

Broadcast is all you need: Robust Multiplayer Tracking in Ice Hockey using Monocular Videos

by

Harish Prakash

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, 2024

© Harish Prakash 2024

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

This thesis contains material from two papers for which I was the lead author. As the lead author, I was responsible for the conceptualization, writing code as well as writing and submitting manuscripts

1. **Prakash, H.**, Shang, J. C., Nsiempba, K. M., Chen, Y., Clausi, D. A., & Zelek, J. S. (2024), "Multi Player Tracking in Ice Hockey with Homographic Projections", Proceedings of the Conference on Robots and Vision.
2. **Prakash, H.**, Chen, Y., Rambhatla, S., Clausi, D. A., & Zelek, J. (2024), "VIP-HTD : A Public Benchmark for Multi-Player Tracking in Ice Hockey", Journal of Computational Vision and Imaging Systems, 9(1), 22–25.

Additionally, I also co-authored two research papers in the areas of computer vision-based sports analytics and 3D human mesh recovery.

1. Balaji, B., Bright, J., **Prakash, H.**, Chen, Y., Clausi, D. A., & Zelek, J. S. (2023), "Jersey Number Recognition using Keyframe Identification from Low-Resolution Broadcast Videos", Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports.
2. Bright, J., Balaji, B., **Prakash, H.**, Chen, Y., Clausi, D. A., & Zelek, J. S. (2024), "Distribution and Depth-Aware Transformers for 3D Human Mesh Recovery", Proceedings of the Conference on Robots and Vision.

Abstract

Multi-object tracking (MOT) in ice hockey pursues the combined task of detecting and associating players across a given sequence to maintain their identities. Tracking players in sports using monocular broadcast videos is an important computer vision problem that enables several downstream analytics and enhances viewership experience. However, existing tracking approaches encounter significant challenges in dealing with occlusions, blurs, camera pan-tilt-zoom effects, and dynamic player movements prevalent in telecast feeds. These challenges are further exacerbated in fast-paced sports such as ice hockey, where existing trackers struggle to maintain identity consistency due to players’ sudden, non-linear motion patterns.

In this thesis, acknowledging the fundamental role of quality datasets, we first present two hockey tracking datasets: our previously developed hockey tracking dataset - 1 (HTD-1) and a newly curated, open-source dataset called hockey tracking dataset - 2 (HTD-2), annotated from broadcast National Hockey League (NHL) games. Based on this new dataset, we establish a reference benchmark by evaluating six state-of-the-art (SOTA) tracking methods to enable performance comparisons in hockey MOT. A detailed study is conducted for each algorithm to understand their merits and drawbacks on tracking players.

Next, to address the present limitations, we propose a novel tracking model formulating MOT as a bipartite graph matching problem cued with homography inputs. Specifically, we disambiguate the positional representation of occluded players as viewed through broadcast footage, by warping them onto a view-invariant overhead rink template and encode their transformations into the graph message passing network. This ensures reliable spatial context for identity-preserved track prediction.

Experimental results demonstrate that our model achieves a **10** times reduction in identity switch score (IDsw) and a **32.45%** improvement in identify F1 score (IDF1) score compared to the existing baseline on HTD-1, establishing a new SOTA. The proposed model also exhibits strong generalization capabilities, achieving **92.8%** IDF1 and only **60** IDsw during cross-validation on HTD-2. Finally, ablation studies are presented to validate our performance and substantiate our approach. s

Acknowledgements

First and foremost, I would like to begin by thanking god for giving me the strength, resilience, and determination to embark on this journey away from home for the first time, and pursue this research to its completion.

When I came to Canada in the fall of 2022 with a head full of dreams, it was Dr. David Clausi who recognized my potential and nurtured me. His words, “If you are not enjoying what you do, you are not doing the right thing,” have been my guiding light during moments of self-doubt. Along with him, Dr. John Zelek has been an incredible guide and mentor throughout this work. I couldn’t have asked for better supervisors to shape my research journey.

The enriching conversations with Dr. Alexander Wong, Dr. Yuhao Chen, and Dr. Sirisha Rambhatla have been instrumental in my growth as a researcher. Their willingness to share their knowledge and time has inspired me to push boundaries and aim higher. My gratitude to Dr. Charbel Azzi and Dr. Elliot Creager for their valuable feedback in reviewing this work, despite their busy schedule.

To my amazing labmates – Jerrin, Bavesh, Yuhao, and Chris – thank you for being my sounding board, collaborators, and friends. You’ve made coming to work something to look forward to every day. I am thankful to the entire SARG team – Jason, Liam, Ken, and Mehrnaz, for always willing to collaborate.

