend system initiates a series of preprocessing tasks. These tasks include resizing the video to ensure uniform input dimensions and extracting individual frames from the video. This preprocessing step is crucial for preparing the video data for analysis by deep learning models.

The backend system is built using Flask, a lightweight and efficient web framework. Flask provides a robust foundation for building scalable and maintainable web applications. It enables the backend to handle multiple concurrent requests and ensures efficient communication with the front-end system.

The interaction between the backend system and the deep learning models is another key functionality. Once the video data is preprocessed, the backend forwards it to the

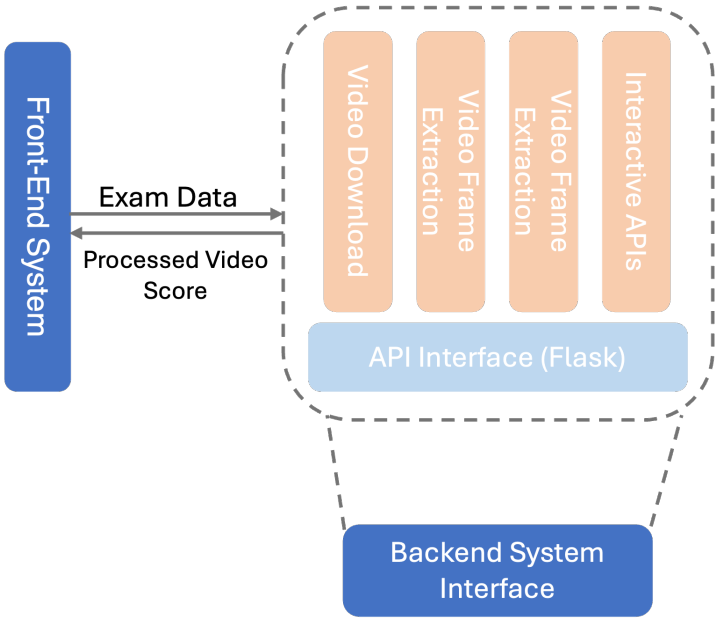

Figure 6.3: This diagram depicts the backend system API structure. The API interface built with Flask can interact with the frontend system, receive video and exam information, preprocess videos, and provide APIs for interacting with the backend deep learning module.

appropriate deep learning models for further analysis. The results from these models are then processed and sent back to the front end, providing users with detailed examination scores and statistics.

By leveraging Flask, the backend system ensures a modular and extensible architecture, allowing for easy integration of additional features and models in the future. This design choice also facilitates the maintenance and scalability of the system, ensuring it can handle increased loads and new examination scenarios as needed.

In summary, the Backend System Interface plays a pivotal role in the overall system architecture by managing the flow of data between the front-end interface and the deep learning models. Its design ensures efficient preprocessing, reliable communication, and scalability, aligning with the system's goals of high availability and adaptability.

Prediction Module. The Prediction Module consists of two key components: the

Human Skeleton Prediction Module and the Object Tracking Module, as shown in Figure 6.4. These components are designed to handle the diverse requirements of various examination scenarios by performing specialized predictions based on the video information and frames received from the backend system.

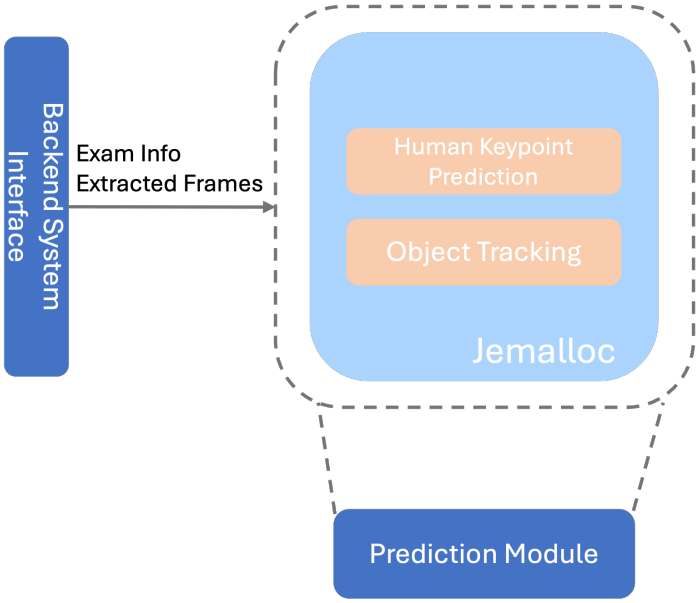

Figure 6.4: This image describes the structure of the Prediction Module. We use two deep learning modules, namely the human skeleton prediction module and the target tracking module, while using Jemalloc as memory assistance to help reclaim memory during the inference process.

The Human Skeleton Prediction Module is responsible for analyzing the human body's movements and postures. This module utilizes advanced deep learning models to predict the skeletal structure from the input video frames. The predicted skeletal data is crucial for assessing the performance and techniques of examinees in sports examinations, providing detailed insights into their movements and form.

The Object Tracking Module focuses on identifying and tracking specific objects within the video frames. The objects to be tracked vary depending on the nature of the examination. For instance, in a basketball examination, the module would track the ball, whereas in a running test, it might track the examinee's position relative to the track.

This module employs sophisticated algorithms to maintain accurate tracking throughout the video, ensuring reliable and consistent results.

Both modules receive video information and frames from the backend system and perform their respective predictions. The outputs of these predictions are then used to generate detailed examination scores and feedback, which are sent back to the front end for user review.

During practical deployment, we encountered issues with memory management, particularly with the deep learning models experiencing delayed memory reclamation, leading to memory crashes. To address this, we integrated Jemalloc, a memory allocator known for its efficient handling of memory fragmentation and allocation overheads. Jemalloc significantly improves memory management, ensuring timely memory reclamation and preventing crashes, thereby enhancing the overall stability and performance of the Prediction Module.

In summary, the Prediction Module is a critical component of the system, designed to provide accurate and reliable predictions through its Human Skeleton Prediction and Object Tracking modules. By addressing memory management challenges with Jemalloc, we ensure that the module operates efficiently, contributing to the system's goals of high availability and robustness.

Rule Module. The Rule Module is a critical component designed to address the limitations of traditional deep learning models in accurately assessing examination performance under diverse conditions. This module encompasses a variety of tasks such as video stream classification and skeleton sequence classification. However, in practical examination scenarios, these models often face challenges such as the inability to precisely evaluate the completion of actions and the requirement for extensive, high-quality datasets, which are difficult to obtain.

Examinations demand not only the evaluation of whether actions are performed correctly but also the assessment of the usage of examination tools. For instance, in a basketball examination, the evaluation involves the examinee's position, gestures, the positions of the basketball and the hoop, and whether the shot is successful. Traditional deep learning models struggle to address these specific needs effectively.

Therefore, our system employs the Rule Module, which consists of a series of heuristic rules to perform these assessments. This module receives the prediction results from the Prediction Module, including the skeleton sequences and object tracking sequences. By aligning these two sequences, the Rule Module evaluates each frame based on predefined rules tailored to the specific examination requirements.

For instance, in a basketball examination, the module would check the alignment of the skeleton sequence (indicating the examinee's movements and gestures) with the object tracking sequence (indicating the positions of the basketball and hoop). The rules might include criteria such as the positioning of the examinee‚Äôs hands relative to the ball, the trajectory of the ball, and whether the ball passes through the hoop. Each frame is assessed to determine if the action is successful or if any violations occur.

By using a rule-based approach, the system can flexibly adapt to various examination scenarios without the need for extensive and precise datasets required by traditional deep learning models. This approach ensures that the module can effectively evaluate complex actions and tool usage, providing accurate and reliable results.

