A Two-Stage Learning Approach for Goalie, Net and Stick Pose Estimation in Ice Hockey

by

Fatemeh Shahi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following paper is used in this thesis. I was co-author with major contributions to the design, analysis, writing and editing.

F. Shahi, D. Clausi, and A. Wong, "GoalieNet: A Multi-Stage Network for Joint Goalie, Equipment, and Net Pose Estimation in Ice Hockey," Women in Computer Vision Workshop (WiCV), The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

Abstract

Accurate pose estimation of ice hockey goaltenders presents a unique challenge due to the dynamic nature of the sport and the intricate interactions among the goalie, equipment, and net. This study introduces a comprehensive investigation into goalie pose estimation using both One-Stage and Two-Stage Learning GoalieNet architectures. The One-Stage Learning GoalieNet predicts all keypoints simultaneously, while the Two-Stage Learning GoalieNet employs a Keypoint Predictor Network (KPN) to predict 26 out of 29 keypoints and a Keyheatmap Fusion Network (KFN) to predict 3 stick-related keypoints. Evaluation on a NHL dataset underscores the effectiveness of both approaches in accurately predicting keypoints. Results on the test data reveal a median percentage of detected keypoints of 71% for the Two-Stage approach and 70% for the One-Stage approach, along with normalized localization errors on detected keypoints of 0.0187 for the Two-Stage and 0.0194 for the One-Stage approach. This work introduces the first-ever goalie pose estimation technique designed specifically for ice hockey, accompanied by a thorough analysis of the obtained results.

Acknowledgements

I would like to express my heartfelt gratitude to Professor David Clausi and Alexander Wong for their invaluable guidance, support, and mentorship throughout my thesis journey. Their expertise and encouragement have been instrumental in shaping the direction of this research.

I am also deeply thankful to my dear family for their unwavering love, understanding, and encouragement. Your constant support has been my driving force and motivation.

Dedication

This is dedicated to the brave and resilient people of Iran, whose courage and spirit inspire me every day.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ice hockey, a dynamic and fast-paced team sport, has captivated audiences worldwide and emerged as a compelling domain for sports analytics research. The advent of computer vision and deep learning technologies has unlocked new possibilities for exploring and understanding sports events at unprecedented levels of detail. Within the realm of ice hockey analytics, a critical yet underrepresented research subject is "goalie pose estimation" – the intricate process of precisely detecting and estimating the positions and orientations of various body parts of goaltenders within the context of gameplay, typically captured in images or videos.

Goalie pose estimation holds critical importance in the realm of sports analytics and computer vision due to its potential to provide invaluable insights into player performance, strategy analysis, and injury prevention. In ice hockey, the goalie's movements and positions directly impact the outcome of the game, making accurate pose estimation crucial for understanding their actions and reactions. By accurately determining the positions and orientations of goaltenders, teams can gain a deeper understanding of their gameplay dynamics, enabling them to strategize more effectively and make data-driven decisions. Furthermore, this information can be instrumental in identifying potential areas for improvement, enhancing player training, and developing advanced coaching techniques. Therefore, accurate goalie pose estimation has the potential to revolutionize the way ice hockey is analyzed, practiced, and played.

This thesis aims to bridge this substantial research gap by introducing and assessing "GoalieNet", a network explicitly designed to facilitate comprehensive pose estimation of goalies, equipment, and the net in ice hockey. By proposing two distinct approaches, one involving learning of various keypoint positions jointly, and the other employing a Two-

Stage Learning process, where each stage is responsible for detecting a specific subset of keypoints, this study pioneers innovative methods tailored to the complexities of goalie pose estimation within the dynamic ice hockey context.

Our endeavor seeks to establish a fundamental benchmark method for the intricate task of goalie pose estimation within the dynamic domain of ice hockey. By introducing and evaluating the innovative GoalieNet architecture, encompassing both the One-Stage Learning and Two-Stage Learning paradigms, we aspire to provide a solid framework for accurate and comprehensive pose estimation encompassing goalies, equipment, and the net. This achievement holds the potential to revolutionize sports analytics, empower computer vision applications, and open pathways to broader insights. The implications span from refining player performance analysis and strategic planning in ice hockey to advancing the broader frontiers of artificial intelligence and machine learning, ultimately enhancing our understanding of human interactions in dynamic contexts.

1.1 Automated Player Evaluation Pipeline

The sports industry has emerged as a crucial driver of economic growth, with a global market value of approximately $509 billion in 2022, according to Sports Global Market Report 2022 [1]. Because it is such a significant contributor to the revenue that clubs bring in, the player transfer market is one of the most important aspects of the sports market. Transferring players between teams and clubs enables the distribution of talent across a wider range of organizations. This gives lower-tier organizations the chance to advance their skills and compete at a higher level.

Accurate player evaluation is paramount in the player trade market, as it can directly impact revenue generation for individual athletes and contribute substantially to the overall value of the sports industry. Additionally, player assessment assists teams in designing effective tactics by providing insights into individual strengths, weaknesses, and overall team performance. Furthermore, it aids in preventing injuries by identifying potential vulnerabilities and enabling preventive actions.

The availability of enormous sports datasets, which were collected through cameras and GPS technology, has led to the widespread and fast-rising usage of machine learning and data-driven methodologies in various aspects of sports analytics, such as the assessment of player performance. The examination of vast and complicated datasets is made possible by machine learning in a manner that would not be feasible using conventional statistical approaches or human analysis. Automated player evaluation approaches with distinct and

clearly stated objectives have the potential to remove human subjectivity from the evaluation process, identify undervalued players with potential for improvement, and provide support for these players. In addition to this, they are able to recognize complex patterns that people might overlook at first glance, which ultimately results in improved predictive models for the outcomes of future games. Real-time analysis of games also empowers coaches to make well-informed decisions during games, adapting their strategies as needed. Likewise, employing machine learning for performance evaluation cut down on the amount of time spent on manual data analysis. This can be accomplished by accurately analyzing vast amounts of data in a quick and efficient manner.

In Figure 1.1, we present an overview of the necessary steps required to initiate the ice hockey player evaluation task. This pipeline represents the collaborative efforts of the Sports Analytics and Research Group (SARG) at the University of Waterloo. It has been meticulously designed to advance the field of player evaluation within the context of ice hockey. Comprising a series of interconnected tasks, this process aims to refine and revolutionize the assessment of players' skills and performance.

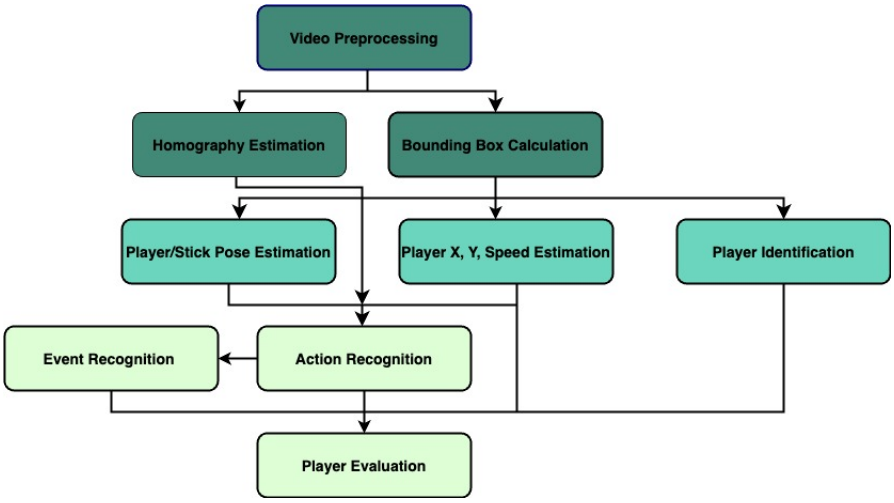

Figure 1.1: The SARG pipeline for automating player evaluation. The completed tasks consist of video preprocessing, homography estimation, and bounding box calculation. Work is currently being done on player/stick pose estimation, player X, Y, speed estimation, and player identification. In the near future, action recognition, event recognition, and player evaluation will be implemented.

The evaluation process requires comprehensive information about players' actions, events, identities, speed, coordination, and more. The initial phase involves video pre-

processing, extracting frames from videos and information related to each task. Homography estimation is performed to map the ice rink from the video frame of a hockey game to an overhead view of the hockey rink. Accurate bounding boxes are determined considering equipment and spatial consistency. Player identification relies on these bounding boxes, enabling tracking and estimation of players' coordinates, trajectories, speed, and poses. This data enhances action recognition, allowing for the classification of various game events and comprehensive player evaluations in ice hockey.

Within this broader framework, the Goalie Pose Estimation task serves as a pivotal step within this expansive pipeline. As the industry grapples with the complexities of assessing and optimizing player performance, specialized tasks like goalie pose estimation play a crucial role. This project, with its focus on accurately determining the positions and orientations of goaltenders during gameplay, contributes to the overarching objective of creating a holistic player evaluation framework. In essence, the Goalie Pose Estimation project represents one piece of the larger puzzle, where each task contributes to the broader goal of elevating player evaluation, team strategies, and overall sports industry outcomes.

1.2 Problem Statement

This thesis centers its attention on the specialized domain of ice hockey goaltender pose estimation. While often overlooked, this research area is of utmost importance. Goaltenders serve as the last line of defense in ice hockey, playing a critical role in preventing opposing teams from scoring. Their positioning, movements, and reactions during gameplay are pivotal in determining the outcome of games. Accurate pose estimation of goaltenders provides a comprehensive understanding of their on-ice behavior, enabling analysts and coaches to delve into the intricacies of their performance.

With precise pose estimation data, researchers can analyze goalie movements in various game situations, such as facing shots, making saves, and positioning for potential threats. Understanding how goalies respond to different plays and offensive strategies helps identify patterns and tendencies in their decision-making process. This knowledge can be utilized by coaches to develop strategic game plans that exploit the goalie's weaknesses or reinforce their strengths.

Moreover, accurate pose estimation allows for detailed analysis of goalie reactions to game events. Coaches can study how goalies adjust their positioning and movements in response to offensive plays, providing insights into their anticipation skills and reflexes. This information is invaluable for improving goalie training and honing their ability to read the game.

Additionally, pose estimation data enables the assessment of goalie performance under pressure. Analyzing how goalies react during critical moments, such as penalty shots or power plays, can offer a deeper understanding of their composure and effectiveness in high-stakes situations. This knowledge can aid coaches in making informed decisions during crucial moments in games.

Overall, accurate goalie pose estimation significantly impacts game dynamics and team strategies. By gaining insights into goalie movements, reactions, and decision-making, coaches and analysts can fine-tune their game plans, maximize goalie performance, and ultimately influence the outcome of ice hockey matches. It provides a competitive advantage in understanding and exploiting the complexities of the goaltender position, making it a vital aspect of sports analytics in ice hockey.

1.3 Challenges

The task of goalie pose estimation in ice hockey faces several complexities that require innovative solutions. One of the primary challenges arises from the goalie's protective gears, as depicted in Figure 1.2, which often conceals key body joints. The bulky pads, masks, and gloves can obscure crucial keypoints, making it challenging for current pose estimation algorithms like Convolutional Pose Machines [48] and Deeppose [45] to precisely locate them. This concealment introduces ambiguity and uncertainty, necessitating the development of specialized techniques to accurately infer the hidden keypoints from partial or obscured visual cues.

Figure 1.2: An overview of ice hockey goalie, equipment and the net. [32]

In addition to the concealed human keypoints, goalie equipment introduces a unique set of non-human keypoints. The distinct shapes and components of goalie gear, such as the

pads and helmet, create keypoints that are not present in typical human pose estimation datasets. These non-human keypoints are essential for capturing the goalie's equipment orientation and position, as they directly impact the goalie's ability to defend the net effectively. Incorporating these non-human keypoints into the pose estimation process requires the design of novel algorithms that can handle the mixed nature of human and non-human keypoints.

Furthermore, the dynamic nature of ice hockey adds complexity to the task of goalie pose estimation. Keypoints related to the moving net present a particular challenge, as the net's position can change rapidly during gameplay. Detecting and tracking these keypoints in real-time requires robust and efficient algorithms capable of handling occlusions, variations in lighting, and fast movements.

Addressing these challenges is crucial for achieving accurate goalie pose estimation in ice hockey. Overcoming the concealment of human keypoints under goalie gear, effectively handling non-human keypoints, and accurately tracking the moving net's keypoints will be instrumental in developing a comprehensive and reliable system for analyzing goalie performance. Innovative solutions in computer vision and deep learning are essential to tackle these complexities and pave the way for advancements in goalie pose estimation, ultimately enhancing sports analytics and understanding the intricacies of goaltending in ice hockey.

1.4 Proposed Solution

In this thesis, we aim to present a comprehensive solution that addresses the intricate challenges of precise keypoint localization for ice hockey goalie. Our proposed approach centers around the development and evaluation of advanced approach designed to enhance keypoint prediction accuracy.

1.4.1 One-Stage Learning GoalieNet

Our initial exploration led us to propose the One-Stage Learning GoalieNet, a neural network architecture tailored for efficient keypoint prediction. The One-Stage GoalieNet, an advanced network inspired by cutting-edge deep learning methodologies. GoalieNet leverages cross-stage aggregation and varying kernel sizes to enhance localization accuracy, making it well-suited for joint goalie, equipment, and net pose estimation. By synthesizing

information from multiple stages, the One-Stage GoalieNet aims to achieve robust keypoint predictions.