Sudhan, you’ve been more than a friend – you’re the brother I found in Canada. Our 307 Hazelglen days together have given me memories I’ll cherish forever.

To my support system – Arvind, Vignesh, Sanjay, Aakash, Adarsh, Darrsan, Sriram, and Naveen – thank you for always being just a call away, especially during tough times. I’m the luckiest man to call you all my friends. Special thanks to Shriram anna and Abi bro for your thoughtful advices on life and work. I cherish to have you both.

I am blessed beyond measure to have my parents, sister, and grandparents by my side throughout my life. Without their kindness and support, I could not have travelled this far. You are my foundation and my strength. Thank you.

Finally, I extend my gratitude to Stathletes for funding this research and the University of Waterloo for providing an environment where I could thrive and grow.

Dedication

I dedicate this thesis to my family without whom I would not be here today. I am what I am because of their unconditional love and support.

To Appa, for your tireless efforts to provide me with the best of everything. You are my first and foremost inspiration.

To Amma, for all the sacrifices you've made to mold me into the best version of myself. Thank you for teaching me kindness, gentleness, and humility.

To Cheenu thatha, for selflessly fulfilling every wish I've had. Thank you for being a pillar of support for our family to this very day.

To Janardhanan thatha, for nurturing my curiosity and igniting in me a passion for lifelong learning. Watching you voraciously read every day made me inquisitive about this world.

To Padma paati, for all the beautiful stories you narrated and the wisdom you imparted. I shall always be your *kanna*.

To Vasantha paati, for laying the foundations that our family firmly stands on today. Though our time together was brief, your impact has been profound.

To Keerthu, for being my partner in crime. You are the best sister I could ask for.

I am truly grateful to have you all. Love you.

Table of Contents

Author’s Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	xi
List of Tables	xiii
List of Abbreviations	xiv
1 Introduction	1
1.1 Motivation	2
1.2 Challenges	3
1.3 Problem Formulation	4
1.4 Contributions	8
1.5 Thesis Structure	8

2	Background	10
2.1	Multi-Object Tracking	10
2.2	Joint Detection and Tracking	11
2.3	Tracking-by-Detection	12
2.3.1	Detection	12
2.3.2	Re-identification	13
2.4	MOT for Sports	13
2.5	Tracking for Team Sports	14
2.6	Tracking for Ice Hockey	15
2.7	Discussion	16
3	Datasets and Evaluation Metrics	17
3.1	Introduction	17
3.2	Hockey Tracking Dataset - 1 (HTD-1)	17
3.3	Hockey Tracking Dataset - 2 (HTD-2)	18
3.3.1	Metadata	19
3.3.2	Frame extraction methodology	23
3.3.3	Identity assignment and annotation strategy	23
3.4	Evaluation Metrics	24
3.5	Conclusion	26
4	Public Tracking Benchmark	27
4.1	Introduction	27
4.2	Tracking Benchmark: Motivation	27
4.2.1	Simple online and realtime tracking (SORT)	28
4.2.2	Observation-centric SORT (OC-SORT)	28
4.2.3	Tracktor	28
4.2.4	ByteTrack	29

4.2.5	HockeyMOT	29
4.2.6	Trackformer	31
4.3	Comparison Study	33
4.3.1	Implementation details	33
4.3.2	Quantitative results	33
4.4	Conclusion	36
5	Part I: Multi-object Tracking Pipeline	37
5.1	Introduction	37
5.2	Detection	39
5.3	Re-identification (reID)	41
5.4	Homography	44
5.5	Conclusion	46
6	Part II: Multi-object Tracking Pipeline	47
6.1	Introduction	47
6.2	Proposed Method	51
6.2.1	Overview	51
6.2.2	Graph formulation	51
6.2.3	Temporal graph association	52
6.2.4	Message passing network	54
6.2.5	Post-processing	57
6.3	Implementation Details	58
6.3.1	Datasets	58
6.3.2	Training details	58
6.4	Experimental Results	60
6.5	Ablation Studies	62
6.5.1	Effect of homography	62

6.5.2	Number of message passing steps, L	64
6.5.3	Video-wise quantitative results	65
6.5.4	t-SNE visualization	66
6.6	Summary	66
7	Conclusion and Future Research	67
7.1	Summary of Contributions	67
7.2	Future Research Directions	68
7.2.1	Unique player identification	68
7.2.2	Multi-view tracking	68
7.2.3	Action recognition	69
7.2.4	Trajectory imputation	69
7.3	Thesis Applicability	69
7.4	Thesis Impact	70
	References	71