In summary, the Rule Module plays a vital role in the system by leveraging heuristic rules to overcome the limitations of deep learning models. It ensures comprehensive assessment by aligning and evaluating the skeleton and object tracking sequences, thereby meeting the specific requirements of various sports examinations.

Visualization Module. The Visualization Module is a crucial component designed to

enhance user interaction by providing a clear and comprehensive visual representation of the examination results. This module directly receives the output from the Rule Module and transforms the raw data into an easily interpretable format.

Upon receiving the scores and penalty information from the Rule Module, the Visualization Module overlays this information onto the corresponding video frames. This includes not only the final scores but also specific deductions or penalties incurred during the examination. By integrating these visual cues directly into the video, the system provides an intuitive understanding of the performance and areas for improvement.

Additionally, the Visualization Module visualizes the skeletal data and object tracking information. This involves rendering the predicted skeletal structures and tracked objects within the video frames, allowing users to see the alignment and accuracy of their movements and the interaction with the examination tools. This visual feedback is crucial for users to understand the precise mechanics of their actions and how they were evaluated.

The key functionalities of the Visualization Module include:

- **Overlaying Scores and Penalties:** The module annotates the video with scores and penalties, highlighting specific moments where points were gained or lost. This provides a clear and immediate understanding of the examinee's performance throughout the video.

- **Visualizing Skeletal Information:** The predicted skeletal data is overlaid on the video, showing the user's movements in a clear and structured manner. This helps in analyzing posture and movement accuracy.

- **Visualizing Object Tracking Information:** The tracked objects, such as balls or other tools, are highlighted within the video frames, illustrating their interaction

with the examinee. This visualization aids in understanding the context of the performance relative to the examination tools.

By providing these visual enhancements, the Visualization Module ensures that users can easily interpret the examination results and understand the detailed aspects of their performance. This module plays a vital role in making the system user-friendly and effective for training and feedback purposes.

In summary, the Visualization Module transforms the raw output from the Rule Module into a comprehensive visual format. It overlays scores, penalties, skeletal data, and object tracking information onto the video, offering users clear and actionable insights into their performance. This visualization is essential for effective feedback and continuous improvement in various sports examinations.

6.3 Dataset

To evaluate our system, we collect a dataset from a real youth level three basketball examination conducted in Shanghai, China. As illustrated in Figure 6.5, the examination comprises nine different items, each recorded using standard cameras with 1080P resolution at 30 FPS. The data collection process is meticulously designed to capture comprehensive information across multiple dimensions of the examination. The dataset, which we collect, is detailed in Chapter 4, Section 4.3. The shooting angle of the video is as depicted in the figure.

The examination takes place across five distinct areas on two courts, with each item being recorded from various angles to ensure a comprehensive dataset. Typically, 2-3 cameras are used per item to capture different perspectives, providing a robust foundation for subsequent analysis. This multi-angle recording setup not only enhances

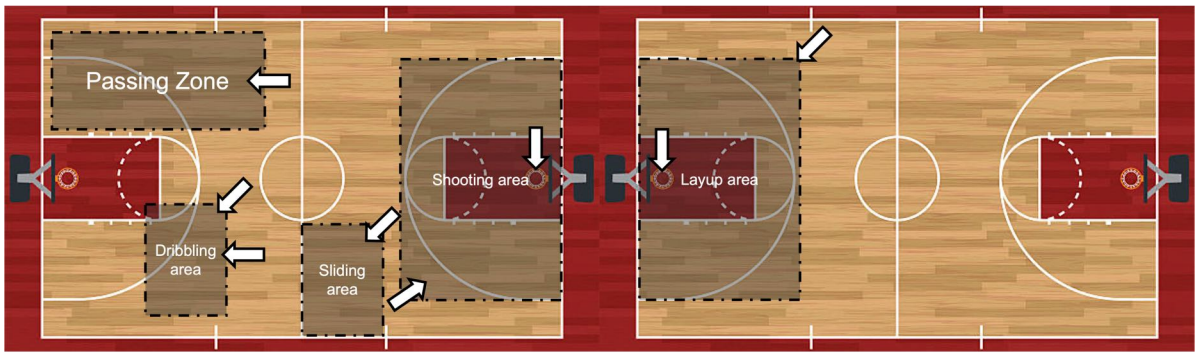

Figure 6.5: Data collection process for a youth level three basketball exam, consisting of 9 items across 5 areas on 2 courts. Each item is recorded from different angles using 2-3 standard cameras.

the accuracy of our predictions but also ensures that critical moments and actions are adequately documented.

Additionally, this dataset includes the manual scoring results from professional examiners. These scores serve as a benchmark against which we compare the system's automated scoring. By comparing the system's results with the manually assigned scores, we evaluate the accuracy and reliability of our automated scoring mechanism. This comparison is crucial for validating the effectiveness of our system in real-world applications.

By leveraging this carefully curated dataset, we are able to test and validate the system under realistic conditions, reflecting the challenges and dynamics of actual basketball examinations. This real-world data is essential for assessing the effectiveness and reliability of our system, ensuring that it performs well in practical applications.

6.3.1 Exam Items and Deduct Score Rules

The exam items include: Triangle Slide Defense, Dribbling Layup, Five-point Spot Shooting, Passing and Catching the Ball, Front and Back Spin Dribble, Stationary In-front Dribble, Stationary Two-hand Dribble, Stationary Behind-the-back Dribble, and

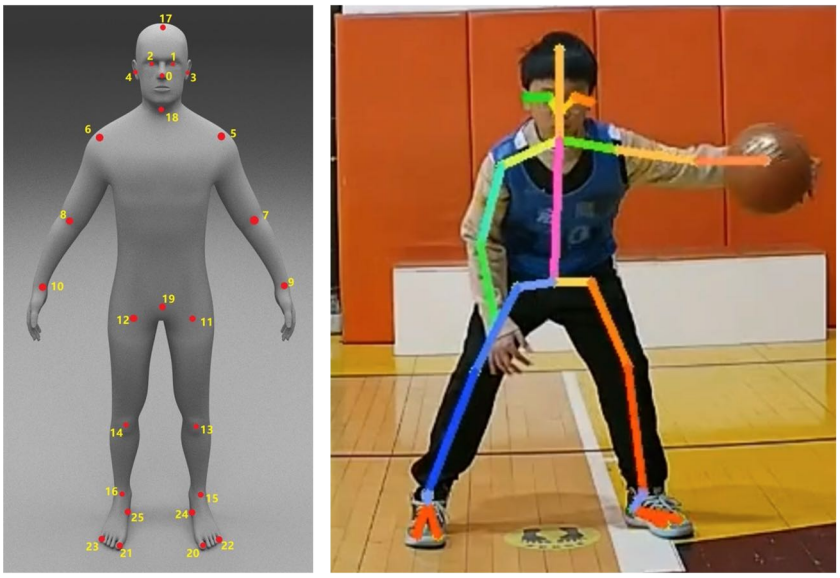

Figure 6.6: We use the Alphapose Halpe 26 keypoints Fang et al. (2022) model.

Stationary Between-the-legs Dribble. In this dataset, we encode each rule based on the deduction rules in real exam scenarios to form a system of rule modules. To track and analyze these movements, we utilize the Alphapose Halpe 26 keypoints model, as shown in Figure 6.6, ensuring accurate detection of critical points for scoring. The deduction rules of each exam item are as follows:

6.3.1.1 Triangle Slide Defense.

As shown in Figure 4.2, the Triangle Slide Defense includes **14** deduction points:

1. Lowering the head in the basic defensive stance: Identified when the angle formed by points 17-18-19 in the skeleton recognition model is less than 130 degrees.