1.4.2 Enhancing Precision with Two-Stage Learning

Recognizing the complexity of this task, particularly in the domain of stick keypoints estimation, we endeavored to design a novel system that could enhance prediction accuracy through a more sophisticated approach, leading to the Two-Stage GoalieNet. In the initial stage, we leverage the Keypoint Predictor Network to estimate all keypoints except for the stick. In the subsequent stage, we integrate the outcomes of the first network with the original image to facilitate the estimation of stick keypoints using Keyheatmap Fusion Network. This strategic integration serves to optimize information fusion across the network, consequently advancing the precision of keypoint localization.

1.4.3 Key Elements and Performance Indicators

Assessing the efficacy of our proposed solution entails a comprehensive examination of performance indicators tailored to the detection and localization of keypoints. The core performance metrics revolve around the accuracy of the estimated keypoint coordinates. To rigorously evaluate the effectiveness of our networks, we employ the following methodologies:

- Detection Accuracy: We analyze the accuracy of keypoint detection by considering a fixed distance threshold. A predicted keypoint is considered a true positive if its Euclidean distance to the ground truth is below a certain threshold.
- Normalized Keypoint Localization Error for Detected Keypoints: We calculate the Euclidean distance between the estimated coordinates and the ground truth Keypoint for each detected keypoint. This value is then divided by the diagonal of the bounding box to ensure comparability across all frames. This metric serves as an indicator of the precision in determining the keypoint's position. Lower normalized localization error indicate superior performance.
- Keypoint Grouping and Joint-Wise Analysis: Given the complexity of body and equipment keypoints, we group related keypoints together, such as those associated with shoulders or leg pads. We calculate the mean accuracy of each group to provide a holistic view of the network's performance on various body parts and equipment

items. Furthermore, we perform a detailed joint-wise analysis to identify any peculiarities or limitations of the model's performance at specific keypoints.

1.5 Contributions

This research makes substantial contributions to the field of sports analytics and computer vision by developing a method to estimate keyoints poses for ice hockey goaltenders. The key contributions include:

- Introducing a technique for goalie pose estimation in ice hockey for the first time, addressing a notable research gap and opening avenues for improved player evaluation.
- Introducing the One-Stage Learning GoalieNet, a framework designed to jointly estimate all keypoints encompassing the goalie, equipment, and net.
- Proposing a Two-Stage Learning methodology, featuring a Keypoint Predictor Network (KPN) and Keyheatmap Fusion Network (KFN), that effectively improves metrics for most of the keypoints.
- Providing detailed quantitative analyses, showcasing the positive and negative aspects of both methods and the potential area of improvements.

1.6 Thesis Outline

The primary aim of this thesis is to develop and evaluate a novel approach to estimate keypoints coordinates for goalie, equipment, and net in ice hockey. Chapter 2 provides an overview of relevant research in pose estimation, sports analytics, and computer vision, identifying key gaps in the field of goalie pose estimation. In Chapter 3, we detail the data collection process, dataset preparation, and the design of GoalieNet, including its network architecture and training procedures. Chapter 4 presents experimental results, discussing the evaluation and performance comparison of two presented approaches of GoalieNet. Finally, Chapter 5 summarizes the contributions of this thesis and emphasizes the significance of goalie pose estimation in advancing sports analytics and computer vision research.

Chapter 2

Literature Review

Human pose estimation, the task of accurately estimating the spatial locations of key body joints in images or videos as illustrated in Figure 2.1, has gained significant attention in the field of artificial intelligence. The ability to precisely infer human poses has profound implications across various domains, including sports analytics, personalized training, and performance evaluation. By analyzing and interpreting athletes' poses during their athletic activities, pose estimation enables a deeper understanding of their movements, facilitating the optimization of training techniques, injury prevention, and the enhancement of overall athletic performance.

Figure 2.1: The goal of human pose estimation is to accurately determine the spatial coordinates of keypoints on the human body. This Figure has been adapted from Figure 1 of the paper introducing Intensive Feature Consistency Networks [40].

In the domain of sports, pose estimation assumes particular significance as it provides valuable insights into athletes' body positions, movements, and technique during sports-specific activities. Accurate pose estimation in sports scenarios offers a wide range of applications, including performance evaluation, biomechanical analysis, skill assessment, and tactical decision-making. By accurately tracking key body joints, such as limbs, torso, and head, pose estimation allows for the quantification and evaluation of various performance metrics, such as posture, joint angles, trajectory, and balance. These insights empower coaches, trainers, and athletes themselves to identify areas for improvement, refine training strategies, optimize movement patterns, and prevent injuries. Furthermore, pose estimation facilitates the development of interactive training systems, virtual coaching platforms, and sports motion analysis tools, enabling athletes to receive real-time feedback, enhance skill acquisition, and elevate their athletic capabilities.

In this chapter, we delve into the literature on human pose estimation, covering both general applications and its specific application in sports. In Section 2.1, we provide an overview of the research and advancements in general human pose estimation, focusing on various techniques and models employed in this field. We discuss the challenges associated with accurate pose estimation and the development of deep learning-based approaches. In Section 2.2, we shift our focus to pose estimation in the sports domain. Sports pose estimation presents unique challenges due to the dynamic and fast-paced nature of athletic movements. We delve into the literature that specifically addresses pose estimation in sports, emphasizing the modifications and adaptations made to existing algorithms to enhance their performance on sports-specific data.

2.1 General Pose Estimation

Human pose estimation, a challenging task in computer vision, has witnessed significant advancements through the use of deep learning techniques. This literature review explores various approaches proposed for pose estimation, highlighting their key contributions and performance.

One prevalent direction focuses on utilizing deep neural networks for pose estimation, formulated as a regression problem towards body joints. By employing a cascade of DNN regressors, these methods achieve high precision pose estimates and capitalize on recent advances in deep learning [45]. Another approach introduces Dual-Source Deep Convolutional Neural Networks (DS-CNN) to integrate both local part appearance and holistic views for accurate human pose estimation. This method combines joint detection and

localization results from image patches, demonstrating effectiveness compared to state-of-the-art methods [11].

Joint detection and pose estimation have been jointly addressed in a different strategy. A partitioning and labeling formulation of body-part hypotheses, generated using CNN-based part detectors, allows for non-maximum suppression and grouping of body parts. This approach achieves state-of-the-art results in both single-person and multi-person pose estimation [36]. Convolutional Pose Machines, shown in Figure 2.2, provide a sequential prediction framework that incorporates convolutional networks to model long-range dependencies between variables. By addressing the issue of vanishing gradients during training, this method demonstrates superior performance on benchmark datasets [48].

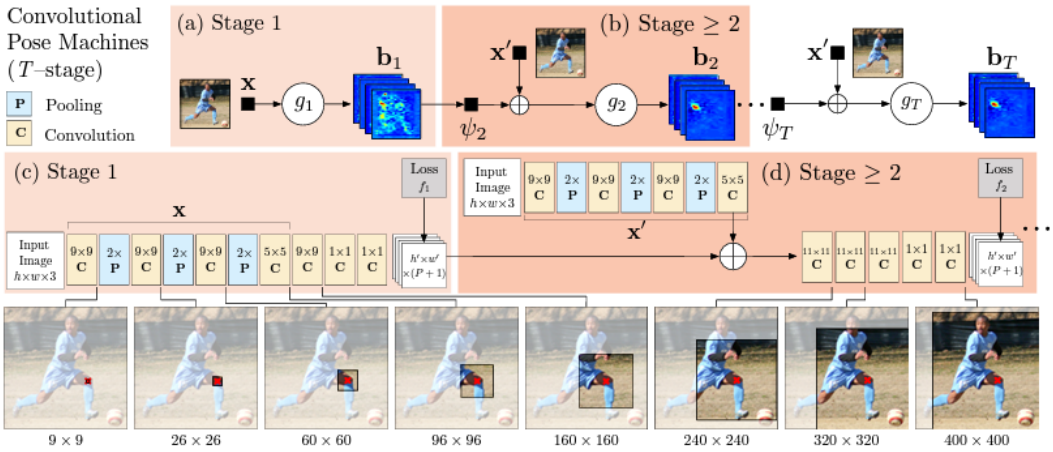

Figure 2.2: A convolutional architecture that incorporates multiple stages and expands the receptive fields across layers for a Convolutional Pose Machine (CPM) with a customizable number of stages, denoted as T, presented in Figure 2.2 in the paper "Convolutional Pose Machines" [48].

To incorporate dependencies in the output space, a framework called Iterative Error Feedback (IEF) introduces top-down feedback in hierarchical feature extractors. IEF achieves excellent performance on articulated pose estimation tasks, even without ground truth scale annotation [7]. Additionally, a cascade architecture involving detection and regression CNNs is proposed, enabling robust pose estimation even in the presence of severe part occlusions. This architecture encodes part constraints, context, and effectively copes with occlusions, achieving top performance on benchmark datasets [5].

Another architecture which is shown in Figure 2.3 referred to as the "stacked hour-

glass" network, leverages repeated bottom-up and top-down processing with intermediate supervision to capture spatial relationships associated with the body [31]. Furthermore, improved body part detectors, novel image-conditioned pairwise terms, and an incremental optimization strategy are proposed to address articulated pose estimation in scenes with multiple people. This approach achieves state-of-the-art results and demonstrates competitive performance on single person pose estimation [19].

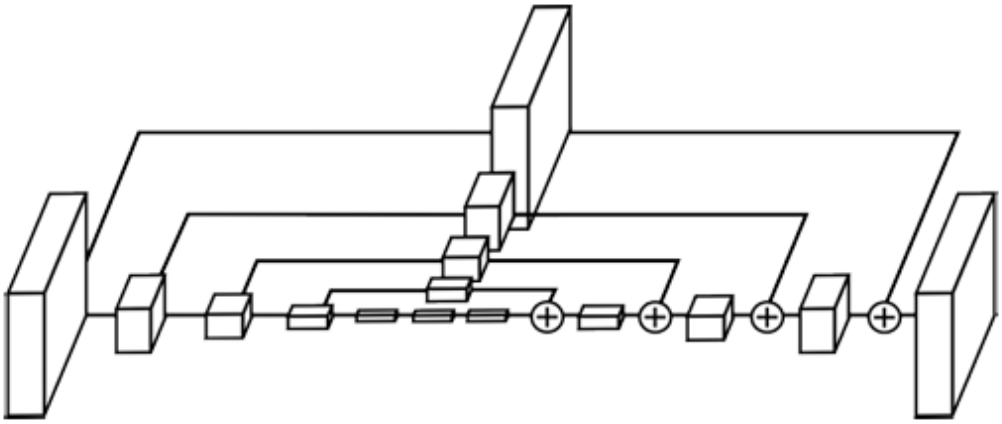

Figure 2.3: A single Hourglass module, as depicted in Figure 3 of the paper introducing Stacked Hourglass architecture [31].

For joint 2D and 3D human pose estimation, an end-to-end architecture called LCR-Net is introduced. The architecture in Figure 2.4 generates pose proposals, scores them, and refines them in both 2D and 3D. By integrating neighboring pose hypotheses, LCR-Net achieves superior performance on controlled and real image datasets [38]. Additionally, a pose refinement network (PoseRefiner) is proposed to address challenging cases and refine incorrect body joint predictions. With a novel data augmentation scheme and evaluation on popular pose estimation benchmarks, PoseRefiner demonstrates systematic improvement over the state of the art [12].

A method based on Part Affinity Fields (PAFs) achieves efficient multi-person detection and 2D pose estimation. This approach uses a nonparametric representation and encodes global context, enabling realtime performance and outperforming previous state-of-the-art results [6]. Another top-down approach employs Faster RCNN for person detection and predicts keypoints using dense heatmaps and offsets. With novel aggregation and confidence scoring techniques, this method achieves state-of-the-art performance on the COCO keypoints task [34].

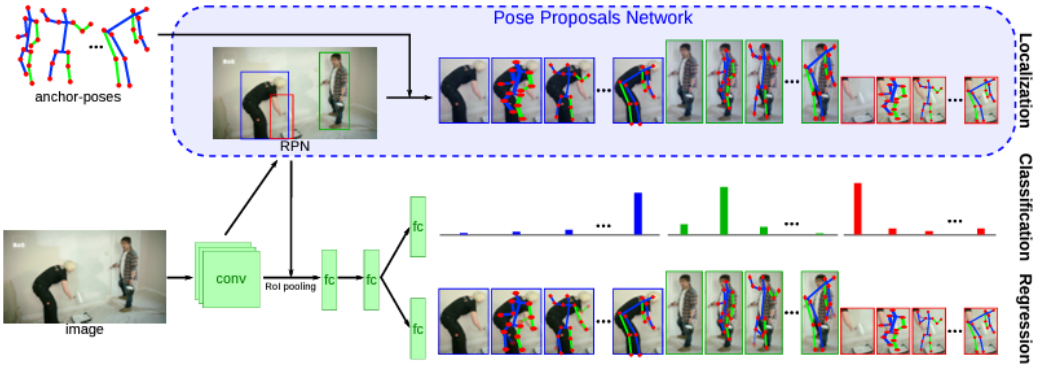

Figure 2.4: An overview of LCR-Net architecture, displayed in Figure 2 of "Localization-Classification-Regression for Human Pose" [38].