List of Figures

1.1	Example of an identity-preserved tracklet	5
1.2	Problem formulation for MOT.	6
3.1	Duration of each clip in HTD-1.	18
3.2	Inaccurate annotations in McGill hockey player tracking dataset (MHPTD)	19
3.3	Duration of each clip in HTD-2.	20
3.4	HTD-2 player instances and frames	22
4.1	HockeyMOT architecture	30
4.2	Trackformer architecture	32
5.1	Overall tracking pipeline	38
5.2	YOLOX architecture	39
5.3	Detection results on the HTD-2 test-set	40
5.4	Architecture of OSNet adapted for hockey player re-identification.	41
5.5	t-SNE visualization of reID player features.	42
5.6	Convolutional feature activation maps from the last layer of OSNet	43
5.7	Warping broadcast frames onto the overhead rink template using homography.	45
6.1	Eliminating identity switches.	48
6.2	What makes our approach novel?	50
6.3	Proposed tracking architecture	53

6.4	Number of IDsw incurred vs. the number of message passing steps, L . . .	64
6.5	t-SNE visualization for the temporal graph association stage.	66

List of Tables

3.1	Metadata for each curated clip from HTD-2	21
4.1	Public reference benchmark on HTD-2	34
4.2	Video-wise IDsw benchmark on HTD-2	34
4.3	Video-wise IDF1 benchmark on HTD-2.	35
6.1	Feature encoder and message passing networks (MPN) layer details	56
6.2	Evaluation results on HTD-1	60
6.3	Cross-dataset validation on HTD-2.	61
6.4	Effect of including homography cues on IDsw and IDF1 scores	63
6.5	Effect of message passing steps on IDsw and IDF1 scores	64
6.6	Sequence-wise results on HTD-1 test set	65

List of Abbreviations

MOT Multi-object tracking

VIP-HTD VIP - hockey tracking dataset

SOTA state-of-the-art

IDsw identity switch score

IDF1 identify F1 score

NHL National Hockey League

NBA National Basketball Association

VR virtual reality

ML machine learning

AI artificial intelligence

IR infrared radiation

reID Re-identification

FoV field-of-view

GNN graphical neural networks

CVAT computer vision annotation tool

MHPTD McGill hockey player tracking dataset

MOTA multi-object tracking accuracy

FP false positive
FN false negative
TP true positives
CNN convolutional neural networks
COCO common objects in context
RoI region of interest
IoU intersection over union
MPN message passing networks
MLP multi-layer perceptron
MHSA multi-head self attention
FFN feed forward network
i.i.d independent and identical distribution
FLOP floating-point operations per second
TBD Tracking-by-detection
JDT Joint detection and tracking
RPN region proposal network
YOLO You Only Look Once
HTD-2 hockey tracking dataset - 2
HTD-1 hockey tracking dataset - 1
PoV point-of-view
RNN recurrent neural network

Chapter 1

Introduction

Consider this scenario – you are given the responsibility to manage your favorite ice hockey team from the [NHL](#) for the 2024-2025 season. This role is a major responsibility with high stakes – 5.5 million viewers [\[86\]](#), \$100 million in operation costs [\[87\]](#), and multiple other actors are directly or indirectly affected by your decisions. To begin with, there are a set of questions that naturally arise: “How has the team performed in the last few seasons?”, “Who are the strongest and weakest players?”, “Are we built around offense, defense, speed, physicality, or a mix of these elements?”, etc. Next, there are specific questions to answer, such as “What is the contribution of player X?”, “How long can each player effectively stay on the ice before their performance begins to drop?”, “Do we need fresh legs for a final push or to hold off a late-game surge from the opponent?” and likewise. Finding the *correct* answers is paramount to your team’s success.

The best possible strategy here is to thoroughly analyze past patterns (data analysis) to make informed decisions. This is formally known as *sports analytics*. In the context of ice hockey, analytics helps optimize player performance, refine team strategies, and make informed choices in scouting and recruitment. By following a data-driven approach, it helps go beyond what is possible through traditional observation, enhancing game tactics, injury prevention, event fan engagement, and business operations to answer the most pressing questions. In recent times, ice hockey has hugely benefited from a wave of new technology [\[88\]](#), including virtual reality ([VR](#)) to train players (mainly, injury prevention, and player safety), advanced statistics to enhance viewership engagement, and sensor-embedded equipments for monitoring game play. Furthermore, the advancing capabilities of artificial intelligence ([AI](#)) and machine learning ([ML](#)) holds immense potential [\[89\]](#) to revolutionize the entire game experience.