2. Failing to open arms in the basic defensive stance: Identified when the angle under the armpit is less than 30 degrees.

3. Incorrect knee angle in the basic defensive stance: Recognized when the left knee angle is greater than 170 degrees or the right knee angle is greater than 140 degrees.

4. Feet not forming an outward 'V' shape in the basic defensive stance: Detected

when the distance between points 15 and 16 is more than 1.3 times the distance between points 20 and 21.

5. Incorrect number of slides in the triangle slide defense: Deducted if the touch sequence 0-1-2-0-2-1-0 is not completed.

6. Incorrect sliding path: Deducted if the touch sequence does not follow the required order within the seven slides.

7. Excessive vertical movement while sliding: Identified when the vertical head movement exceeds 50 units between frames.

8. Lowering the head while sliding: Detected similarly to the basic stance using the angle of points 17-18-19.

9. Feet not forming an outward 'V' shape and touching each other during the horizontal slide: Identified using the same criteria for the 'V' shape and if the distance between points 24 and 25 is less than 30 units.

10. Failing to raise hands to interfere during the horizontal slide: Identified if the arms are not opened as specified.

11. Feet not forming a 'T' shape and touching each other during the upward slide: Recognized when the foot angle is less than 30 degrees with the heel as the vertex.

12. Toes not pointing in the direction of the backward step: Identified if the toe direction angle relative to the line connecting points 2 and 0 is greater than 45 degrees.

13. Failing to raise hands to interfere during the upward slide.

14. Failing to raise hands to interfere during the backward slide.

6.3.1.2 Dribbling Layup.

As shown in Figure 4.3, the dribbling layup includes **9** deduction points:

1. Missed layup: Judged as a layup action, but the ball did not go in.

2. Missed shot: Judged as a shot action, but the ball did not go in.

3. Layup violation - carrying: Judged at the moment of the layup, tracing back 10 frames after bypassing the cone. First, determine the dribbling hand; if the ball is above the hand and both the hand and ball are above the waist, it is considered a carry.

4. Layup violation - traveling: Judged at the moment of the layup, tracing back 10 frames after bypassing the cone. Count the number of dribbles and steps (looking for peaks in the distance between the feet for each step taken). Compare the dribble times with the steps; if there are three steps between two dribbles, it triggers a traveling violation.

5. Layup with incorrect hand: Judged at the moment of the layup, tracing back 10 frames after bypassing the cone. First, determine if the cone is the near or far one. During this process, measure the distance between the ball and each hand. For each frame, if the ball is closer to the left hand, add 1 to the left hand score; if closer to the right hand, add 1 to the right hand score. If the dribbling hand is not the outside hand, it triggers a violation.

6. Incorrect takeoff foot for layup: Judged at the moment of the layup, tracing back 10 frames after bypassing the cone. First, determine if the cone is the near or far one. If bypassing the far cone and the left knee is lower than the right at the moment of the layup, it triggers a violation. If bypassing the near cone and the right knee is lower than the left, it triggers a violation.

7. Layup with incorrect hand: Judged at the moment of the layup, tracing back 10 frames after bypassing the cone. First, determine if the cone is the near or far one. If bypassing the far cone and the left hand is lower than the right at the moment of the layup, it triggers a violation. If bypassing the near cone and the right hand is lower than

the left, it triggers a violation.

8. Failing to bypass the cone from the outside: Draw the candidate's X-axis chart to find the valley, indicating the turning point. Determine whether the candidate is to the left or right of the cone. If to the right, it triggers a violation for not bypassing the cone from the outside.

9. Excessive distance from the cone: After determining the candidate has bypassed the cone from the outside, check the difference between the candidate's x-coordinate and the cone's x-coordinate. If the difference exceeds 80, it is considered too far.

6.3.1.3 Five-point Spot Shooting.

As shown in Figure 4.4, the dribbling layup includes **4** deduction points:

1. Incorrect foot position in standard shooting stance: When the candidate finishes shooting from position N, but the next shot is not from position N+1, it triggers a violation.

2. Incorrect knee angle in standard shooting stance: If the candidate's knee angle does not drop below 175 degrees from arriving at each spot to initiating the shooting action, it triggers a violation. This parameter can be custom-adjusted in the configuration file.

3. No wrist flexion upon release: Due to hand recognition issues, accurate judgment is difficult. Currently, if the predicted angle between the hand and the arm exceeds 180 degrees 5 frames after the ball is released, it is considered incorrect.

4. Missed shot: The ball is shot but does not go in.

6.3.1.4 Passing and Catching the Ball.

As shown in Figure 4.5, the passing and catching the ball includes **9** deduction points:

1. Incorrect number of passes: Not satisfying the requirement of 2 aerial passes or 2

bounce passes. This parameter can be custom-adjusted in the configuration file.

2. Incorrect knee angle in basic stance: Before starting the pass, if the candidate's knee angle in the skeleton recognition model is greater than 170 degrees, it is considered incorrect. This parameter can be custom-adjusted in the configuration file.

3. Failure to shift weight forward in basic stance: Before starting the pass, if the side waist angle in the skeleton recognition model is greater than 170 degrees, it triggers a violation. This parameter can be custom-adjusted in the configuration file.

4. Elbows not abducted in basic stance: Based on predicted values, if the underarm angle is set greater than 5 degrees, it is considered incorrect. This parameter can be custom-adjusted in the configuration file.

5. Pass deviating from the correct trajectory: If at the moment of catching the ball, the horizontal distance between the catcher's feet and the passer's feet exceeds 70 units, it is considered that the ball's trajectory is incorrect. This parameter can be custom-adjusted in the configuration file.

6. Arms not extended to receive the ball: If the angle of the receiver's elbow joint is less than 90 degrees when catching the ball, it triggers a violation. This parameter can be custom-adjusted in the configuration file.

7. Failure to cushion the ball to the chest or abdomen after touching the ball: Tracing back to the last catch at the moment of passing the ball. If the distance between the ball's center and any point on the hip joint is consistently greater than 100 units, it is considered that the ball was not cushioned to the chest or abdomen. This parameter can be custom-adjusted in the configuration file.

8. Failure to return to basic stance after catching the ball: Tracing back to the last catch at the moment of passing the ball. If the basic stance requirements (knee angle and forward weight shift) are not met, it triggers a violation. This parameter can be

custom-adjusted in the configuration file.

9. Dropping the ball after catching: Tracing back to the last catch at the moment of passing the ball. If the distance between the ball and the left heel exceeds 300 units, it is considered a dropped ball. This parameter can be custom-adjusted in the configuration file.

6.3.1.5 Front and Back Spin Dribble, Stationary In-front Dribble, Stationary Two-hand Dribble, Stationary Behind-the-back Dribble, and Stationary Between-the-legs Dribble.

As shown in Figures 4.6, 4.7, 4.8, 4.9, and 4.10, these items each include **9** deduction points:

1. Incorrect dribbling action: Triggered when the corresponding action is performed incorrectly.

2. Incorrect number of dribbles: Triggered when the required number of dribbles is insufficient.

3. Basic stance holding the ball - feet not shoulder-width apart: If the shoulder width is 3.5 times greater than the distance between the heels, it is considered incorrect. This parameter can be custom-adjusted in the configuration file.

4. Basic stance holding the ball - incorrect hip angle: If the hip angle is greater than 175 degrees, it is considered incorrect. This parameter can be custom-adjusted in the configuration file.

5. Basic stance holding the ball - incorrect knee angle: If the left knee angle is greater than 175 degrees or the right knee angle is greater than 170 degrees, it is considered incorrect. This parameter can be custom-adjusted in the configuration file.