Regression-based methods have emerged as a popular approach for 2D human pose estimation. One approach, shown in Figure 2.5, formulates pose estimation as a sequence prediction task and employs a Transformer network to directly regress the keypoint coordinates from images. This end-to-end differentiable framework achieves state-of-the-art performance and outperforms heatmap-based methods [30]. Similarly, a fully end-to-end multi-person pose estimation framework, PETR, utilizes Transformers to reason about sets of full-body poses, eliminating the need for hand-crafted modules and achieving favorable accuracy and efficiency [41].

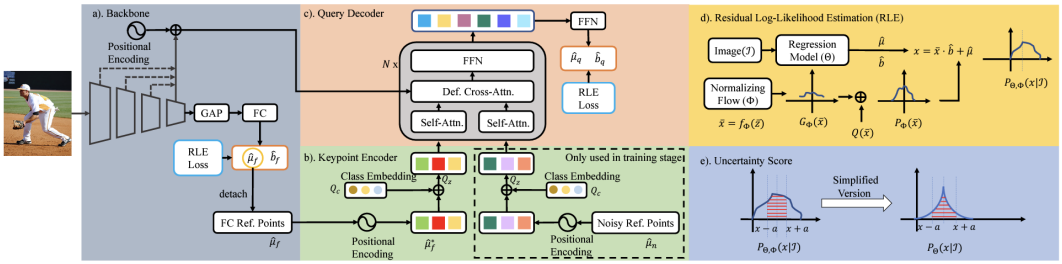

Figure 2.5: The Poseur architecture, as shown in Figure 3 of the paper introducing Poseur [30], directly acquires knowledge of mapping image inputs to the coordinates of keypoints. This approach bypasses the need for intermediate representations like heatmaps.

An alternative approach, YOLO-pose, integrates joint detection and 2D multi-person pose estimation within the YOLO object detection framework. By optimizing the Object

Keypoint Similarity (OKS) metric, YOLO-pose achieves state-of-the-art results without relying on post-processing steps like NMS and grouping [29]. Efficient architecture design is also explored, aiming to deploy pose estimation models on resource-constrained edge devices. LitePose, a single-branch architecture, achieves real-time multi-person pose estimation with reduced latency while maintaining performance [47].

BAPose introduces a bottom-up approach that leverages a disentangled multi-scale waterfall architecture and adaptive convolutions to handle occlusions and improve accuracy in crowded scenes. This efficient framework achieves state-of-the-art results on challenging datasets [3]. Additionally, a neural module is proposed to enhance fast and lightweight 2D human pose estimation CNNs by encoding global spatial and semantic information, leading to increased accuracy without sacrificing runtime computational efficiency [33].

In the quest for accurate pose estimation, limb direction cues and differentiated Cauchy labels are leveraged in the LDCNet. This network effectively suppresses uncertainties and outperforms state-of-the-art methods on benchmark datasets [26]. Furthermore, the application of Vision Transformers (ViTs) in 2D human pose estimation is explored. A method for reducing ViT's computational complexity is introduced, selecting informative patches and achieving improved speed without significant performance degradation [22].

In conclusion, the literature on general human pose estimation has witnessed a plethora of approaches aiming to accurately infer the positions of keypoints on the human body. While there are several commonalities among these approaches, such as leveraging deep learning techniques and utilizing large-scale annotated datasets, there exist notable differences in their methodologies and performance.

Many approaches rely on heatmap-based methods, which generate heatmaps indicating the likelihood of keypoints at each spatial location. These methods often employ graphical models or sequential prediction frameworks to estimate keypoints based on the heatmaps. While heatmap-based approaches have demonstrated strong performance, they tend to be computationally expensive. On the other hand, some of the reviewed approaches adopt a regression-based strategy, directly predicting the coordinates of keypoints from input images. These methods often leverage convolutional neural networks (CNNs) and exploit techniques like attention mechanisms and cascaded architectures to improve accuracy. They have shown promising results in terms of pose estimation accuracy and real-time performance.

Furthermore, variations can be observed in the specific architectural designs, data augmentation techniques, and loss functions employed by different approaches. Some models incorporate multi-scale representations, hierarchical feature extractors, or top-down feedback mechanisms to capture spatial relationships and contextual information. Others focus

on efficient architecture design, reducing computational complexity, or addressing domain-specific challenges. Overall, the literature showcases the diversity of techniques and highlights the need for further exploration to enhance the accuracy, efficiency, and applicability of human pose estimation methods.

2.2 Pose Estimation in Sports

The field of pose estimation in sports has gained significant attention as it offers valuable insights into athletes' movements and performance during various sporting disciplines. Accurate pose estimation in sports enables the analysis of key anatomical keypoints, capturing essential information about body positions, joint angles, and dynamics. By leveraging advanced computer vision techniques, researchers have developed specialized approaches to tackle the unique challenges posed by sports scenarios. These approaches aim to enhance the understanding of athletes' motions, optimize training techniques, provide real-time feedback, and facilitate performance evaluation. In this section, we delve into the literature on pose estimation in sports, exploring the specific methodologies, and applications that have contributed to advancements in this field.

One common approach is the use of hierarchical spatial models. These models aim to capture high-order dependencies among body parts while maintaining a compact representation. For example, a hierarchical spatial model was proposed that employed a mixture representation on each part and utilized latent nodes to represent spatial relationships. This model demonstrated the ability to capture poses, accurately reconstruct unseen poses, and outperform previous hierarchical models on challenging datasets [44]. Another study introduced a novel technique for exploiting dependencies between images in a collection. By sharing appearance models, this approach improved pose estimation results, particularly in sports scenarios [10].

Controlling variations in training datasets is another important aspect in sports-specific pose estimation. To address this limitation, a technique was proposed to generate synthetic samples with controlled pose and shape variations. By leveraging computer graphics advancements, this approach achieved state-of-the-art results in articulated human detection and pose estimation tasks [37].

In the context of team sports videos, where players often wear helmets and engage in various activities, specific techniques have been developed. For instance, a method was proposed to estimate head and upper body poses accurately by utilizing both pelvis and head tracking. This approach incorporated random decision forest classifiers and focused

on the upper body region, enabling pose estimation even with intensive movement and without temporal filtering techniques [17].

To improve articulated pose estimation, researchers have explored various appearance representations and combined flexible spatial models with image-conditioned spatial models. The experiments showed that these enhancements resulted in state-of-the-art performance on benchmark datasets, such as the "Leeds Sports Poses" and "Parse" benchmarks [35].

Other studies focused on specific body regions, such as the lower body. One proposed method used a label-grid classifier to estimate lower body joint positions, even in challenging scenarios, such as motion-blurred and low-resolution images. This approach outperformed traditional part-based models and demonstrated robustness in different poses and scales in team sports videos [16].

Efficiency and scalability are also important considerations in sports pose estimation. To address these challenges, a fast pose distillation model learning strategy was proposed. By transferring pose structure knowledge from a strong teacher network to a lightweight student network, this approach achieved superior cost-effectiveness on standard benchmark datasets [51].

One approach focuses on the use of LSTM-Attention models, which leverage two-branch multi-stage CNNs to extract human joint features in the spatial dimension. These models ensure real-time performance while maintaining high accuracy. Experimental results demonstrate the effectiveness of this method in achieving high-performance pose estimation for sports-related applications [49].

Researchers have explored fine-tuning techniques in sports pose estimation using a few labeled images plus a set of unlabelled images, Figure 2.6. These methods, including pseudo labeling and mean teacher approaches, allow for competitive results with limited labeled data. They bridge the gap between fully supervised training and fine-tuning on a few labeled poses [27].

Researchers have developed a MobileNet CNN-based model that accurately identifies 18 anatomical key points in athletes without specialized sensors. This model facilitates analysis of running behavior and gait parameters, such as cadence, knee angle, and velocity. Its easy implementation reduces reliance on personal trainers and expensive equipment, aiding athletes in optimizing performance [39].

A novel method improves training of 2D pose estimators for extreme poses in sports. Leveraging a sports-specific dataset and data augmentation, it achieves accurate 2D pose estimation during acrobatic movements, surpassing state-of-the-art performance [23].

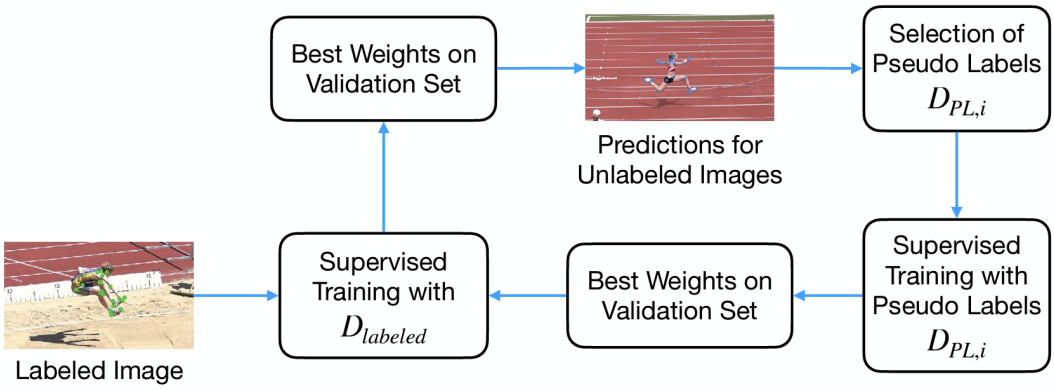

Figure 2.6: The procedure for pseudo label training, illustrated in Figure 2 of the paper introducing self-supervised learning for human pose estimation in the sports context [27].

A top-down pose estimation approach combines self-calibrating networks and graph convolutional neural networks to enhance accuracy. It focuses on human body detection for athletes in sports competitions, utilizing coordinate regression and heatmap testing. Deconvolution with high-resolution feature maps improves single-target pose estimation recognition [9].

In the context of bowling, a deep-learning approach called BowlingDL is proposed for pose estimation and classification of bowling players. The model utilizes the MoveNet model for pose estimation and the BowlingDL model for classifying the detected poses. The proposed approach achieves high accuracy on a custom dataset and is deployed in a smart mobile application for bowling players [20].

For ski jumping, an image-based pose estimation method is proposed using an efficient channel attention (ECA) module embedded in a high-resolution network (HRNet), Figure 2.7. The method achieves high precision in estimating ski jumper poses and allows for analyzing motion characteristics such as hip and knee angles. Transfer learning is employed to leverage feature knowledge from the COCO2017 dataset, and the proposed approach demonstrates promising results [4].

In swimming, a marker-less 2D swimmer pose estimation system called swimmerNET is introduced. This system combines computer vision algorithms and fully convolutional neural networks to estimate the pose of swimmers during exercise using a single wide-angle camera. The proposed approach achieves accurate pose estimation without the need for wearable sensors or optical markers, providing a non-intrusive solution for analyzing

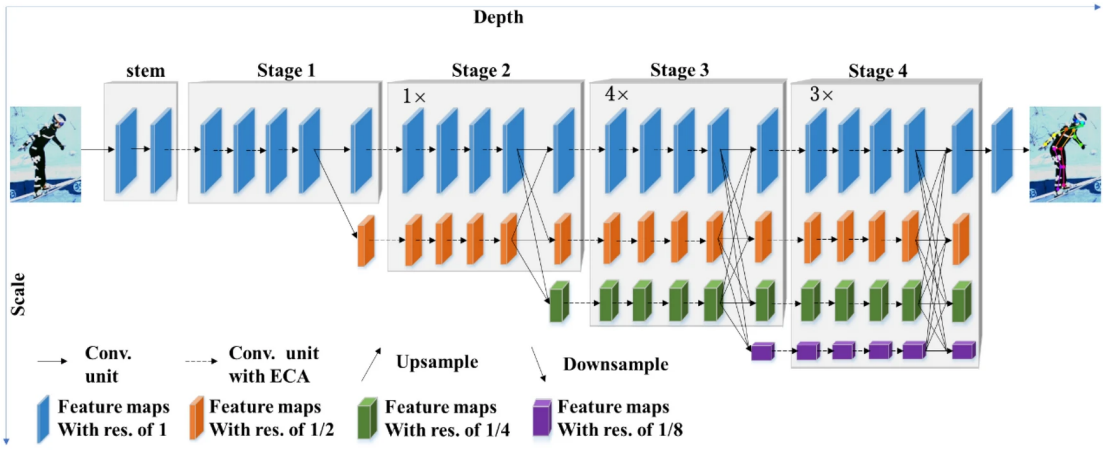

Figure 2.7: The ECA-HRNET architecture, showcased in Figure 5 of the paper introducing the architecture [4].

swimming technique [14].

In the domain of fitness applications and mobility activities, a unique approach for single-person pose estimation and action recognition is presented. The framework shown in Figure 2.8 consists of a base network for initial pose estimation and an Intensive Feature Consistency (IFC) network for refinement. The IFC network enforces high-level constraints on global body intensity correction and local body part adjustments, improving pose estimation accuracy with real-time processing speed on the CPU platform [40].