In this direction, a crucial aspect of modern hockey analytics is the continuous tracking of players on the ice, as accurately identifying player locations during each moment of the game is a fundamental requirement for several downstream tasks, such as players’ performance analysis, action recognition, injury prevention, tactical analysis, and shift detection¹. A significant milestone in tracking came with the [NHL](#)’s introduction of player and puck tracking technologies during the 2021-2022 season, with infrared radiation ([IR](#)) emitters embedded in puck and player jerseys, and 18-20 [IR](#) detectors around the rink [90]. However, there exists important limitations in this approach such as: (i) high installation and maintenance costs, (ii) player resistance to wearing equipment that tracks them continuously, and (iii) inability of this technology to model advanced tasks like player action interpretation.

1.1 Motivation

Given these present setbacks, we ask: *Can we utilize readily available broadcast hockey videos to overcome the present limitations with sensor-based tracking?* Before answering this question, it is first necessary to understand *why* this question is being raised.

After each game, [NHL](#) teams conduct post-match analysis, assessing their strengths and weaknesses to gain tactical and strategic insights. At present, significant time and effort is required to scrutinize all three periods of play time (60 minutes total) for each game to generate meaningful information. This is an extremely cumbersome task, especially considering the limited window within consecutive games during an ongoing season. In addition, human analysis has certain inherent caveats including subjectivity and bias, incorrect interpretations (accidental), varying attention span, and other unforeseen errors. But, there exists significant potential for automation in this area, with the fundamental first step being the ability to track all players on the ice at each instant, formally known as multi-object tracking ([MOT](#)). This is, therefore, the core focus of this work.

Currently, the sensor-based tracking technology used by [NHL](#) incurs significant setup and maintenance costs [91], along with other operational overheads. The raw positional data from these sensors should be manually post-processed to derive actionable insights, requiring large computational infrastructure [92]. This deepens the divide between [NHL](#) and other junior hockey leagues, which are unable to afford such expensive technologies for performance insights. In the long run, this disparity is problematic, as players advanc-

¹Here, shift detection refers to the rapid and spontaneous substitution of multiple players during game-play, commonly seen in ice hockey games.

ing from junior leagues do so without comprehensive performance insights during their developmental phases.

Therefore, developing an automated, cost-effective solution to consistently track players is an important research objective to pursue. A computer vision-based MOT framework for ice hockey using standard broadcast videos could enable several downstream tasks including data-driven performance analysis and strategic decision-making, while significantly reducing the manual effort required for player tracking and game analysis. The objective of this thesis thus, is to develop a comprehensive pipeline for tracking ice hockey players from broadcast video feeds.

1.2 Challenges

While the potential benefits of automated player tracking are significant, there exists several technical challenges that arise from both the dynamic nature of the sport and limitations of monocular broadcast videos. These include:

1. **Occlusion:** A characteristic feature of hockey is the presence of multiple players in close proximity, competing for possession of the puck. From a monocular broadcast point-of-view (PoV), players are often visually obstructed either by other players or dasher boards on the rink. These occlusions significantly impact the reliability of computer vision algorithms in accurate player localization and maintaining player identities.
2. **Dynamic player movements:** The highly dynamic nature of ice hockey sets it apart from other team sports like soccer, basketball, and volleyball, where player positions tend to be more structured. Players exhibit rapid, spontaneous movements across the rink, making conventional motion tracking assumptions insufficient. Therefore, ice hockey necessitates tracking algorithms to model complex motion patterns.
3. **Camera motion and resolution:** Broadcast cameras are subjected to pan-tilt-zoom effects to capture the gameplay and present an interesting telecast for the audience. This leads to rapid camera movements and sudden scene shifts when switching between multiple angles, leading in blurs and inconsistent player coverage. Further, the standard resolution of broadcast feeds (1280×720 p) constrains the ability to clearly identify and track individual players, particularly during fast-paced action

sequences where a higher frame resolution would provide fine-grained features for player identification and tracking.

4. **Lack of distinguishable features:** Ice hockey players wear bulky protective equipment (gloves, pads, guards, helmets) that covers most of their bodies. This poses significant identification challenges, as the protective gear obscures distinguishing physical characteristics, making players on the same team appear nearly identical and limiting the visibility of traditional identifiers like jersey numbers.
5. **Need for reproducible benchmarks:** Benchmarks serve as a reference for future research progress and advancing the [SOTA](#). They enable fair comparison between different approaches, identify current limitations, and establish clear metrics for measuring improvements in tracking performance. But, currently there exists no *reproducible* benchmark with recent [SOTA](#) advancements evaluated on ice hockey. Without such benchmarks, quantifying progress and validating new methodologies becomes challenging.