6. Basic stance holding the ball - elbows not abducted: If the underarm angle is less

than 5 degrees while holding the ball, it is considered incorrect. This parameter can be custom-adjusted in the configuration file.

7. Dribbling with head down: If the angle formed by points 17-18-19 in the skeleton recognition model is less than 130 degrees, it is considered head down. This parameter can be custom-adjusted in the configuration file.

8. Dribbling violation - carrying: If the ball appears above the waist and the distance between the hand and the ball's center is less than the ball's radius plus 10 units, with the ball above either hand, it is considered carrying. This parameter can be custom-adjusted in the configuration file.

9. Dropping the ball: If the distance between the ball and the left foot exceeds 300 units, it is considered a dropped ball. This parameter takes into account the dynamic dribbling distance and is only triggered for significant deviations. This parameter can be custom-adjusted in the configuration file.

6.4 Experiments and Evaluation

Our experiments are conducted on an NVIDIA GeForce RTX 3090 GPU with 24GB of memory. The computational tasks are set up to run in parallel, with a maximum of four tasks running simultaneously. When the number of tasks exceeds four, they are queued and processed sequentially. This setup ensures efficient utilization of the GPU's capabilities while managing resource constraints effectively. Compared to the time taken for manual scoring, our program is on average more than **five times** faster.

The results of our automated scoring system are compared against the manual scoring results provided by professional examiners. The comparative analysis is presented in Table 6.1. The table illustrates the correlation and discrepancies between the automated and manual scores, highlighting the accuracy and reliability of our system.

Action	**Accuracy**			
	Y+BoT+H26	**CT+H26**	**Y+BoT+KE**	**CT+KE (Ours)**
Triangle Slide Defense (14)	0.81	0.81	0.88	0.95
Five-point Spot Shooting (4)	0.40	0.78	0.57	0.90
Passing and Catching the Ball (9)	0.66	0.85	0.69	0.92
Front and Back Spin Dribble (9)	0.46	0.81	0.59	0.88
Stationary In-front Dribble (9)	0.63	0.74	0.70	0.91
Stationary Two-hand Dribble (9)	0.85	0.85	0.85	0.95
Stationary Behind-the-back Dribble (9)	0.60	0.77	0.57	0.87
Stationary Between-the-legs Dribble (9)	0.55	0.79	0.51	0.89
Dribbling Layup (9)	0.49	0.73	0.63	0.80

Table 6.1: Accuracy Comparison Across Different Methods in Basketball Skill Evaluation. Evaluated on 810 videos from 30 youth testers aged 8-14, each performing 9 testing items. Methods compared include: Y+BoT+H26 (YoloV8+BoT-SORT+Halpe26), CT+H26 (Context Track+Halpe26), Y+BoT+KE (YoloV8+BoT-SORT+Key Point Enhance), and CT+KE (Context Track+Key Point Enhance (Ours)). The numbers in parentheses indicate the number of deduction items for each action.

6.5 Conclusions

In this chapter, we design and implement an intelligent sports examination system that is fair, objective, and efficient. The system is built using a modular design, ensuring ease of expansion and adaptability to various sports examination scenarios. By integrating advanced computer vision and machine learning techniques, we address key challenges in sports assessments, such as interference-resistant tracking, accurate classification of motion sequences, and error point identification.

Our system is rigorously tested using data from real-world scenarios, specifically a youth level three basketball examination in Shanghai, China. The dataset includes comprehensive video recordings and manual scoring results, providing a robust basis for evaluating the system's performance. The Rule Module, tailored to the specific requirements of the basketball examination, aligns the skeletal and object tracking sequences to evaluate each frame, ensuring precise scoring and identification of violations.

We conduct experiments on an NVIDIA GeForce RTX 3090 GPU with 24GB of memory, utilizing parallel task execution to enhance processing efficiency. The results, compared against manual scoring, demonstrate the system's high accuracy and significant time savings, processing each video in approximately half its duration.

By visualizing the scoring results and motion data directly on the video frames, the system provides clear and actionable feedback to users. This visualization includes overlaying scores, penalties, skeletal information, and object tracking data, making the assessment results easily interpretable.

The development and testing of this intelligent sports examination system highlight its practical applicability and potential for real-world deployment. The system's modular design ensures it can be extended and adapted to other sports disciplines, addressing the diverse requirements and challenges of various examination scenarios.

Overall, this chapter demonstrates the feasibility and effectiveness of a comprehensive, modular intelligent sports examination system. By connecting theoretical advancements with practical implementation, we offer a robust solution for objective sports assessments, laying the foundation for further research and development in this area.

CHAPTER 7

CONCLUSION AND FUTURE RESEARCH

This thesis addresses several significant challenges in the realm of video-based intelligent sports examination systems. By using basketball skill assessments as a primary example, this research illustrates the broader applicability of our solutions to various sports disciplines. The primary contributions of this research are organized as follows:

Effective Human Skeleton Keypoint Completion. We proposed a robust method for keypoint prediction enhancement by integrating STGNP with existing human pose estimation methods. This approach addresses the issue of occluded keypoints by robustly inferring and supplementing missing keypoints and providing precise uncertainty estimates. It also corrects anomalies in predicted keypoints, such as sudden coordinate changes, ensuring more accurate and reliable predictions.

Advanced Object Tracking in Specific Scenarios. Our development of advanced object tracking algorithms incorporates positional information to enhance accuracy amidst frequent occlusions and complex interactions. This ensures precise tracking of examination props by integrating the spatial context of candidates and incorporating body information, effectively addressing the challenges posed by traditional object tracking

methods.

Accelerated Computational Processing. We introduced a novel method for efficient computational processing that leverages Lie Algebra to represent human skeletal structures, significantly reducing computational overhead. In conjunction with this, our use of Memristor-Augmented LSTM and CNN technologies further accelerates computation and decreases the power requirements. This integrated approach not only achieves faster inference but also ensures suitability for real-time applications in resource-constrained environments. Additionally, it involves embedding trained network weights into a memristor-based structure, opening new avenues for enhanced computational acceleration.

Modular and Practical Intelligent Examination System. We designed a comprehensive modular intelligent sports examination system, using basketball skill assessments as a primary example. The system integrates robust keypoint prediction, accurate object tracking, and efficient computational processing technologies into a cohesive unit. We evaluated the system's performance in real-world basketball skill assessments, providing insights into its effectiveness and robustness. Additionally, we ensured the scalability and adaptability of the system, facilitating its application to different sports and dynamic environments. Cost, ease of use, and integration with existing sports examination frameworks were also considered to ensure the system's real-world applicability and scalability.

7.1 Research Significance

The theoretical and practical significance of this thesis are profound and multifaceted:

Theoretical Significance: This research provides a nuanced and standardized framework for identifying and addressing the challenges inherent in video-based intel-

ligent sports examination systems, particularly in dynamic and complex sports environments. Our innovative methods for enhancing human keypoint prediction through STGNP, advanced object tracking, and the incorporation of Lie Algebra for efficient representation of human skeletal structures address significant theoretical gaps. Additionally, integrating Memristor-Augmented LSTM and CNN for computational efficiency contributes to the development of energy-efficient deep learning models suitable for real-time applications in resource-constrained settings.

Practical Significance: The development of a modular, scalable intelligent sports examination system, exemplified through basketball skill assessments, has been rigorously validated in real-world settings. This system not only ensures precise tracking and rapid inference but also demonstrates substantial adaptability to various sports disciplines under different conditions. The practical implementations of our research have laid a foundation for extending these advanced methodologies to broader applications in sports and dynamic environments, proving the system's reliability for performance evaluation and skill assessment.