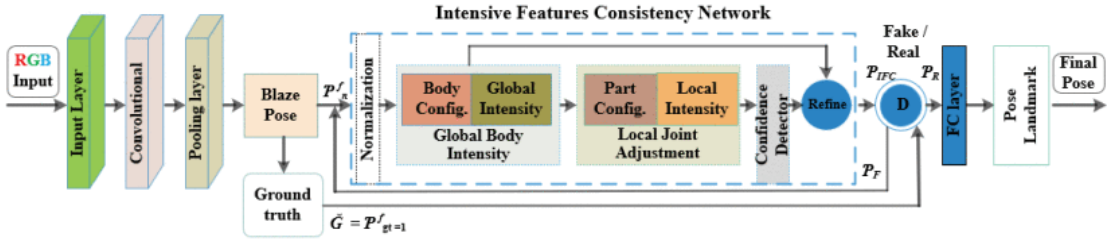

Figure 2.8: Pose estimation and refinement process using Intensive Feature Consistency (IFC) network, illustrated in Figure 2 of the paper introducing IFC [40].

Lastly, some studies explored the use of human pose estimation in sports video analysis applications. These applications included creating AI coach systems for personalized athletic training experiences and enabling content browsing and searching through video tagging and summarization. These techniques demonstrated promising results in provid-

ing better user training experiences and enabling applications such as action recognition, retrieval, and detection in sports videos [46, 15, 13, 18].

In conclusion, various approaches have been developed to address the challenges of pose estimation in sports. While all the methods aim to accurately detect and track human poses in sports scenarios, they exhibit similarities and differences in their approaches.

One commonality among the approaches is the utilization of deep learning techniques, such as convolutional neural networks (CNNs), to extract meaningful features from sports-related images or videos. This allows for the effective representation and analysis of human poses in dynamic sports activities. Additionally, many methods incorporate sophisticated architectures, such as hourglass networks or pose machines, to capture the hierarchical structure of human poses and model the spatial relationships between body parts.

Despite these commonalities, there are notable differences in the specific strategies employed. Some approaches focus on top-down estimation, where a bounding box or region proposal is first detected to localize the athlete, followed by pose estimation within that region. Others adopt a bottom-up approach, detecting individual keypoints and then grouping them to form complete poses. The choice of keypoint representation, whether it is based on heatmap estimation or direct regression, also varies among the methods. Moreover, certain techniques address the challenges of extreme poses or occlusions commonly encountered in sports scenarios, introducing novel data augmentation strategies or attention mechanisms to improve accuracy.

Overall, the diverse approaches to pose estimation in sports reflect the ongoing efforts to enhance the accuracy and robustness of human pose analysis in dynamic and challenging sports environments. By leveraging deep learning techniques and innovative architectures, researchers continue to explore new avenues for advancing the field and enabling applications such as sports analytics, performance optimization, and injury prevention.

Chapter 3

Methodology

This chapter presents a comprehensive overview of the methodology employed in this research to develop the GoalieNet model. In Section 3.1, the dataset employed for goalie pose estimation in ice hockey is introduced, outlining its distinctive incorporation of both human and non-human keypoints. Challenges such as occlusions and instances of missed annotations that impact the presence of keypoints in frames are highlighted. Section 3.2 outlines the two proposed approaches: the One-Stage Learning approach utilizing the Multi-Stage Pose Network (MSPN), and the Two-Stage Learning approach involving sequential Keypoint Predictor and Fusion Networks. The formulation of the loss function is expounded upon in Section 3.2.3, while the method for coordinate estimation is detailed in Section 3.2.4. The chapter sets the foundation for subsequent sections that explore training, evaluation, and results of GoalieNet, emphasizing its contributions to sports analytics using computer vision.

3.1 Dataset

The dataset utilized in this research consists of 34 National Hockey League (NHL) video clips, each containing various sequences of ice hockey gameplay. The dataset was specifically annotated to address the task of goalie pose estimation in ice hockey with relevant keypoints. These annotations were crucial for accurately determining the positions of goalies, equipment and net. The data annotation process was carried out in collaboration with Stathletes [1], a Canadian company that specializes in improving the process of player

[1] https://www.stathletes.com/

evaluation in sports. There are a total 44912 frames in the data, out of which 5040 frames containing a goalie.

The keypoints identified for pose estimation in the dataset are listed in Table 3.1. The keypoints include 22 positions related to the goalie's body, four keypoints associated with the net, and three keypoints corresponding to the goalie's stick.

The combination of human and non-human keypoints makes our dataset unique compared to other common pose estimation datasets, such as COCO [25]. While most pose estimation datasets primarily focus on human keypoints, our dataset introduces the additional complexity of incorporating non-human keypoints, specifically related to the equipment (net and stick) in the ice hockey context.

Figure 3.1 provides a visual representation of the annotated keypoints on a sample frame. The visualization helps to understand the relevance position of keypoints in the pose estimation task.

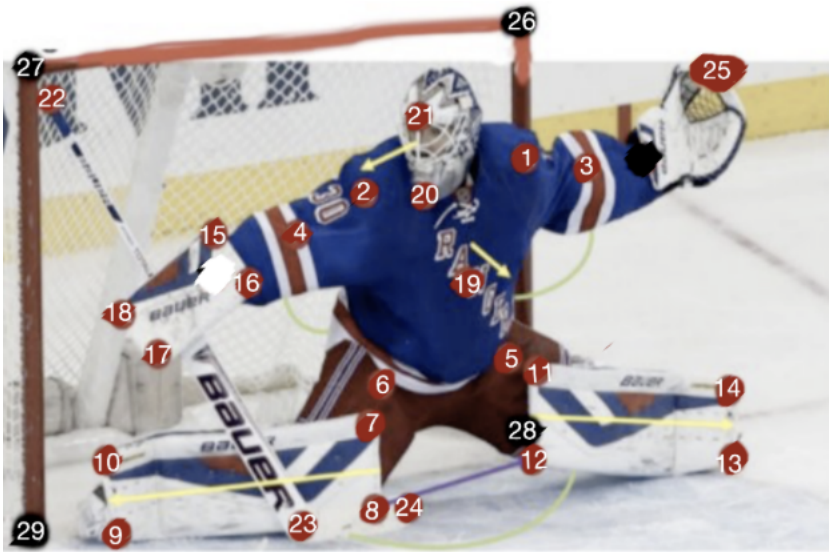

Figure 3.1: The illustration[2]depicts the keypoints' positions for the goalie, equipment, and net. Each keypoint is labeled with an assigned number and name, as shown in Table 3.1.

[2]This picture is created by Mehrnaz Fani and Stathletes company.

Table 3.1: This table presents the specific numbers and names assigned to the keypoints for the goalie, equipment, and net. The positions of each keypoint, along with its corresponding number, are illustrated in Figure 3.1.

1	Left Shoulder	11	Left Legpad (0)	21	Mask-High
2	Right Shoulder	12	Left Legpad (1)	22	Stick-Upper
3	Left Elbow	13	Left Legpad (2)	23	Stick-Lower
4	Right Elbow	14	Left Legpad (3)	24	Stick-Blade-Tip
5	Left Hip	15	Blocker (0)	25	Mit-Top
6	Right Hip	16	Blocker (1)	26	Net-Top-Left
7	Right Legpad (0)	17	Blocker (2)	27	Net-Top-Right
8	Right Legpad (1)	18	Blocker (3)	28	Net-Bottom-Left
9	Right Legpad (2)	19	Torso Center	29	Net-Bottom-Right
10	Right Legpad (3)	20	Mask-Low		

3.1.1 Challenges

One notable challenge presented by this dataset is that not all frames contain all keypoints. The presence of occlusion or missed annotations, results in certain keypoints being not visible or annotated in some frames. Consequently, the percentage of frames containing specific keypoints exhibits significant variation, a trend visually illustrated in Figure 3.2.

To delve deeper into this variability, we offer a more detailed visualization in Figure 3.3. This representation highlights a distinct lack of keypoints, particularly in the lower regions of the goalie's body. Given the rapid movements of these lower parts during gameplay, accurately estimating the coordinates of their keypoints becomes a challenge[21]. The scarcity of data in these regions could potentially impact the model's capability to comprehensively capture the goalie's complete pose, a crucial consideration for GoalieNet's performance in a fast-paced sport like ice hockey, making it important for the robustness of our model.

Figure 3.4 further contributes to our understanding of challenges related to this dataset. Each video clip corresponds to a specific game, and the figure illustrates the distribution of the number of frames exclusively featuring goalies during these games. Understanding this distribution is crucial as frames within a single video clip are often both sequential

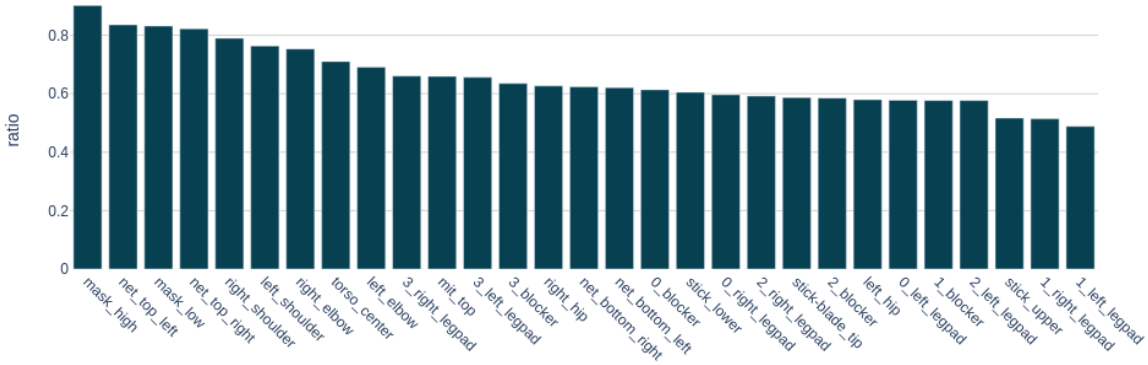

Figure 3.2: Proportion of frames featuring specific keypoints.

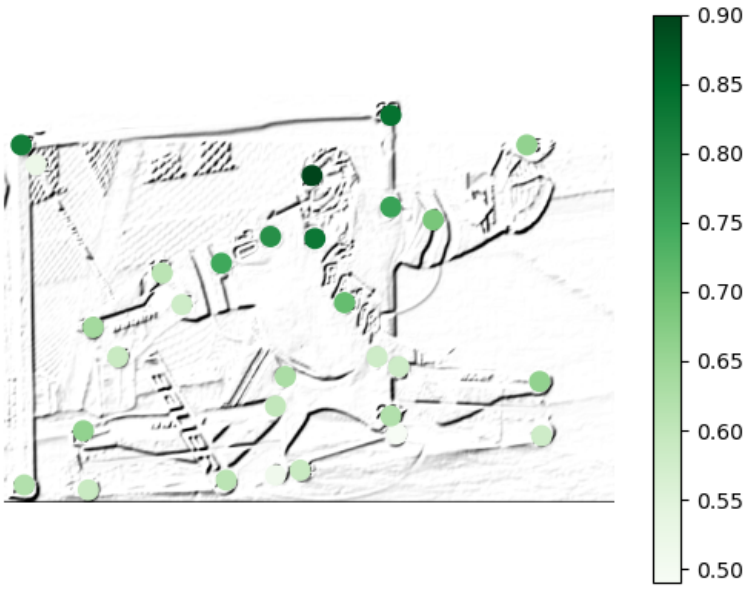

Figure 3.3: This graph depicts the ratio of frames containing specific visible keypoints based on their respective positions on the goalie body, equipment or net.

and visually correlated. It is essential to be aware of this correlation to avoid introducing bias during data splitting for evaluations. Therefore, careful consideration of the video clip structure is imperative to ensure objective and reliable evaluation of our GoalieNet model.

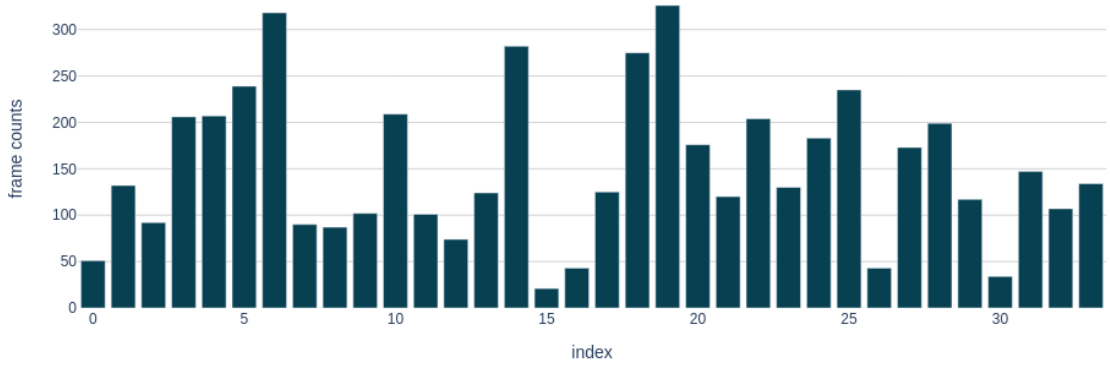

Figure 3.4: The number of frames featuring goalies per video clip.

3.2 Proposed Method

To estimate the positions of 29 keypoints across the goalie, equipment, and net, two distinct approaches are employed. The initial approach involves utilizing the Multi-Stage Pose Network architecture [24], where a joint goalie, equipment, and net pose estimation training methodology is applied. This strategy capitalizes on the interplay between the goalie's position, equipment placement, and net orientation, enabling the network to leverage the spatial relationships for accurate predictions.