1.3 Problem Formulation

In light of the aforementioned challenges, this thesis focuses on developing a robust methodology to consistently track players from broadcast video feeds while preserving their identities. Following current [MOT](#) conventions, we treat each player as an *object* to be tracked. The term *players* henceforth, also encompasses referees and goalies, as they actively influence the tracking of team members within the rink. To proceed further, we define the following important terms:

1. *Identity-preserved tracklets:* A tracklet $\mathcal{T}_P$ is a temporal sequence of associations belonging to player P with a unique identity throughout the tracklet’s duration. However, when players cross paths or overlap (occlusions), their trajectories intersect causing identity switches – errors where $\mathcal{T}_P$ incorrectly contains detections from multiple players with different IDs (Fig. [6.1](#)). Our goal is to minimize these identity switches while tracking players, given that they remain *detectable* during occlusions.
2. *Consistent tracking:* In hockey [MOT](#), we define *consistent tracking* as maintaining a player’s identity throughout their tracklet predictions while they remain both *detectable* and *available* within the broadcast camera’s field-of-view ([FoV](#)). If a player becomes either undetectable or leaves the [FoV](#), we end their current tracklet. When



Figure 1.1: Example of an identity-preserved tracklet, $\mathcal{T}_v$. Here, the track ID (eight) in the top right corner is consistent for the entire tracklet and does not correspond to the player jersey number. Frames shown are sampled at varying intervals to demonstrate the tracklet’s progression.

they reappear, we create a new tracklet with a new ID rather than attempting to connect the gaps (Fig. 1.2). This helps prevent misidentification errors that could arise from trying to interpolate player positions or using longer temporal relationships, since the duration of a player’s absence is non-deterministic.

Given the above definitions, consider a broadcast ice hockey sequence S_t parsed into frames, $S_t = \{f^i \mid i = 1, 2, \dots, n\}$ where $n = \frac{t}{\Delta t}$ denotes the total number of frames; t denotes the total duration of the sequence (in seconds), and Δt represents the duration per frame (in seconds). We define the following three modules which forms the basis of our approach:

Definition 1.1 (Detection) For each frame f^i assuming that there exists at least one player $\{P_j^i \mid j = 1, 2, \dots, m\}$, where m represents the number of players, the problem of detection (localization) deals with finding $P_{j=1:m}^{i=1:n}$ players across all n frames, deriving the following information for each instance:

$$\tilde{\mathcal{D}} = \{f^i, b_j^i, c\} \quad (1.1)$$

where f^i represents frame number (ID), $b_j^i = \{x, y, wd, ht\}$ defines the bounding box coordinates for the j^{th} player in the i^{th} frame, and c denotes the detection confidence score (probability). When there are no players to localize on the ice ($P_j \mid j = 0$), those frames are skipped since $\tilde{\mathcal{D}} = \{\phi_i\}$ (empty-set).