In conclusion, this thesis presents a comprehensive and effective strategy for exploiting the complexities of developing intelligent sports examination systems. By exploiting STGNP for robust human keypoint completion, utilizing advanced object tracking techniques, and adopting energy-efficient computational methods like Lie Algebra and Memristor-Augmented neural networks, this research significantly enhances the capabilities of intelligent systems in dynamic sports settings. Moreover, the development and real-world validation of a modular, scalable system underscore its practical utility and flexibility, establishing a new standard for future advancements in the field. These contributions not only improve the precision and efficiency of sports examination systems but also guarantee their reliable performance, laying a solid foundation for continued innovation and wider application in various sporting and dynamic environments.

7.2 Future Research

This thesis identifies the following directions for future research:

Advanced Tracking Algorithms with Large Language Models.

One of the future research directions is to develop more advanced tracking algorithms that leverage the understanding capabilities of large language models (LLMs). By integrating LLMs, we aim to automatically and accurately identify the correct examination props or subjects in sports examinations. This will provide a solid foundation for downstream tasks, such as performance analysis and skill assessment. The incorporation of LLMs will enhance the system's ability to interpret complex scenarios, improving tracking precision and robustness in dynamic sports environments.

Designing More Advanced Neural Network Structures with Memristors.

Future research will focus on designing memristor circuits to achieve more advanced and efficient neural network structures. While current neural network models have rapidly evolved, the application of memristors has been hindered by architectural design challenges. Addressing these challenges will enable the development of more powerful and energy-efficient solutions, enhancing the scalability and performance of intelligent sports assessment systems for real-time evaluations.

Integration with Large Language Models for Rule Generation.

Another important direction is the integration of large language models to assist in writing downstream decision rules for the intelligent sports examination system. By utilizing the knowledge and flexibility of LLMs, we can develop more adaptable and transferable examination systems that can be easily migrated to different sports or assessment tasks. This integration will simplify the process of customizing the examination system for various sports disciplines, ensuring that the system remains versatile and effective across diverse applications.

References

Jake K Aggarwal and Michael S Ryoo.
Human activity analysis: A review.
Acm Computing Surveys (Csur), 43(3):1–43, 2011.

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky.
Bot-sort: Robust associations multi-pedestrian tracking.
arXiv preprint arXiv:2206.14651, 2022.

Emre Aksan and Otmar Hilliges.
STCN: stochastic temporal convolutional networks.
In *International Conference on Learning Representations, ICLR*, 2019.

Gabriel Appleby, Linfeng Liu, and Li-Ping Liu.
Kriging convolutional networks.
In *The Thirty-Fourth Conference on Artificial Intelligence AAAI*, volume 34, pages 3187–3194, 2020.

Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua.
Multiple object tracking using k-shortest paths optimization.
IEEE transactions on pattern analysis and machine intelligence, 33(9):1806–1819, 2011.

Christopher M Bishop and Nasser M Nasrabadi.
Pattern recognition and machine learning.
Springer, 2006.

Faisal Budiman, Detiza Goldianto Octensi Hernowo, Reetu Raj Pandey, Hirofumi Tanaka, et al.
Recent progress on fabrication of memristor and transistor-based neuromorphic devices for high signal processing speed with low power consumption.
Japanese Journal of Applied Physics, 57(3S2):03EA06, 2018.

Delong Cai, Zhaoyun Zhang, and Zhi Zhang.
Corner-point and foreground-area iou loss: Better localization of small objects in bounding box regression.
Sensors, 23(10):4961, 2023.

Wei Cao, Xiaoyong Wang, Xianxiang Liu, and Yishuai Xu.
A deep learning framework for multi-object tracking in team sports videos.
IET Computer Vision, 2024.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh.
Openpose: Realtime multi-person 2d pose estimation using part affinity fields.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh.
Realtime multi-person 2d pose estimation using part affinity fields.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko.
End-to-end object detection with transformers.
In *European conference on computer vision*, pages 213–229. Springer, 2020.

Joao Carreira and Andrew Zisserman.
Quo vadis, action recognition? a new model and the kinetics dataset.
In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu.
Transformer tracking.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021.

Shuai Cheng, Yonggang Cao, Junxi Sun, and Guangwen Liu.
Visual tracking with online incremental deep learning and particle filter.
International Journal of Signal Processing, Image Processing and Pattern Recognition, 8:107–120, 2015.

Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang.

A neural attention model for urban air quality inference: Learning the weights of monitoring stations.
In *AAAI*, pages 2151–2158, 2018.

Leon Chua.
Memristor-the missing circuit element.
IEEE Transactions on circuit theory, 18(5):507–519, 1971.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio.
A recurrent latent variable model for sequential data.
Annual Conference on Neural Information Processing Systems, NeurIPS, pages 2980–2988, 2015.

MMPose Contributors.
Openmmlab pose estimation toolbox and benchmark.
https://github.com/open-mmlab/mmpose, 2020.

Corinna Cortes and Vladimir Vapnik.
Support-vector networks.
Machine learning, 20:273–297, 1995.

Menglin Cui and Yang Zhang.
Memristive synaptic circuits for deep convolutional neural networks.
In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu.
Target transformed regression for accurate tracking.
arXiv preprint arXiv:2104.00403, 2021.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu.
Mixformer: End-to-end tracking with iterative mixed attention.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022.

Navneet Dalal and Bill Triggs.
Histograms of oriented gradients for human detection.
In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.
An image is worth 16x16 words: Transformers for image recognition at scale.
arXiv preprint arXiv:2010.11929, 2020.

Yong Du, Wei Wang, and Liang Wang.
Hierarchical recurrent neural network for skeleton based action recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

Weiming Fan, Jiahui Yu, and Zhaojie Ju.
Fast object detection leveraging global feature fusion in boundary-aware convolutional networks.
Information, 15(1):53, 2024.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu.
Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani.
Person re-identification by symmetry-driven accumulation of local features.
In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2360–2367. IEEE, 2010.

Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan.
Random forests: from early developments to recent advancements.
Systems Science & Control Engineering: An Open Access Journal, 2(1):602–609, 2014.

Pedro Felzenszwalb, David McAllester, and Deva Ramanan.
A discriminatively trained, multiscale, deformable part model.
In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

Chelsea Finn, Ian Goodfellow, and Sergey Levine.
Unsupervised learning for physical interaction through video prediction.

Advances in neural information processing systems, 29, 2016.

Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard Turner.
Meta-learning stationary stochastic process prediction with convolutional neural processes.
NeurIPS, 33:8284–8295, 2020.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik.
Recurrent network models for human dynamics.
In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015.

Dr. Gajendra.
Artificial intelligence in sports.
International Journal of Future Research in Management and Research, 5(4):5657, 2023.
doi: 10.36948/ijfmr.2023.v05i04.5657.
URL https://dx.doi.org/10.36948/ijfmr.2023.v05i04.5657.

Cunzhang Gao, Haitao Gu, Siquan Yu, and Xingzhen Li.
Unifying classification and bounding box regression head for object detection.
In *Journal of Physics: Conference Series*, volume 2216, page 012106. IOP Publishing, 2022.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh.
Neural processes.
In *ICML Workshop*, 2018.

Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell.
Actionvlad: Learning spatio-temporal aggregation for action classification.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–980, 2017.

Ross Girshick.
Fast r-cnn.
In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik.
Rich feature hierarchies for accurate object detection and semantic segmentation.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner.
Convolutional conditional neural processes.
In *ICLR*, 2020.