One-Stage Learning GoalieNet has its own challenges due to the complexity of adhering the complex interactions and spatial dependencies among various elements within the scene. It turns out that intricate nature of the stick, which always moves quickly with the goalie and changes the direction quickly, introduces an additional layer of difficulty in the development of accurate and robust joint pose estimation algorithms tailored to the specific demands of ice hockey scenarios. In response this complexity, the second approach, which is a Two-Stage Learning paradigm within the same network paradigm, comes to play. In this method, a network is first trained to estimate all keypoints except for the

stick. Subsequently, leveraging the latent poses and original images, a separate network is trained specifically to estimate stick keypoints. This bifurcated strategy is devised to enhance the model's ability to tackle the specific challenges inherent to the estimation of goalie, equipment, and net poses while maintaining a coherent and adaptable network architecture.

Both approaches are carefully tailored to tackle the distinctive challenges presented by our dataset and the intricate spatial correlations among goalie, equipment, and net keypoints. In both instances, the network's output comprises a collection of heatmaps, with each heatmap corresponding to a specific keypoint on the goalie, equipment, or net.

3.2.1 GoalieNet: One-Stage Learning

The architecture of the proposed One-Stage Learning GoalieNet draws inspiration from the framework introduced in Li et al. [24], as illustrated in Figure 3.5. This architecture, referred to as Multi-Stage Pose Network (MSPN), embraces a multi-stage design, comprising interconnected stages, each with distinct functionalities. This configuration equips the network to effectively capture intricate spatial dependencies and gradually refine the pose estimation process. The incorporation of cross-stage aggregation between adjacent stages further contributes to the preservation of image quality, an indispensable property given the considerable distance from which the videos are captured.

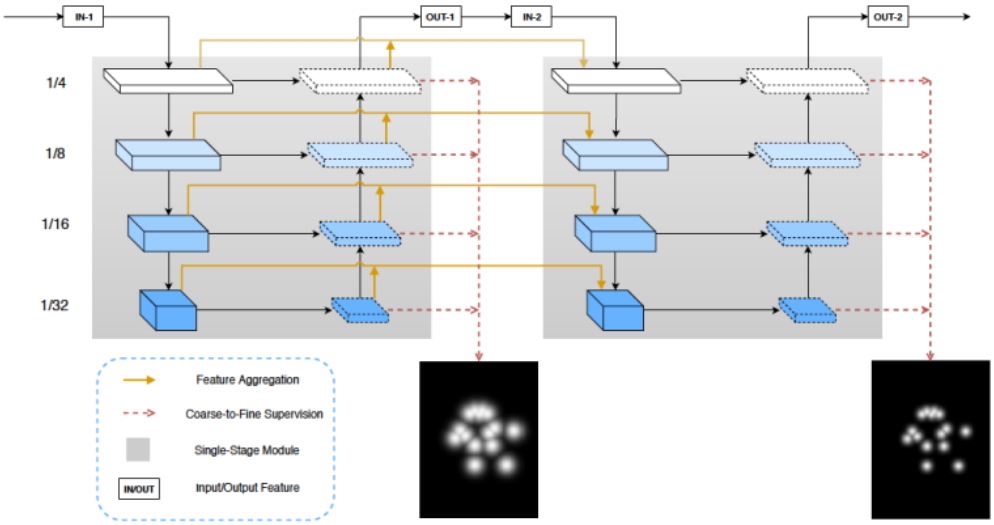

Figure 3.5: Multi-Stage Pose Network [24] architecture.

To enhance localization accuracy, MSPN employs varying kernel sizes across different levels. This approach proves particularly vital for accurately localizing keypoints, such as those on the leg pads, where the proximity of keypoints presents a challenge for traditional pose estimation methods. The adoption of diverse kernel sizes empowers the model to meticulously capture these nuanced details.

The intrinsic capability of MSPN to leverage cross-stage aggregation and using variyng kernel sizes aligns seamlessly with the demands of our pose estimation task within ice hockey scenarios. The dynamics of goalie poses, coupled with the varying positions of the net and equipment across frames, necessitate a heightened localization precision. In this context, MSPN emerges as a fitting solution, poised to meet the intricate demands of accurate pose estimation in the realm of ice hockey.

To leverage the strong spatial relationship between goalie, equipment, and net in real ice hockey scenarios, the network is trained in a joint fashion, encompassing all 29 keypoints across these entities. Joint training allows the network to benefit from the collective information of all keypoints and their interactions, leading to more accurate and coherent pose estimations compared to training with separate entities.

3.2.2 GoalieNet: Two-Stage Learning

The endeavor to predict all 29 keypoints concurrently within the context of the One-Stage Learning GoalieNet revealed its inherent challenge. Especially, the intricate nature of the stick, characterized by its swift movement with the goalie and slender structure, increases the difficulty for the network in estimating its keypoints alongside the other crucial keypoints. Furthermore, this simultaneous prediction of all keypoints could potentially impede the learning process for other keypoints.

In light of these considerations, a viable solution has been formulated in the form of a Two-Stage Learning framework, visualized in Figure 3.6. Notably, how goalies stand, their orientation, the direction the net faces, and exactly where the stick is placed are closely connected parts in ice hockey keypoint estimation for goalie. Specifically, most of the time the net's positioning remains relatively stable, while the goalie's posture significantly influences the stick's orientation. Thus, a fine estimation of goalie, their gear, an the net's keypoints serves as a foundational step for accurate stick's keypoint prediction.

Recognizing this interplay, our method follows a sequential approach that begins by looking closely at how the goalie stands, the way their gear is set up, and the precise configuration of the net. This approach strategically initiates with the estimation of all keypoints except those associated with the stick, subsequently addressing the intricate

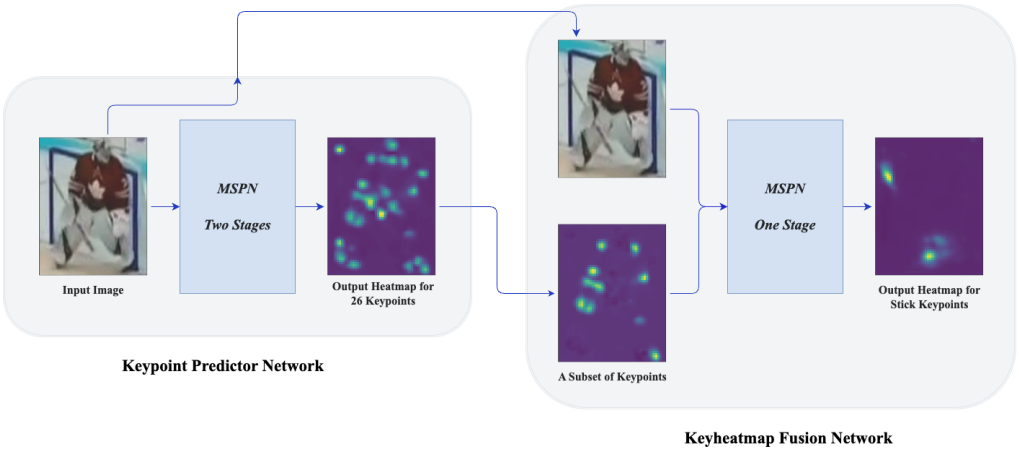

Figure 3.6: Illustration of the Two-Stage Learning GoalieNet architecture, comprising two integral components: the Keypoint Predictor Network (KPN) and the Keyheatmap Fusion Network (KFN). The input image is processed by the KPN to generate heatmaps for 26 out of 29 keypoints. Subsequently, the initial image is concatenated with a heatmap derived from a chosen subset of keypoints, forming the input to the Keyheatmap Fusion Network. This final stage produces heatmaps for 3 stick-related keypoints.

task of predicting stick keypoints. This sequential order is motivated by the fact that the goalie's positioning profoundly impacts the spatial arrangement of all other elements on the ice.

This integrated design not only enhances prediction accuracy, at least for some keypoints, but also ensures that the predictions are meaningfully embedded within the broader context of the game, facilitating a more informed understanding of player movements and interactions.

Embedded within this Two-Stage Learning framework are two distinct networks: the Keypoint Predictor Network and the Keyheatmap Fusion Network, which we will explain comprehensive explanation of them in the subsequent sections.

3.2.2.1 Keypoint Predictor Network

A foundational component within the architecture of Two-Stage Learning GoalieNet is the Keypoint Predictor Network (KPN). This network serves as a pivotal cornerstone that contributes significantly to the overall framework. At its fundamental core, the KPN is

tasked with the complex mission of jointly predicting 26 out of the total 29 keypoints that are integral to the overall system.

Specifically, operating within the framework of MSPN's heatmap-based architecture, the images serve as inputs to the KPN, which then undertakes the pivotal role of generating 26 heatmaps realted to all keypoints except the ones corresponding to the stick. This process concludes in the creation of latent features characterized by a dimension of $26 \times H \times W$. Here, H and W represents the height and width of the output heatmaps, respectively. This approach empowers the KPN to encapsulate the spatial intricacies of keypoint estimation by focusing solely on the goalie, gear, and net positions, effectively paving the way for the subsequent stages of fusion and stick keypoints prediction within the GoalieNet framework.

3.2.2.2 Keyheatmap Fusion Network

To address the complexity of accurately predicting the positions of the stick keypoints while maximizing the effective utilization of available information, we introduce the keyheatmap fusion network (KFN) in the subsequent stage of learning, building upon the latent features generated by the Keypoint Predictor Network.

Within this framework, the KFN undertakes the responsibility of predicting the stick keypoints, leveraging the fused information to achieve a refined and comprehensive prediction outcome. This network receives a concatenated four-channel input of $4 \times H \times W$, where three channels correspond to the original image input itself, and the additional channel contains the cumulative sum of selected latent poses heatmaps generated by KPN. This selection focuses on a subset of keypoints that we deem primarily influential in the estimation of stick keypoints. This fusion is strategically accomplished by concatenating the latent features with image-based features, effectively infusing the network with multi-modal insights.

Importantly, this step involves freezing the weights of KPN, leveraging the accumulated knowledge from the earlier prediction stage. The structural backbone of the KFN aligns with the MSPN architecture, thereby expanding its versatility to this critical stage.

This comprehensive strategy ensures that the nuanced interdependencies among keypoints, particularly the complex interplay between the positions of the goalie, net, and stick, are effectively harnessed to enhance the accuracy and contextual relevance of the predictions within the dynamic context of ice hockey.

3.2.3 Loss Function

The loss function employed in this study takes cues from Li et al. [24], which delved into intermediate supervision as a means of enhancing accuracy. In each specific stage s and for each sample i, at each level l of the network, the loss $L_i^{s,l}$ is calculated as the mean square of the L2 distance between the predicted and actual heatmaps across all keypoints, as indicated by (3.1). Here, $gt_{i,j}^{s,l}$ and $p_{i,j}^{s,l}$ denote the actual and predicted heatmap of the j^{th} keypoint for the i^{th} sample at level l and stage s of the network. When dealing with invisible keypoints, the Multi-Stage Pose Network (MSPN) push the intermediate levels to generate zero pixel values by assigning a ground truth value of zero to these specific keypoints ($gt_{i,j}^{s,l} = 0$). Another noteworthy aspect of the loss calculation for the final output, i.e. $l = L$, is the utilization of the online hard keypoint mining loss strategy [8]. This approach involves considering only the top k visible keypoints with the highest L2 distance from their actual heatmaps in the gradient calculation. This approach optimizes the training process by prioritizing challenging keypoints and effectively focusing the network's learning on areas that require improvement.

$$L_i^{s,l} = \sum_{j \in \text{set}} \|gt_{i,j}^{s,l} - p_{i,j}^{s,l}\|_2^2 \quad \text{set is} \begin{cases} \text{top } k \text{ visible keypoints with the highest L2 loss if } l = L \\ \text{all keypoints otherwise} \end{cases} \tag{3.1}$$

Ultimately, (3.2) is employed to compute the overall loss, L_{tot}, and subsequently update the network's weights. In this equation, N represents the number of samples, S denotes the number of stages, and L signifies the number of levels.

$$L_{tot} = \frac{1}{N} \sum_{n=1}^{N} \sum_{s=1}^{S} \sum_{l=1}^{L} L_i^{s,l} \tag{3.2}$$

This comprehensive approach encapsulates the essence of the training process and guides the network's learning.

3.2.4 Coordinates Estimation

During the inference stage, the predicted coordinates of a given keypoint are derived by locating the point that lies at 25% of the length along a vector connecting the heatmap's first and second maximum values [24]. This heatmap-based approach enables precise and

spatially-aware keypoint localization, accounting for variations and interactions among keypoints.

3.3 Summary

In this chapter, we presented the methodology employed in developing the GoalieNet framework for joint goalie, equipment, and net pose estimation in ice hockey. The approach consists of a Two-Stage Learning strategy, encompassing the Keypoint Predictor Network (KPN) and the Keyheatmap Fusion Network (KFN).

We began by highlighting the challenges posed by the One-Stage Learning approach due to the complexity of stick movement and the need to predict all keypoints simultaneously. In response, the Two-Stage Learning approach was introduced, addressing these challenges through specialized networks. KPN effectively predicts goalie, gear, and net keypoints, while KFN, building upon fused information, focuses on predicting stick keypoints.