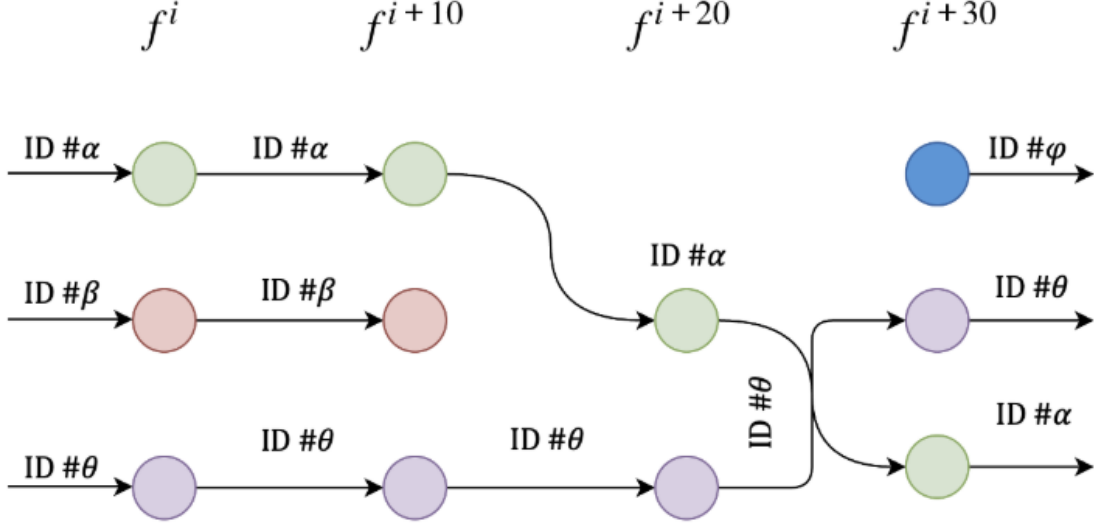


Figure 1.2: MOT for ice hockey. This example shows four player tracks (IDs: $\alpha, \beta, \theta, \varphi$) across frames f^i to f^{i+30} , sampled every 10 frames. It demonstrates two key tracking scenarios we handle: (i) Maintaining IDs for players α and θ when their paths intersect (f^{i+20} to f^{i+30}) given that they are detectable and within the camera FoV, (ii) Handling player β 's disappearance (undetectable or out-of-FoV) at f^{i+10} by terminating their tracklet and assigning a new ID (φ) upon reappearance at f^{i+30} to avoid misidentification errors.

Definition 1.2 (Re-identification). Given the bounding box coordinates for player P_j^i in frame f^i , $b_j^i = \{x_{min}, y_{min}, \text{width}, \text{height}\}$, where $b_j^i \in \mathbb{R}^4$, x_{min} and y_{min} represent the top-left corner of the bounding box, and width and height denote its horizontal and vertical extents, respectively. These define the RGB input with fixed downsampling, $\mathcal{I}_c \in \mathbb{R}^{256 \times 128 \times 3}$, which is passed into the reID network for encoding player appearance features into a 512-dimensional embedding for identity matching (association) with the closest target in frame f^{i+1} , based on a combination of cosine and L_2 distance metrics.

This is defined as:

$$\mathcal{F}_j^i = \text{reID}(\mathcal{I}_c) \quad (1.2)$$

where $\mathcal{F}_j^i \in \mathbb{R}^{512}$. In monocular tracking, this association happens *frame-by-frame* for all

players throughout the entire sequence. **reID** networks are typically convolutional neural networks (**CNN**) trained with cross-entropy [121], contrastive [34], or triplet [82] loss functions. These networks are pivotal in discerning appearance features between different players and helping with identity-preservation throughout a tracklet’s duration.

Definition 1.3 (Tracking). Given the player associations across a sequence of k frames, where player tracklet $\mathcal{T}_P = \{P^1, P^2, \dots, P^k \mid k \leq n\}$ encompasses k detections of player P , tracking involves the assignment of unique ID for the given tracklet such that, each tracklet represents one unique player throughout its duration.

$$\text{Tracking} = \{f^i, \mathbf{t}^{\text{id}}, b_j^i, c\} \quad (1.3)$$

where $\mathbf{t}^{\text{id}}$ denotes the player ID (track ID) and c denotes the detection confidence score. For a sequence of n frames containing m players, the set of all tracklets $\mathcal{T} = \bigcup_{j=1}^m \bigcup_{i=1}^n \{\mathcal{T}_j^i\}$, where each tracklet $\mathcal{T}_j$ corresponds to exactly one unique player and spans at most n frames. This formulation follows the definitions of consistent tracking and identity-preserved tracklets (defined in 1.3).

Problem Statement. Consolidating the above three tasks of player detection ($\tilde{\mathcal{D}}$), **reID** ($\mathcal{F}$), and tracking ($\mathcal{T}$), our core objective is to consistently track hockey players by minimizing identity switches, particularly during occlusion events. The problem can be formally expressed as finding the mapping:

$$\mathcal{M} : S_t \rightarrow \mathcal{T} \quad (1.4)$$

where the mapping $\mathcal{M}$ transforms the input sequence S_t into a set of consistent player tracklets $\mathcal{T}$, while maintaining temporal continuity and preserving tracklet identities across the sequence duration.

1.4 Contributions

This thesis contributes the following to the objective of MOT in ice hockey:

1. A novel MOT approach formulated as a bipartite graph matching problem cues with homography inputs. It utilizes graphical neural networks (GNN) to outperform all existing approaches, especially in reducing IDsw (defined in Subsection 3.4) during occlusions and blurs. Ablation studies are presented to substantiate its SOTA performance (Section 6).
2. A newly curated *open-source* tracking dataset titled HTD-2² is introduced. This dataset is sampled from eight broadcast NHL games to extend upon the existing HTD-1. A detailed metadata analysis is presented along with a discussion on relevant evaluation metrics to follow (Section 3).
3. The first *public* MOT reference benchmark for ice hockey, featuring evaluations on six different tracking algorithms, and a thorough analysis of their merits and drawbacks in our context (Section 4).