Zhitao Guo, Linlin Zhao, Jinli Yuan, and Hengyong Yu.
Msanet: Multiscale aggregation network integrating spatial and channel information for lung nodule detection.
IEEE Journal of Biomedical and Health Informatics, 26(6):2547–2558, 2021.

Ziqi Guo, Chu He, Lian Zhou, Qingyi Zhang, and Shilei Sun.
Robust bounding box regression for small object detection.
In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2290–2294. IEEE, 2023.

Will Hamilton, Zhitao Ying, and Jure Leskovec.
Inductive representation learning on large graphs.
Advances in neural information processing systems, 30, 2017.

Qilong Han, Dan Lu, and Rui Chen.
Fine-grained air quality inference via multi-channel attention model.
In *the International Joint Conference on Artificial Intelligence, IJCAI*, pages 2512–2518, 2021.

Raqibul Hasan, Tarek M Taha, and Chris Yakopcic.
On-chip training of memristor crossbar based multi-layer neural networks.
Microelectronics journal, 66:31–40, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick.

Mask r-cnn.
In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick.
Masked autoencoders are scalable vision learners.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista.
High-speed tracking with kernelized correlation filters.
IEEE transactions on pattern analysis and machine intelligence, 37(3):583–596, 2014.

Hung-Min Hsu, Tsung-Wei Huang, Gaoang Wang, Jiarui Cai, Zhichao Lei, and Jenq-Neng Hwang.
Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models.
In *CVPR workshops*, pages 416–424, 2019.

Junfeng Hu, Zhencheng Fan, Jun Liao, and Li Liu.
Predicting long-term skeletal motions by a spatio-temporal hierarchical recurrent network.
arXiv preprint arXiv:1911.02404, 2019.

Junfeng Hu, Yuxuan Liang, Zhencheng Fan, Yifang Yin, Ying Zhang, and Roger Zimmermann.
Decoupling long-and short-term patterns in spatiotemporal inference.
arXiv preprint arXiv:2109.09506, 2021.

Junfeng Hu, Yuxuan Liang, Zhencheng Fan, Hongyang Chen, Yu Zheng, and Roger Zimmermann.
Graph neural processes for spatio-temporal extrapolation.
In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 752–763, 2023.

Anping Huang, Xinjiang Zhang, Runmiao Li, and Yu Chi.
Memristor neural network design.
Memristor and Memristive Neural Networks, pages 1–35, 2018.

Hsiang-Wei Huang, Cheng-Yen Yang, Jiacheng Sun, Pyong-Kun Kim, Kwang-Ju Kim, Kyoungoh Lee, Chung-I Huang, and Jenq-Neng Hwang.
Iterative scale-up expansioniou and deep features association for multi-object tracking in sports.
In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 163–172, 2024.

Qinghua Huang, Lizhi Jia, Guanqing Ren, Xiaoyi Wang, and Chunying Liu.
Extraction of vascular wall in carotid ultrasound via a novel boundary-delineation network.
Engineering Applications of Artificial Intelligence, 121:106069, 2023.
ISSN 0952-1976.
doi: https://doi.org/10.1016/j.engappai.2023.106069.
URL https://www.sciencedirect.com/science/article/pii/S0952197623002531.

Yu-Chuan Huang, I-No Liao, Ching-Hsuan Chen, Tsì-Uí İk, and Wen-Chih Peng.
Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications.
In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

Yuhua Huang.
The role of artificial intelligence technology in promoting the development of my country's sports industry.
In *2nd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2022)*, volume 12348, pages 226–230. SPIE, 2022.

Sergey Ioffe and Christian Szegedy.
Batch normalization: Accelerating deep network training by reducing internal covariate shift.
In *International conference on machine learning*, pages 448–456. pmlr, 2015.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu.
Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments.
IEEE transactions on pattern analysis and machine intelligence, 36(7):1325–1339, 2013.

Marina Ivasic-Kos, Kristina Host, and Miran Pobar.
Application of deep learning methods for detection and tracking of players.
Deep Learning Applications, 2021.

Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena.
Structural-rnn: Deep learning on spatio-temporal graphs.
In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016.

Jianfeng Jiang and Xiaojing Zhang.
Research on moving object tracking technology of sports video based on deep learning algorithm.
In *2021 4th International Conference on Information Systems and Computer Aided Education*, pages 2376–2380, 2021.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo.
Whole-body human pose estimation in the wild.
In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020.

Sung Hyun Jo, Kuk-Hwan Kim, and Wei Lu.
High-density crossbar arrays based on a si memristive system.
Nano letters, 9(2):870–874, 2009.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al.
ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation.
Zenodo, 2022.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh.
Attentive neural processes.
In *ICLR*, 2019.

Mingyu Kim, Kyeongryeol Go, and Se-Young Yun.
Neural processes with stochastic attention: Paying more attention to the context dataset.
In *ICLR*, 2022.

Diederik P. Kingma and Max Welling.
Auto-encoding variational bayes.
In *ICLR*, 2014.

Koung-Suk Ko, Woo-Jin Ahn, Geon-Hee Kim, Myo-Taeg Lim, Tae-Koo Kang, and Dong-Sung Pae.
Re-identification for multi-object tracking using triplet loss.
In *2021 International Conference on Information Networking (ICOIN)*, pages 525–527. IEEE, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton.
Imagenet classification with deep convolutional neural networks.
Advances in neural information processing systems, 25, 2012.

Corey Lammie, Wei Xiang, Bernabé Linares-Barranco, and Mostafa Rahimi Azghadi.
Memtorch: An open-source simulation framework for memristive deep learning systems.
Neurocomputing, 485:124–133, 2022.

Tuan Anh Le, Hyunjik Kim, Marta Garnelo, Dan Rosenbaum, Jonathan Schwarz, and Yee Whye Teh.
Empirical evaluation of neural process objectives.
In *NeurIPS workshop on Bayesian Deep Learning*, page 71, 2018.

Chuankun Li, Yonghong Hou, Pichao Wang, and Wanqing Li.
Joint distance maps based action recognition with convolutional neural networks.
IEEE Signal Processing Letters, 24(5):624–628, 2017a.

Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li.
Skeleton-based action recognition using lstm and cnn.
In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 585–590. IEEE, 2017b.

Jia Wen Li, Shovan Barma, Peng Un Mak, Fei Chen, Cheng Li, Ming Tao Li, Mang I Vai, and Sio Hang Pun.
Single-channel selection for eeg-based emotion recognition using brain rhythm sequencing.
IEEE Journal of Biomedical and Health Informatics, 26(6):2493–2503, 2022.

Jun Li, Chong Xie, Sizheng Wu, and Yawei Ren.
Uav-yolov5: A swin-transformer-enabled small object detection model for long-range uav images.
Annals of Data Science, pages 1–30, 2024.

Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu.
Pose recognition with cascade transformers.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1944–1953, 2021.

Naiqi Li, Wenjie Li, Jifeng Sun, Yinghua Gao, Yong Jiang, and Shu-Tao Xia.
Stochastic deep gaussian processes over graphs.
NeurIPS, 33:5875–5886, 2020a.

Wei Li, Xiatian Zhu, and Shaogang Gong.
Harmonious attention network for person re-identification.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

Yongjun Li, Shasha Li, Haohao Du, Lijia Chen, Dongming Zhang, and Yao Li.
Yolo-acn: Focusing on small target and occluded object detection.
IEEE access, 8:227288–227303, 2020b.

Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li.
Person re-identification by local maximal occurrence representation and metric learning.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

Haitao Lin, Zhangyang Gao, Yongjie Xu, Lirong Wu, Ling Li, and Stan Z. Li.
Conditional local convolution for spatio-temporal meteorological forecasting.
In *The Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 7470–7478, 2022.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie.
Feature pyramid networks for object detection.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Jianbo Liu, Ying Wang, Yongcheng Liu, Shiming Xiang, and Chunhong Pan.
3d posturenet: A unified framework for skeleton-based posture recognition.
Pattern Recognition Letters, 140:143–149, 2020a.

Li Liu, Li Cheng, Ye Liu, Yongpo Jia, and David S Rosenblum.
Recognizing complex activities by a probabilistic interval-based model.
In *Thirtieth AAAI conference on artificial intelligence*, 2016a.

Peng Liu, Zhigang Zeng, and Jun Wang.
Multistability of recurrent neural networks with nonmonotonic activation functions and mixed time delays.
IEEE Transactions on Systems, Man, and Cybernetics: Systems, 46(4):512–523, 2015.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg.
Ssd: Single shot multibox detector.
In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016b.

Xiaoyang Liu, Zhigang Zeng, and Donald C Wunsch II.
Memristor-based lstm network with in situ training and its applications.
Neural Networks, 131:300–311, 2020b.

Jaakko Luttinen and Alexander Ilin.
Efficient gaussian process inference for short-scale spatio-temporal modeling.
In *Artificial Intelligence and Statistics*, pages 741–750. PMLR, 2012.

Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool.
Learning target candidate association to keep track of what not to track.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13444–13454, 2021.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt.
Vnect: Real-time 3d human pose estimation with a single rgb camera.
Acm transactions on graphics (tog), 36(4):1–14, 2017.

Sakorn Mekruksavanich and Anuchit Jitpattanakul.

Multimodal wearable sensing for sport-related activity recognition using deep learning networks.
Journal of Advances in Information Technology, 13, 2022.

Henrique Morimitsu, Isabelle Bloch, and Roberto M Cesar-Jr.
Exploring structure for long-term tracking of multiple objects in sports videos.
Computer Vision and Image Understanding, 159:89–104, 2017.

Hyeonseob Nam and Bohyung Han.
Learning multi-domain convolutional neural networks for visual tracking.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4293–4302, 2016.

Alejandro Newell, Kaiyu Yang, and Jia Deng.
Stacked hourglass networks for human pose estimation.
In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

Junhyug Noh, Wonho Bae, Wonhee Lee, Jinhwan Seo, and Gunhee Kim.
Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9725–9734, 2019.

Bernt Øksendal.
Stochastic differential equations.
Stochastic differential equations, pages 65–84, 2003.

Chongkeun Paik and Hyunwoo J Kim.
Improving object detection, multi-object tracking, and re-identification for disaster response drones.
arXiv preprint arXiv:2201.01494, 2022.

George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy.
Towards accurate multi-person pose estimation in the wild.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4903–4911, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al.
Pytorch: An imperative style, high-performance deep learning library.
Advances in neural information processing systems, 32, 2019.

Zeel B. Patel, Palak Purohit, Harsh M. Patel, Shivam Sahni, and Nipun Batra.
Accurate and scalable gaussian processes for fine-grained air quality inference.
In *AAAI*, pages 12080–12088, 2022.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis.
Coarse-to-fine volumetric prediction for single-image 3d human pose.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

Ronald Poppe.
Vision-based human motion analysis: An overview.
Computer vision and image understanding, 108(1-2):4–18, 2007.

Shenghao Qin, Jiacheng Zhu, Jimmy Qin, Wenshuo Wang, and Ding Zhao.
Recurrent attentive neural process for sequential data.
NeurIPS Workshop, 2019.

Joseph Rafferty, Chris D Nugent, Jun Liu, and Liming Chen.
From activity recognition to intention recognition for assisted living within smart homes.
IEEE Transactions on Human-Machine Systems, 47(3):368–379, 2017.

Oan Gheorghe Ratiu, Dana Badau, Claudia Georgeta Carstea, Adela Badau, and Florin Paraschiv.
Artificial intelligence (ai) in sports.
In *Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases*, pages 93–97, 2010.

Joseph Redmon and Ali Farhadi.
Yolov3: An incremental improvement.
arXiv preprint arXiv:1804.02767, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi.
You only look once: Unified, real-time object detection.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun.
Faster r-cnn: Towards real-time object detection with region proposal networks.
Advances in neural information processing systems, 28, 2015.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese.
Generalized intersection over union: A metric and a loss for bounding box regression.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

Kaziwa Saleh, Sándor Szénási, and Zoltán Vámossy.
Occlusion handling in generic object detection: A review.
In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000477–000484. IEEE, 2021.

Esraa Samkari, Muhammad Arif, Manal Alghamdi, and Mohammed A Al Ghamdi.
Human pose estimation using deep learning: A systematic literature review.
Machine Learning and Knowledge Extraction, 5(4):1612–1659, 2023.

Syed Shakib Sarwar, Syed An Nazmus Saqueb, Farhan Quaiyum, and ABM Harun-Ur Rashid.
Memristor-based nonvolatile random access memory: Hybrid architecture for low power compact memory design.
IEEE Access, 1:29–34, 2013.

Jacopo Secco, Mauro Poggio, and Fernando Corinto.
Supervised neural networks with memristor binary synapses.
International Journal of Circuit Theory and Applications, 46(1):221–233, 2018.

Matthias Seeger.
Gaussian processes for machine learning.
International journal of neural systems, 14(02):69–106, 2004.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang.
Ntu rgb+ d: A large scale dataset for 3d human activity analysis.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake.
Real-time human pose recognition in parts from single depth images.
In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

Divya Singh and Rajeev Srivastava.
An end to end trained hybrid cnn model for multi-object tracking.
Multimedia Tools and Applications, 81(29):42209–42221, 2022.

Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn.
Sequential neural processes.
In *NeurIPS*, pages 10254–10264, 2019.

Kamilya Smagulova and Alex Pappachen James.
A survey on lstm memristive neural network architectures and applications.
The European Physical Journal Special Topics, 228(10):2313–2324, 2019.

DC Suman.
Artificial intelligence in sport: An ethical issue.
Unity Journal, 3(01):27–39, 2022.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang.
Deep high-resolution representation learning for human pose estimation.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

Yi Sun, Zhe Zhang, Ioannis Kakkos, George K Matsopoulos, Jingjia Yuan, John Suckling, Luoyi Xu, Shuxia Cao, Wenjuan Chen, Xingyue Hu, et al.
Inferring the individual psychopathologic deficits with structural connectivity in a longitudinal cohort of schizophrenia.
IEEE Journal of Biomedical and Health Informatics, 2022.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang.
Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline).
In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

Tarek M Taha, Raqibul Hasan, Chris Yakopcic, and Mark R McLean.

Exploring the design space of specialized multicore neural processors.
In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2013.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon.
Csdi: Conditional score-based diffusion models for probabilistic time series imputation.
Advances in Neural Information Processing Systems, 34:24804–24816, 2021.

Tamilvizhi Thanarajan, Youseef Alotaibi, Surendran Rajendran, and Krishnaraj Nagappan.
Improved wolf swarm optimization with deep-learning-based movement analysis and self-regulated human activity recognition.
AIMS Mathematics, 8(5):12520–12539, 2023.

Z Tian, C Shen, H Chen, and T He.
Fcos: Fully convolutional one-stage object detection. arxiv 2019.
arXiv preprint arXiv:1904.01355, 1904.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri.
Learning spatiotemporal features with 3d convolutional networks.
In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu.
Wavenet: A generative model for raw audio.
In *The Speech Synthesis Workshop, ISCA*, page 125, 2016.

Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa.
Human action recognition by representing 3d skeletons as points in a lie group.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014.

Michael Volpp, Fabian Flürenbrock, Lukas Grossberger, Christian Daniel, and Gerhard Neumann.
Bayesian context aggregation for neural processes.
In *ICLR*, 2020.

Bin Wang, Jie Lu, Zheng Yan, Huaishao Luo, Tianrui Li, Yu Zheng, and Guangquan Zhang.

Deep uncertainty quantification: A machine learning approach for weather forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2087–2095, 2019.

Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou.
Learning discriminative features with multiple granularities for person re-identification.
In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018a.

Peng Wang, Jun Wen, Chenyang Si, Yuntao Qian, and Liang Wang.
Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition.
IEEE Transactions on Image Processing, 31:6224–6238, 2022a.

Pichao Wang, Wanqing Li, Chuankun Li, and Yonghong Hou.
Action recognition based on joint trajectory maps with convolutional neural networks.
Knowledge-Based Systems, 158:43–53, 2018b.

Xiangyu Wang, Chao Liu, and Laishuang Sun.
[retracted] the design of sports games under the internet of things fitness by deep reinforcement learning.
Computational Intelligence and Neuroscience, 2022(1):4623561, 2022b.

Xufei Wang and Jeongyoung Song.
Iciou: Improved loss based on complete intersection over union for bounding box regression.
IEEE Access, 9:105686–105695, 2021.

Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Roger Zimmermann, and Yuxuan Liang.
Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models.
arXiv preprint arXiv:2301.13629, 2023.

Nu Wen, Renzhong Guo, Ding Ma, Xiang Ye, and Biao He.
Aiou: Adaptive bounding box regression for accurate oriented object detection.
International Journal of Intelligent Systems, 37(1):748–769, 2022.

Shiping Wen, Rui Hu, Yin Yang, Tingwen Huang, Zhigang Zeng, and Yong-Duan Song.
Memristor-based echo state network with online least mean square.
IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(9):1787–1796, 2018a.

Shiping Wen, Shuixin Xiao, Yin Yang, Zheng Yan, Zhigang Zeng, and Tingwen Huang.
Adjusting learning rate of memristor-based multilayer neural networks via fuzzy method.
IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 38 (6):1084–1094, 2018b.

Shiping Wen, Huaqiang Wei, Yin Yang, Zhenyuan Guo, Zhigang Zeng, Tingwen Huang, and Yiran Chen.
Memristive lstm network for sentiment analysis.
IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(3):1794–1804, 2019.

Shiping Wen, Jiadong Chen, Yingcheng Wu, Zheng Yan, Yuting Cao, Yin Yang, and Tingwen Huang.
Ckfo: Convolution kernel first operated algorithm with applications in memristor-based convolutional neural network.
IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 40 (8):1640–1647, 2020.

Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun.
Inductive graph neural networks for spatiotemporal kriging.
In *AAAI*, volume 35, pages 4478–4485, 2021.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang.
Graph wavenet for deep spatial-temporal graph modeling.
In *IJCAI*, pages 1907–1913, 2019.

Andreu Girbau Xalabarder.
Sports broadcasting and multiple object tracking with deep learning methods.
PhD thesis, Universitat Politècnica de Catalunya (UPC), 2021.

Su Xianguo and Wang Cong.
Research on the application of artificial intelligence technology in physical training.
In *2021 2nd International Conference on Big Data and Informatization Education (ICBDIE)*, pages 261–264. IEEE, 2021.

Bin Xiao, Haiping Wu, and Yichen Wei.
Simple baselines for human pose estimation and tracking.
In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

Fei Xie, Chunyu Wang, Guangting Wang, Wankou Yang, and Wenjun Zeng.
Learning tracking representations via dual-branch fully transformer networks.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2688–2697, 2021.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao.
Vitpose: Simple vision transformer baselines for human pose estimation.
Advances in Neural Information Processing Systems, 35:38571–38584, 2022.

Xuan Xuan and Hui Xu.
Complex sports target tracking with machine learning: Take basketball as an example.
Mathematical Problems in Engineering, 2022, 2022.

Chris Yakopcic, Md Zahangir Alom, and Tarek M Taha.
Memristor crossbar deep network implementation based on a convolutional neural network.
In *2016 International joint conference on neural networks (IJCNN)*, pages 963–970. IEEE, 2016.

Chris Yakopcic, Md Zahangir Alom, and Tarek M Taha.
Extremely parallel memristor crossbar architecture for convolutional neural network implementation.
In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1696–1703. IEEE, 2017.

Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu.
Learning spatio-temporal transformer for visual tracking.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021.

Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao.
Attention driven person re-identification.
Pattern Recognition, 86:143–155, 2019.

Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang.
Transpose: Towards explainable human pose estimation by transformer.
arXiv preprint arXiv:2012.14214, 2(6), 2020.

Xiao Yang.
Memristor based neural networks: Feasibility, theories and approaches.
University of Kent (United Kingdom), 2014.

Peng Yao, Huaqiang Wu, Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, J Joshua Yang, and He Qian.
Fully hardware-implemented memristor convolutional neural network.
Nature, 577(7792):641–646, 2020.

Jaesik Yoon, Gautam Singh, and Sungjin Ahn.
Robustifying sequential neural processes.
In *ICML*, pages 10861–10870, 2020.

Bin Yu, Ming Tang, Linyu Zheng, Guibo Zhu, Jinqiao Wang, Hao Feng, Xuetao Feng, and Hanqing Lu.
High-performance discriminative tracking with transformers.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9856–9865, 2021.

Fisher Yu and Vladlen Koltun.
Multi-scale context aggregation by dilated convolutions.
In *4th International Conference on Learning Representations, ICLR*, 2016.

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon.
Temporal regularized matrix factorization for high-dimensional time series prediction.
Advances in neural information processing systems, 29, 2016.

Tianzhu Zhang, Bernard Ghanem, and Narendra Ahuja.
Robust multi-object tracking via cross-domain contextual information for sports video analysis.
In *2012 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 985–988. IEEE, 2012.

Yao Zhang, Zhiyong Chen, and Bohan Wei.

A sport athlete object tracking based on deep sort and yolo v4 in case of camera movement.
In *2020 IEEE 6th international conference on computer and communications (ICCC)*, pages 1312–1316. IEEE, 2020.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang.
Bytetrack: Multi-object tracking by associating every detection box.
In *European conference on computer vision*, pages 1–21. Springer, 2022.

Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu.
Learn to match: Automatic matching network design for visual tracking.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13339–13348, 2021.

Zhonghan Zhao, Wenhao Chai, Shengyu Hao, Wenhao Hu, Guanhong Wang, Shidong Cao, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang.
A survey of deep learning in sports applications: Perception, comprehension, and decision.
arXiv preprint arXiv:2307.03353, 2023.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah.
Deep learning-based human pose estimation: A survey.
ACM Computing Surveys, 56(1):1–37, 2023.

Zhedong Zheng, Liang Zheng, and Yi Yang.
Unlabeled samples generated by gan improve the person re-identification baseline in vitro.
In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017.

Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz.
Joint discriminative and generative learning for person re-identification.
In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019.

Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis.

Monocap: Monocular human motion capture using a cnn coupled with a geometric prior.
IEEE transactions on pattern analysis and machine intelligence, 41(4):901–914, 2018.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl.
Objects as points.
arXiv preprint arXiv:1904.07850, 2019.

Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, and Zhenyu He.
Saliency-associated object tracking.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9866–9875, 2021.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye.
Object detection in 20 years: A survey.
Proceedings of the IEEE, 111(3):257–276, 2023.