The loss function was detailed, drawing inspiration from intermediate supervision principles and online hard keypoint mining. It ensures accurate training by prioritizing challenging keypoints and considering both visible and invisible keypoints. Finally, we outlined the process of coordinate estimation from generated heatmaps, underscoring its significance in accurate localization.

The methodology, encompassing Two-Stage Learning, specialized networks, loss function, and precise coordinate estimation, forms the foundation of GoalieNet's capability to estimate player poses comprehensively, contributing to enhanced sports analytics in the context of ice hockey.

Chapter 4

Experimental Results

This chapter presents the outcomes of an in-depth investigation into GoalieNet pose estimation method. Through meticulous testing and assessment, it showcases the effectiveness of a Two-Stage Learning strategy that addresses the distinctive complexities of this particular task. By thoroughly examining network design, training techniques, and performance evaluation, this chapter provides a comprehensive understanding of the model's strengths and weaknesses.

The essence of this chapter is to present a coherent depiction of the model's capabilities in deciphering the intricate movements and orientations of goaltenders in the dynamic sport of ice hockey. By incorporating quantitative metrics, qualitative evaluations, and focused analyses, this chapter functions as a guide that navigates us through the multifaceted landscape of goalie pose estimation. This aids in comprehending the potential influence of the model and avenues for future enhancements.

4.1 Network Design and Training

The Keypoint Predictor Network (KPN) is composed of two stages, both utilizing 256 upsampling channels. In contrast, the Keyheatmap Fusion Network (KFN) consists of a single stage with 64 upsampling channels. Both networks have four levels. During the One-Stage Learning phase, the model parameters employed align with those of KPN. For loss computation, the top-k values are set at 8 for KPN and 2 for KFN. The same top-k value as KPN is employed for the One-Stage network.

The weights of the networks were initialized following the same approach as in Li et al.'s work [24]. The training process involves iterative training for a specific number of iterations: 1500 iterations for the One-Stage GoalieNet, 3500 iterations for KPN, and 2500 iterations for KFN. These iteration counts have been determined as yielding optimal performance based on validation results.

The goalie bounding boxes were manually annotated in the dataset. Across all networks, the input dimensions remain constant at 256×192, while the output dimensions are set to 64×48. To optimize the networks, the Adam optimizer is utilized with an initial base learning rate of 0.0005. The chosen batch size for training is set at 32.

As discussed in Section 3.1.1, the correlation among different frames within a specific video clip was identified as a challenge we aimed to address in this task. To mitigate the impact of frame correlation on test accuracy, we adopted a strategy of splitting the dataset for training, testing, and validation based on the individual video clips. This approach entails assigning the entire set of frames from a particular video clip to either the train, test, or validation subsets. While implementing this strategy, we maintained a distribution ratio of approximately 80%, 10%, and 10% among the total frames within each group. The training data consisted of 28 video clips, while each of the test data and validation data each contained 3 video clips.

4.2 Performance Evaluation

Given the unique nature of the goalie pose estimation task, which involves non-human joints unlike conventional human pose estimation tasks, the application of conventional prevalence indicators, such as object keypoint similarity metric [25], as described in (4.1), is not feasible. In this formula, d_i represents the euclidean distance between ground truth and predicted keypoint, and s is the scale of the bounding box divided by the total area of the image. Notably, k_i is a constant value per-keypoint, which is typically based on human joints scale and is not applicable to non-human keypoints of our dataset.

$$OKS = exp(-\frac{d_i^2}{2s^2k_i^2}) \tag{4.1}$$

As an alternative approach, we adopt a simplified version of the percentage of detected joints, which is used in Deeppose [45]. Doing so, we initially calculate the normalized localization error for each keypoint, denoted as NLE_{kp}, as demonstrated in (4.2). In this equation, the vectors P_{kp} and A_{kp} represent the predicted and actual keypoint positions

as (P_x, P_y) and (A_x, A_y), respectively, and d_b refers to bounding box diagonal. Then, the criterion described in (4.3) is adopted as our keypoint detection metric. Based on this criterion, if the normalized localization error falls below a pre-difined threshold, the keypoint is deemed detected. Within this framework, a threshold value of 0.05 is uniformly applied across all keypoints to ascertain their detection status.

$$NLE_{kp} = \frac{\|P_{kp} - A_{kp}\|_2}{d_b} \tag{4.2}$$

$$\text{Keypoint Detection Criterion: } NLE_{kp} \leq \text{Threshold} \tag{4.3}$$

Once the classification is established for each keypoint, we assess the detection accuracy of a specific keypoint, denoted as DA_{kp}, by calculating the percentage of successfully detected keypoints, as shown in (4.4). In this equation, N_{kp}^{det} represents the number of detected keypoints of type kp, and N_{kp}^{vis} is the count of visible keypoints of the same type.

$$DA_{kp} = \frac{N_{kp}^{det}}{N_{kp}^{vis}} \times 100 \tag{4.4}$$

Furthermore, gaining insights into the magnitude of prediction errors among the detected keypoints offers valuable information. Building upon the threshold-based detection in (4.3), a deeper analysis of normalized localization error, but only for detected keypoints, provides a more nuanced understanding of the model's precision. This performance metric is calculated based on 4.5, where $NLE_{kp,i}^{det}$ refers to the normalized localization error for i^{th} detected keypoint of type kp.

$$NLE_{kp}^{det} = \frac{1}{N_{kp}^{det}} \sum_{i=1}^{N_{kp}^{det}} NLE_{kp,i}^{det} \tag{4.5}$$

Given the extensive presence of over 29 keypoints, the utilization of a consolidated metric becomes imperative to succinctly summarize results and facilitate the comparison of outcomes across multiple joints. In order to provide a comprehensive overview of our framework's performance, we present both median and average of detection accuracy, and the average of normalized localization error across diverse classes.

4.3 Latent Feature Selection

A pivotal consideration in the Two-Stage Learning GoalieNet is the strategic selection of keypoints for input to the second learning stage, the Keyheatmap Fusion Network. In this context, we have drawn from intuitive reasoning by focusing on the premise that the stick is consistently held by the goalie, and in the same manner throughout a game. Consequently, keypoints situated on the hands, such as the elbows, shoulders, blocker and mit-top, as it is illustrated with blue points in Figure 4.1, were identified as potentially suitable candidates for input to the Keyheatmap Fusion Network (KPN).

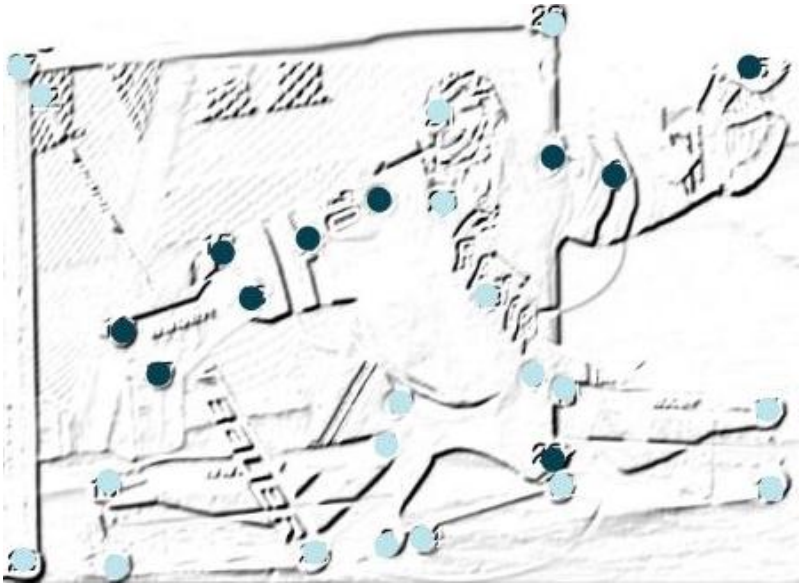

Figure 4.1: Keypoints located on the shoulders (1, 2), elbows (3, 4), blocker (15, 16, 17, 18), mit-top (25), and a single keypoint positioned on the net (28) were deliberately selected as inputs to the Keyheatmap Fusion Network (KFN) for the purpose of predicting stick keypoints.

Moreover, through experimentation, we observed employing a single keypoint associated with the net contributed to an enhancement in accuracy. This prompted us to include this specific keypoint in the set of inputs provided to the Keyheatmap Fusion Network.

4.4 Results

The performance results on both the test and validation datasets, each comprising three distinct video clips, are presented in Table 4.1. This table provide a comprehensive overview of the keypoint estimation outcomes achieved by the two methods. Upon close examination of the numerical values, it is evident that the Two-Stage Learning approach outperforms the One-Stage Learning method, showing a 1% higher median accuracy for keypoint estimation and a reduced localization error of 0.0187 for test data, in contrast to the One-Stage approach's 0.0194. However, for a more granular understanding of these summarized metrics, a detailed analysis of the accuracy across all 29 keypoints is warranted.

Metrics	**Test**		**Validation**	
	One-Stage	**Two-Stage**	**One-Stage**	**Two-Stage**
Detection Accuracy (Median)	70%	**71**%	86%	**88**%
Detection Accuracy (Mean)	71%	71%	85%	**87**%
Normalized Localization Error (Mean)	0.0194	**0.0187**	0.0179	**0.0168**

Table 4.1: Overall comparison between One-Stage Learning and Two-Stage Learning GoalieNet for both test and validation video clips. The evaluation metrics encompass all 29 available keypoints, with the detection threshold in (4.3) set at 0.05. Additionally, normalized localization error is calculated using (4.5).

Given the primary responsibility of the Keypoint Predictor Network (KPN) in estimating 26 out of the total 29 keypoints, we proceed to make a comparison of its performance with that of the same keypoints in the One-Stage GoalieNet. The results are presented in Table 4.2. Across both the median of the accuracy values of these 26 keypoints, and the localization error of detected keypoints, KPN consistently outperforms the One-Stage GoalieNet, increasing the median accuracy from 72% to 75% and decreasing localization error from 0.0195 to 0.0188 for test data. This observation underscores the effectiveness of this network in the Two-Stage GoalieNet approach. A closer examination of the dynamics reveals that excluding the stick keypoints from the set of keypoints during the first stage assists the network in enhancing the precision of estimating the remaining keypoints associated with the goalie, net, and their protective gear. This strategic division of labor within the Two-Stage framework contributes to the improved overall performance observed in the Table 4.1.

Regarding the stick keypoints, a notable observation arises from the data presented in Table 4.3. It becomes apparent that the Keyheatmap Fusion Network faces difficulties in

Metrics	Test		Validation	
	One-Stage	**Two-Stage**	**One-Stage**	**Two-Stage**
Detection Accuracy (Median)	72%	**75%**	87%	**90%**
Detection Accuracy (Mean)	73%	73%	85%	**88%**
Normalized Localization Error (Mean)	0.0195	**0.0188**	0.0170	**0.0168**

Table 4.2: Performance Comparison between One-Stage Learning results for all keypoints except stick-related keypoints and Keypoint Predictor Network (KPN) of Two-Stage Learning in test and validation video clips. The threshold in (4.3) is set at 0.05. Additionally, normalized localization error is calculated using (4.5).

accurately estimating stick-related keypoints. This is reflected in the results where the detection accuracy of stick keypoints is noticeably higher in the One-Stage Net compared to the Keyheatmap Fusion Network, the median detection accuracy stands at 60% for the former, while it reaches 49% for the latter, as observed in the test data. Interestingly, the KFN in Two-Stage GoalieNet exhibits a more precise localization accuracy for stick keypoints, with values of 0.0171 for the Two-Stage approach and 0.0181 for the One-Stage approach. This disparity between localization accuracy and detection accuracy emphasizes the nuanced challenges involved in predicting these intricate stick keypoints within the context of ice hockey scenarios.

Metrics	Test		Validation	
	One-Stage	**Two-Stage**	**One-Stage**	**Two-Stage**
Detection Accuracy (Median)	**60%**	49%	**74%**	71%
Detection Accuracy (Mean)	**55%**	53%	**73%**	72%
Normalized Localization Error (Mean)	0.0181	**0.0171**	0.0193	**0.0167**

Table 4.3: Performance Comparison between One-Stage Learning results for stick-related keypoints and Keyheatmap Fusion Network (KFN) of Two-Stage Learning in test and validation video clips. The threshold in (4.3) is set at 0.05. Additionally, normalized localization error is calculated using (4.5).

4.4.1 Joint-wise Comparison

Given that the ultimate goal of our approach is to determine the precise coordinates of keypoint positions, it becomes essential to conduct a comparison of joint-wise detection accuracy across different methodologies. To achieve this, we begin by categorizing individual keypoints into groups that correspond to the same body parts, such as shoulders, or equipment items like leg pads. We then calculate the mean of detection accuracy across members for each of these groups and assign this value to the overall group accuracy. Subsequently, we delve into a detailed discussion about each joint, aiming to identify any anomalies or weaknesses present in each method. This comprehensive analysis will provide us with a deeper understanding of the performance of various approaches and offer insights into their suitability for accurately pinpointing keypoints.