1.5 Thesis Structure

This thesis is structured as follows:

- Chapter 2 provides the background literature on MOT, its paradigms, application to other team sports and ice hockey.
- Chapter 3 describes the existing hockey tracking dataset - 1 (HTD-1) and introduces the *newly curated* hockey tracking dataset - 2 (HTD-2), and presents the relevant MOT evaluation metrics to be used.
- Chapter 4 establishes a new hockey *public* reference benchmark based on HTD-2, using six different tracking methods, their fine-tuning results, and a study of their performances.
- Chapter 5 details the first three components of our tracking pipeline namely detection, reID, and homography. A study on their architecture, choices, implementation details, and results are also presented.

²<https://github.com/harshap-ai/VIP-HTD>

- Chapter 6 explains our graph-based tracking model – its architecture, design choices, loss function, implementation details, results and ablations.
- Finally, Chapter 7 summarizes the contributions of this thesis and outlines potential application areas and future work directions.

Chapter 2

Background

This chapter is organized as follows: first, we introduce multi-object tracking **MOT** (section 2.1) and its two paradigms – joint detection and tracking (Joint detection and tracking (**JDT**)) and tracking-by-detection (Tracking-by-detection (**TBD**)). Following a review of the **JDT** (section 2.2) paradigm, we detail the **TBD** paradigm (section 2.3), examining detection models (subsection 2.3.1) and **reID** networks (subsection 2.3.2). We then review **MOT** for team sports (section 2.5), discussing relevant literature from soccer, basketball, volleyball, and handball, followed by ice hockey-specific **MOT** methods (section 2.6). We conclude this chapter with an analysis of current methodological gaps and our proposed direction (section 2.7).

2.1 Multi-Object Tracking

MOT is an important computer vision challenge subsuming the tasks of object localization, **reID**, association and tracking across a temporal sequence. It is a highly studied and established problem due to its plethora of applications in robotics, autonomous vehicles, industrial automation, surveillance, and sports. Current **SOTA MOT** methods are mostly focused on dealing with crowded pedestrians [22, 47, 65], group dancing [94] and autonomous driving [29] scenarios, which pose significant challenges due to high-density scene interactions and frequent occlusions.

MOT has two major paradigms: (i) Joint detection and tracking (**JDT**) [5, 59, 69, 109, 120], where a detector is converted into a tracker by predicting the location of bounding boxes in adjacent frames autoregressively; and (ii) Tracking-by-detection (**TBD**) [7, 9, 55,

111, 119, 120] which consists of four steps – detection, feature extraction, data association and identity assignment. Conventionally, the first two stages are off-the-shelf models for detection and reID borrowed from existing literature, while the last two stages of association and identity assignment are novel and represent the core tracking objective. Note that, we discuss the first two stages as subsections within the TBD section and discuss tracking as separate sections.

2.2 Joint Detection and Tracking

In this relatively new paradigm, an object detector is converted into a tracker by autoregressively estimating the positions of candidate bounding box locations in adjacent frames. This paradigm is reasoned to explore potential model blocks that can be shared between the detection and association steps to increase runtime efficiencies and reduce model complexities. Wang *et al.* [109] propose a shared MOT system that allows target detection and appearance embeddings to be jointly optimized using a multi-task loss function. To address the optimization bias between detection and re-identification tasks, Zhang *et al.* [120] introduce FairMOT, which builds upon CenterNet’s [125] anchor-free architecture to achieve balanced training of both networks.

Tracktor [5] presents an alternative by taking frame f^t as input and regresses it into frame f^{t+1} using Faster RCNN-based [76] bounding box detections, while utilizing a separate Siamese network [37] for reID feature extraction. CTracker [69] proposes an end-to-end, chained network approach where the detector, reID and inference (tracker) steps are combined into a single step, regressing target bounding boxes using two adjacent frames at a time. RetinaTrack [59] modifies the RetinaNet [80] detector by splitting the post-FPN computational pathways into different anchor shapes for successful predictions during occlusions.

The main advantage with these JDT methods is their high inference speeds, joint optimization capabilities, and their lightweight and fast detector setups. This enables them to combine detection, association, and tracking into an end-to-end trainable model using a multi-task loss function. Nevertheless, these methods face several challenges: they introduce increased architectural complexity due to condensing the tracking pipeline into one model, struggle with balancing multiple competing loss objectives, suffer from error propagation in identity management due to biased joint optimization, and exhibit limited adaptability since they are designed as monolithic, end-to-end frameworks.

2.3 Tracking-by-Detection

The [TBD](#) paradigm contains the following three modules: Detection, Re-identification, and Tracking.