As illustrated in Figure 4.2, a more detailed examination of specific keypoints reveals interesting insights. Notably, for keypoints such as mit-top, which is a keypoint on the tip of the goalie's glove, stick, blocker, legpad, and hip, the accuracy achieved through the Two-Stage Learning approach appears to be comparatively lower than that of the One-Stage Learning approach. The most pronounced difference is observed in the case of "mit-top" keypoint. Unfortunately, the annotation quality for "mit-top" within our dataset poses challenges. Specifically, the actual position of mit-top is not consistently annotated as being exclusively at the tip of the goalie's mit. This variability is highlighted in Figure 4.3, which illustrates instances of significant positional changes in mit-top even between two consecutive frames.

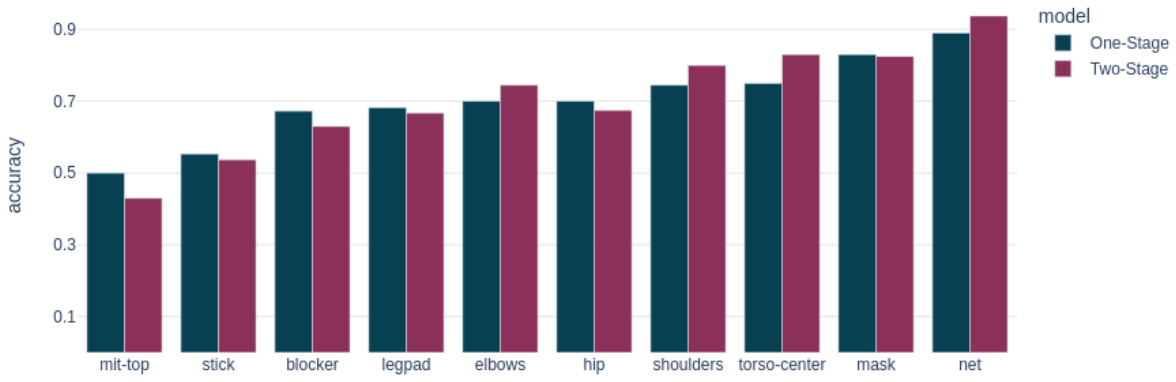

Figure 4.2: Mean detection accuracy for test videos in Two-Stage and One-Stage GoalieNet across different keypoints group.

(a) The annotated position of the "mit-top" keypoint in the first frame.

(b) The annotated position of the "mit-top" keypoint in the next frame of the same video.

Figure 4.3: In 4.3b, it appears that the "mit-top" keypoint is not annotated at the same position as it is in 4.3a.

Despite extensive manual efforts to clean and refine the data, mitigating this challenge proved to be complex. Removing all instances of loosely annotated keypoints might lead to substantial data loss, an outcome that was deemed undesirable. This issue becomes particularly relevant when considering the separation of stick-related keypoints from the rest. In the context of the Keyheatmap Fusion Network within the second stage of GoalieNet, having keypoints on both hands assumes paramount significance as latent heatmaps, ensuring that the network receives comprehensive input information to improve the estimation of stick keypoints. This close connection between how well we label keypoints, the cleanliness of our data, and what information we give to the network highlights the complexity of dealing with these challenges to make our predictions more accurate and reliable, especially for the more detailed keypoints.

However, considering that the positions of keypoints are closely tied to the stick's placement, since both are held by the hands, excluding the stick from the group of keypoints given to the Keypoint Predictor Network could potentially lead to a decrease in the accuracy of detecting the stick's position.

Analyzing the stick keypoints, as shown in Figure 4.4, it becomes evident that the Two-Stage GoalieNet exhibits enhancements in accurately detecting both the upper and lower portions of the stick compared to the One-Stage approach. Surprisingly, despite this improvement, the accuracy metric experiences a decrease of approximately 10% when the second stage network is utilized.

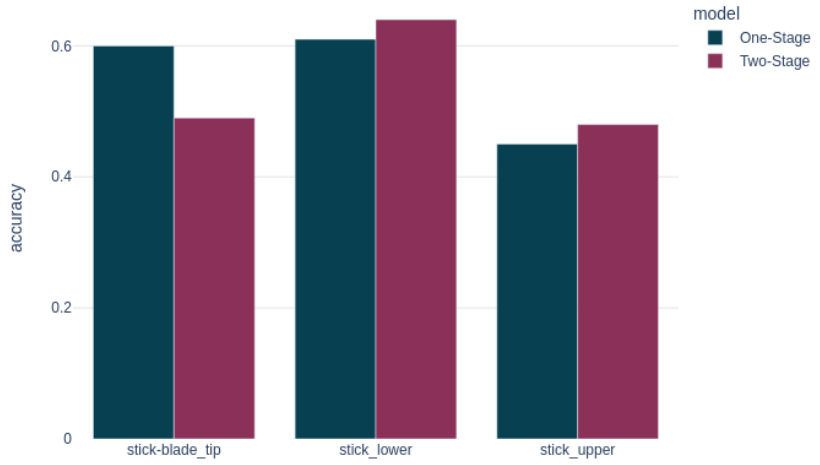

Figure 4.4: Detection accuracy for test videos in Two-Stage and One-Stage GoalieNet for 3 stick-related keypoints.

The decrease in accuracy when employing the second stage network for stick keypoints could be attributed to the model's attempt to fine-tune its predictions. While the Two-Stage GoalieNet succeeds in providing accurate estimations for the upper and lower portions of the stick, the subsequent refinement process might inadvertently introduce noise, leading to a reduction in accuracy. Especially because the Keyheatmap Fusion Network (KFN) is a One-Stage network which may result in lacking the sufficient expertise to precisely refine its estimations. The complexity arises from the rapid movement and rotational changes of the stick's tip. Such intricate movements can be challenging to capture effectively in a One-Stage network, where nuanced adjustments might lead to unintended inaccuracies. This underscores the need for a more sophisticated strategy to refine predictions for keypoints that involve swift and multi-dimensional movements.

The comprehensive assessment of the overall accuracy of both methods is visually sum-

marized in Figure 4.5. A closer inspection reveals that certain keypoints exhibit lower accuracy within the Two-Stage Learning method. Notably, keypoints such as legpads and blocker stand out as areas where the Two-Stage Learning method falls short. These specific keypoints hold significant importance, as highlighted in Section 4.3, where we elucidated that their latent heatmaps constitute vital inputs to the Keyheatmap Fusion Network. The accuracy of predictions for these heatmaps bears profound implications for the subsequent stages of the network's estimation process.

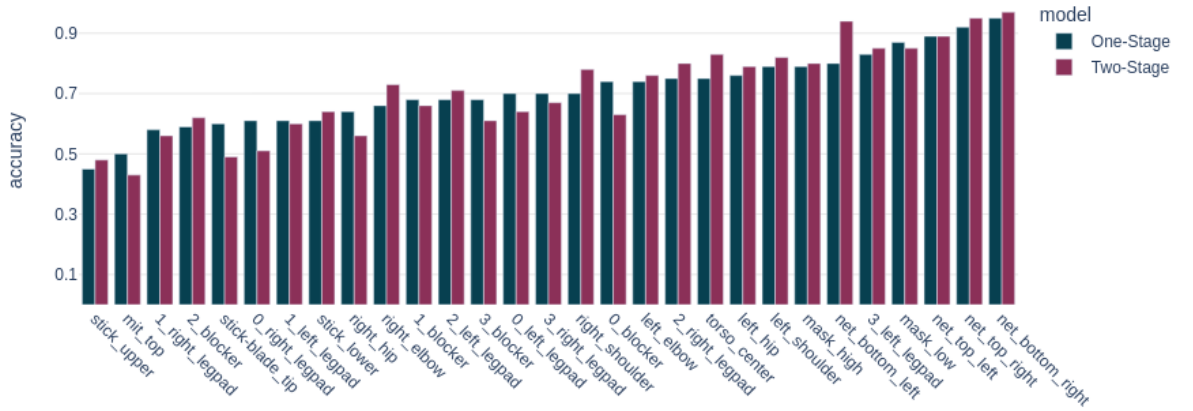

Figure 4.5: Keypoint detection accuracy for all keypoints in One-Stage and Two-Stage GoalieNet.

Interestingly, a pivotal observation arises from the performance of the Keypoint Predictor Network. The removal of stick-related keypoints from its input set appears to have contributed to a decrease in detection accuracy for these keypoints. This outcome underscores the intricate interplay between different keypoints and their collective influence on prediction quality. Specifically, this highlights that stick-related keypoints are not only important in their own right but also play an integral role in refining the estimation process for other keypoints.

In essence, the accuracy trends across different keypoints underscore the necessity of a nuanced approach that considers both individual and collective influences within the Two-Stage framework. This analysis serves as a valuable guide for refining network architecture, input selection, and optimization strategies to enhance the overall accuracy of the Two-Stage Learning GoalieNet, further emphasizing the significance of a holistic approach in

goalie pose estimation.

4.5 Discussion and Future Research

In this study, we delve into the intricate realm of goalie pose estimation, a task that poses unique challenges compared to conventional human pose estimation. While the One-Stage Learning GoalieNet offers insights, it has its limitations, especially when dealing with swift-moving and intricately related elements such as the stick. The introduction of the Two-Stage Learning GoalieNet proves promising, demonstrating enhanced accuracy in most keypoints. However, it is evident that KFN's refinement might lead to slight inaccuracies for the tip of the stick, highlighting the delicate balance between correction needed for different keypoints in Two-Stage Learning.

One of the critical factors contributing to the performance of our Two-Stage approach is the selection of keypoints that we input along with the original frame into the Keyheatmap Fusion Network to predict the stick keypoints. Intuitively, we observe that during a game, the goalie always holds the stick in one of their hands. Therefore, we selected all keypoints on both hands of the goalie as our latent pose heatmaps, as illustrated in Figure 4.1. Through experimentation, we have empirically verified that this intuition holds, and any removal of these keypoints from the set, even the noisiest keypoint like mit-top, results in a degradation of the Two-Stage method's detection accuracy for the stick. Further exploration of other keypoints revealed that adding the upper left keypoint of the net improves the detection accuracy. This keypoint consistently demonstrates high accuracy in our Two-Stage model, as depicted in 4.5. We interpret this finding as an indication of our Keyheatmap Fusion Network's sensitivity to the quality of latent pose estimated by Keypoint Predictor Network.

Another crucial aspect to address when interpreting the outcomes lies in the meticulous nature of dataset annotations. Our approach delves into understanding the interplay between model detection accuracy for test data and the proportion of visible keypoints within the training dataset. This exploration is motivated by our intent to study the correlation between the adequacy of data and the resulting detection accuracy. To enhance the comprehensibility of our findings, we've organized the keypoints into coherent groups representing specific body parts. By computing the mean detection accuracy for individual keypoints within each group, we've mirrored the methodology employed in Figure 4.2. Another reason for grouping keypoints based on their related body parts is that having adequate samples for different sides of a body part, like left and right shoulders, contributes

to better recognition of all variations. This is particularly beneficial as having ample samples for the left shoulder, for example, aids in estimating keypoints for the right shoulder as well, enhancing the model's ability to distinguish between both sides effectively.

For each group, we select the model with the highest mean detection accuracy, aiming to base our discussion on the performance of the most proficient model within each keypoints group. Subsequently, we define the visibility rate of each keypoint group as the mean ratio of visible keypoints across 4,166 training frames. The outcome of this analysis is graphically depicted in Figure 4.6.

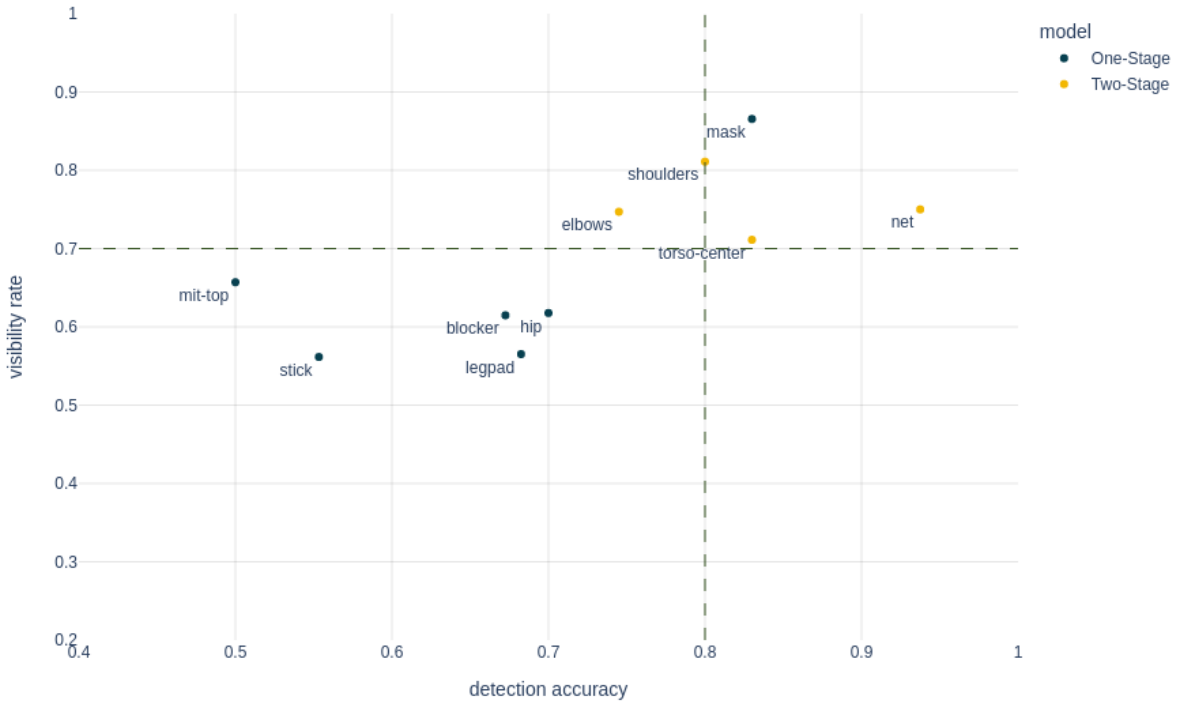

Figure 4.6: Visibility rate of each keypoints group vs the accuracy of the best performing model for that group.

In our evaluation, a detection accuracy of 80% or below is deemed suboptimal, while a visibility rate below 70% is considered limited. A visual representation of these thresholds is presented in Figure 4.6. The empirical analysis yields the discernment of three distinct clusters. The first cluster is characterized by heightened accuracy and a robust visibility

rate, encompassing keypoints groups like "mask", "net", "center of torso", and "shoulders". The second cluster exhibits lower accuracy and reduced visibility, encompassing "mit-top", "blocker", "hip", "stick", and "legpad". Lastly, the third cluster demonstrates sufficient visibility but low detection accuracy, solely comprising the "elbows" group.

While the decreased accuracy of the "elbows" group, despite a sufficient visibility rate, presents a potential area for future investigation, we hypothesize a connection between reduced accuracy and the limited volume of training data accessible for these particular keypoints, as observed in the second cluster. This correlation indicates the possibility of enhancing accuracy through the augmentation of meticulously annotated training data tailored to these specific keypoint groups. Conversely, within the first mentioned cluster, we observe that despite a lower visibility rate, the accuracy for the net keypoints remains notably high. This phenomenon can be attributed to the consistent positioning of the net, coupled with its simple rigid shape, which renders keypoint estimation less challenging.

A fundamental aspect of the MSPN[24] refinement process lies in the implementation of variable kernel sizes, facilitating a transition from coarse to fine localization. Notably, the localization requirements for different keypoints may vary significantly. For instance, keypoints located on blockers are near together and they may greatly benefit from the utilization of smaller kernels, enabling more precise localization. Recognizing the distinct needs of different keypoints and tailoring kernel sizes accordingly holds the potential to significantly enhance the accuracy and reliability of the overall estimation process. To further improve specific keypoint estimation, like legpads, the integration of priors over keypoint positions could offer better detection accuracy. This approach represents a pivotal avenue for further refinement within the architecture, accentuating the significance of tailored strategies in optimizing the performance of the Two-Stage Learning GoalieNet.

The strategic implementation of the Two-Stage Learning GoalieNet framework was driven by the quest to unravel the intricate interplay between various keypoints, particularly in the context of swiftly moving elements like the stick. A notable aspiration was to enable the network to autonomously discern the keypoints of utmost relevance for accurate estimation, thus inherently addressing the dynamic challenges posed by this task.

The utilization of extended bounding boxes within GoalieNet, which encapsulates all three entities of goalie, stick and net together, has exposed the architecture to a considerable amount of noisy background information. This complexity further underscores the significance of an adaptable framework capable of determining which keypoints warrant closer attention. In light of this, future endeavors could delve into more sophisticated network architectures carefully designed to match the changing dynamics of ice hockey.The integration of attention mechanisms, for instance, could potentially empower the network

to concentrate its processing on critical keypoints such as the stick, ultimately enhancing both accuracy and efficiency.

In summary, while our study showcases the potential of Two-Stage Learning for goalie pose estimation, there's ample room for further exploration and refinement. As ice hockey dynamics continue to evolve, advancing the accuracy of pose estimation methods will play a pivotal role in enhancing player analysis, strategy development, and performance evaluation.

Chapter 5

Conclusions

In the pursuit of enhancing the accuracy and comprehensiveness of goalie pose estimation in the context of ice hockey, this study embarked on a rigorous exploration, methodical experimentation, and systematic analysis. Through a meticulous journey of model development, training, and evaluation, we have endeavored to provide valuable insights and contributions to the field.

The journey began with the realization of the intricate challenges posed by the unique nature of goalie pose estimation. Unlike conventional human pose estimation tasks, the presence of non-human joints and the interplay between multiple entities – goalie, stick, and net – necessitated innovative approaches. The One-Stage GoalieNet is an initial architecture within the Multi Stage Pose Network [24] framework that attempts to predict all keypoints simultaneously. It tackles the challenges of estimating non-human keypoints in the dynamic context of ice hockey. While this approach provides a holistic perspective, accurately predicting certain keypoints, such as those associated with the swiftly moving stick, remains challenging.

The limitations and complexities encountered by the One-Stage approach paved the way for the development of the more specialized Two-Stage approach, comprising the Keypoint Predictor Network (KPN) and the Keyheatmap Fusion Network (KFN), which seeks to enhance accuracy by addressing the specific intricacies of each keypoint's estimation. Our Two-Stage Learning architecture, was conceived to tackle these challenges effectively. The Two-Stage Learning GoalieNet comprises two sequential stages aimed at refining the accuracy of keypoints estimation for ice hockey goaltenders. In this approach, the first stage involves the Keypoint Predictor Network (KPN), which focuses on predicting 26 out of the total 29 keypoints, excluding the keypoints associated with the stick. This initial stage

leverages the relationships among goalie, equipment, and net positions to enhance accuracy. Subsequently, the second stage introduces the Keyheatmap Fusion Network (KFN), which takes the fused information from the KPN, along with the original image, and predicts the stick keypoints. The key aspect here is the deliberate exclusion of stick-related keypoints in the first stage to allow a more precise estimation of the stick's position. This model aimed to leverage the spatial relationships between keypoints and infuse multi-modal insights for enhanced predictions.

Experimental results unveiled the nuanced interplay between accuracy and challenges. The efficacy of GoalieNet was demonstrated through comprehensive evaluations, highlighting its capacity to predict keypoints with a focus on goalie, gear, and net positions. Accuracy and localization error metrics provided an insightful overview of our model's performance, showcasing its potential across diverse classes of keypoints.

However, the journey also underscored the subtleties of annotation quality, data cleanliness, and network input selection. The varying accuracy across different keypoints underscored the importance of meticulous data annotation and model tuning. The trade-off between the specificity of keypoints' annotations and the model's accuracy became evident, suggesting the need for a delicate balance.

Looking ahead, the journey does not conclude, but rather evolves. Further refinements and extensions are essential to enhance the model's robustness. Exploring kernel sizes tailored to individual keypoints, implementing prior information, and addressing extended bounding box challenges offer avenues for future investigations. Additionally, attention mechanisms and advanced network architectures could hold the potential to further bolster our model's accuracy and adaptability.

In closing, this study has offered a comprehensive journey through the intricate realm of goalie pose estimation. As the dynamic world of sports and technology continues to advance, our pursuit of accurate and versatile pose estimation remains ever-evolving, guided by a commitment to innovation and a deeper understanding of the nuances within the sport of ice hockey.

References

[1] Sports Global Market Report 2022. https://www.prnewswire.com/news-releases/sports-global-market-report-2022-301500432.html.

[2] Top 5 Clubs that Generated Most Money from Player Sales in the Last Decade. https://notjustok.com/sports/top-5-clubs-that-generated-most-money-from-player-sales-in-the-last-decade-see-list/.

[3] Bruno Artacho and Andreas Savakis. Bapose: Bottom-up pose estimation with disentangled waterfall representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 528–537, 2023.

[4] Wenxia Bao, Tao Niu, Nian Wang, and Xianjun Yang. Pose estimation and motion analysis of ski jumpers based on eca-hrnet. *Scientific Reports*, 13(1):6132, 2023.

[5] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 717–732. Springer, 2016.

[6] 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.

[7] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[8] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[9] Chengpeng Duan, Bingliang Hu, Wei Liu, and Jie Song. Motion capture for sporting events based on graph convolutional neural networks and single target pose estimation algorithms. *Applied Sciences*, 13(13):7611, 2023.

[10] Marcin Eichner and Vittorio Ferrari. Appearance sharing for collective human pose estimation. In *Asian Conference on Computer Vision*, pages 138–151. Springer, 2012.

[11] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1347–1355, 2015.

[12] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to refine human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 205–214, 2018.

[13] Seyed Abolfazl Ghasemzadeh, Gabriel Van Zandycke, Maxime Istasse, Niels Sayez, Amirafshar Moshtaghpour, and Christophe De Vleeschouwer. Deepsportlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes. *arXiv preprint arXiv:2112.00627*, 2021.

[14] Nicola Giulietti, Alessia Caputo, Paolo Chiariotti, and Paolo Castellini. Swimmernet: Underwater 2d swimmer pose estimation exploiting fully convolutional neural networks. *Sensors*, 23(4):2364, 2023.

[15] Daniel Groos, Heri Ramampiaro, and Espen AF Ihlen. Efficientpose: Scalable single-person pose estimation. *Applied intelligence*, 51:2518–2533, 2021.

[16] Masaki Hayashi, Kyoko Oshima, Masamoto Tanabiki, and Yoshimitsu Aoki. Lower body pose estimation in team sports videos using label-grid classifier integrated with tracking-by-detection. *Information and Media Technologies*, 10(2):246–258, 2015.

[17] Masaki Hayashi, Taiki Yamamoto, Yoshimitsu Aoki, Kyoko Ohshima, and Masamoto Tanabiki. Head and upper body pose estimation in team sport videos. In *2013 2nd IAPR Asian Conference on Pattern Recognition*, pages 754–759, 2013.

[18] James Hong, Matthew Fisher, Michaël Gharbi, and Kayvon Fatahalian. Video pose distillation for few-shot, fine-grained sports action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9254–9263, 2021.

[19] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 34–50. Springer, 2016.

[20] Nourah Fahad Janbi and Nada Almuaythir. Bowlingdl: A deep learning-based bowling players pose estimation and classification. In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, pages 1–6, 2023.

[21] Zhehan Kan, Shuoshuo Chen, Ce Zhang, Yushun Tang, and Zhihai He. Self-correctable and adaptable inference for generalizable human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5537–5546, 2023.

[22] Kaleab A Kinfu and René Vidal. Efficient vision transformer for human pose estimation via patch selection. *arXiv preprint arXiv:2306.04225*, 2023.

[23] Takumi Kitamura, Hitoshi Teshima, Diego Thomas, and Hiroshi Kawasaki. Refining openpose with a new sports dataset for robust 2d pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 672–681, 2022.

[24] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[26] Tingting Liu, Hai Liu, Bing Yang, and Zhaoli Zhang. Ldcnet: Limb direction cues-aware network for flexible human pose estimation in industrial behavioral biometrics systems. *IEEE Transactions on Industrial Informatics*, pages 1–11, 2023.

[27] Katja Ludwig, Sebastian Scherer, Moritz Einfalt, and Rainer Lienhart. Self-supervised learning for human pose estimation in sports. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, 2021.

[28] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13264–13273, 2021.

[29] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022.

[30] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *European Conference on Computer Vision*, pages 72–88. Springer, 2022.

[31] 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.

[32] NHL.com. Nhl top players: Top 10 goalies. https://www.nhl.com/news/nhl-top-players-top-10-goalies/c-335454948. Accessed July 26, 2023.

[33] Christos Papaioannidis, Ioannis Mademlis, and Ioannis Pitas. Fast single-person 2d human pose estimation using multi-task convolutional neural networks. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[34] 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.

[35] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[36] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[37] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185, 2012.

[38] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017.

[39] Pranshu Sharma, Bishesh Bikram Shah, and Chandra Prakash. A pilot study on human pose estimation for sports analysis. In *Pattern Recognition and Data Analysis with Applications*, pages 533–544. Springer, 2022.

[40] Ming-Hwa Sheu, S. M. Salahuddin Morsalin, Chung-Chian Hsu, Shin-Chi Lai, Szu-Hong Wang, and Chuan-Yu Chang. Improvement of human pose estimation and processing with the intensive feature consistency network. *IEEE Access*, 11:28045–28059, 2023.

[41] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022.

[42] Zhihui Su, Ming Ye, Guohui Zhang, Lei Dai, and Jianda Sheng. Cascade feature aggregation for human pose estimation. *arXiv preprint arXiv:1902.07837*, 2019.

[43] 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.

[44] Yuandong Tian, C Lawrence Zitnick, and Srinivasa G Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 256–269. Springer, 2012.

[45] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[46] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In

Proceedings of the 27th ACM international conference on multimedia, pages 374–382, 2019.

[47] Yihan Wang, Muyang Li, Han Cai, Wei-Ming Chen, and Song Han. Lite pose: Efficient architecture design for 2d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2022.

[48] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[49] Chuanlei Zhang, Lixin Liu, Qihuai Xiang, Jianrong Li, and Xuefei Ren. Sports pose estimation based on lstm and attention mechanism. In *6GN for Future Wireless Networks: Third EAI International Conference, 6GN 2020, Tianjin, China, August 15-16, 2020, Proceedings 3*, pages 538–550. Springer, 2020.

[50] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020.

[51] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[52] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760*, 2019.

[53] Zhe Zhang, Jie Tang, and Gangshan Wu. Simple and lightweight human pose estimation. *arXiv preprint arXiv:1911.10346*, 2019.