2.3.1 Detection

Localizing target objects in the given frame is the fundamental first step towards [MOT](#) via the [TBD](#) paradigm. Early object detection methods such as SIFT [56] and HOG [21] focused more on extracting distinctive local features from images. SIFT [56] identified key-points invariant to scale and rotation, creating descriptors that allowed matching against known objects through gradient patterns. HOG [21] improved upon SIFT, by computing gradients across localized image regions to create histograms of orientations, effectively capturing shape information for object detection. But, these techniques rely on handcrafted features such as explicitly defined sliding windows on images (hyperparameter), which limits their adaptability to diverse and complex scenes, variations in environment (lighting, scale, resolution), occlusions, and non-rigid object transformations predominantly found in sports.

With the advent of modern deep learning methods [8, 27, 31, 49, 73, 74, 76, 105, 122], there has been significant advancements in object detection both in terms of speed and accuracy. Fast R-CNN [31] works by proposing multiple region of interest ([RoI](#)) inputs to a fully [CNN](#), where each [RoI](#) is max pooled into a fixed-size feature map and encoded into feature embeddings by a fully-connected multi-layer perceptron ([MLP](#)). The output consists of both softmax probabilities and per-class bounding boxes for each [RoI](#). But, the limitation with Fast R-CNN is its reliance on external region proposal methods (selective search) creating a computational bottleneck. Faster R-CNN [76] addresses this by introducing the region proposal network ([RPN](#)) that shares convolutional features with the detection network, enabling end-to-end training and significantly faster inference.

You Only Look Once ([YOLO](#)) [97] family of algorithms differ from [RPN](#)-based methods by first directly regressing the bounding box coordinates and pursuing classification thereafter. In general, [RPN](#)-based methods handle various object scales due to their variable anchor box sizes, leading to a better accuracy, while the [YOLO](#) family tends to be faster due to their single-stage architecture predicting bounding boxes and class probabilities in one forward pass. YOLOX [122], a variant of the [YOLO](#) family strikes a balance through an anchor-free detection approach. This is carried out by decoupling classification and regression heads using two parallel branches, leading to both better inference speeds and

omni-scale detection accuracy. This makes it a highly suitable choice for current sports-based tracking methods [?, 20, 26, 33, 101]. A more detailed explanation of YOLOX [122] and its suitability for ice hockey is presented in Chapter 5.

2.3.2 Re-identification

The second step in TBD paradigm is re-identification (reID), which accomplishes the task of extracting discriminative appearance features from two consecutive input frames F_i and F_{i+1} to find matching correspondences. In the context of monocular tracking, it is framed as an instance-level recognition problem that helps associate similar objects *frame-by-frame* temporally across the given sequence. reID networks typically use CNN backbones for feature extraction, trained with triplet [82], contrastive [34], or cross-entropy [121] loss functions for contrastive learning. DeepReID [50] proposes a filter pairing CNN that encodes photometric and geometric transforms for cross-view associations. Varior *et al.* [100] present a Siamese CNN trained with contrastive loss [34] which clusters similar feature embeddings together and dissimilar features apart. BraidNet [106] performs a cross-summation of the intermediate features from source and target inputs, to extract comparison features between the input pairs. This approach is specifically aimed at being robust to misalignments and color differences. However, a major drawback with all these approaches is their fixed-size convolutional kernels and rigid receptive fields that do not effectively capture discriminative features when the same object appears at significantly different scales across the sequence.

The ability to learn features at multiple scales is an irrefutable advantage, especially for our use case, where broadcast cameras are subjected to pan-tilt-zoom effects to capture interesting moments. OSNet [124] propose CNNs with residual blocks composed of multiple convolutional streams, each learning features at different scales. Also called as *omni-scale* learning, the learned features are fused together through a gating mechanism, leading to a light-weight and SOTA reID model. A more detailed explanation of OSNet [124] and its suitability for ice hockey is presented in Chapter 5.

2.4 MOT for Sports

Sports tracking is a pivotal task in vision due to its numerous downstream applications in game analytics and statistics, strategic planning [61], performance evaluations, injury prevention, and event detection. Automated player tracking reduces the need for manual

labor and human effort by streamlining game analysis and player performance evaluations. However, naively extending methods originally designed for tracking in other domains (pedestrians, vehicles) to team sports is a non-trivial task, due to the domain-specific challenges in modeling dynamic player movements, occlusions, fast broadcast camera motion, and perspective projection errors. Several important steps have been taken with the advent of deep networks [44], to adapt trackers for various team sports including soccer [40, 66, 98, 112, 115], handball [10], basketball [1, 3, 57, 117], and volleyball [96]. Public MOT sports datasets [19, 20] for the aforementioned sports have also been developed to support principled evaluations.

2.5 Tracking for Team Sports

Initial methods in soccer tracking utilized Kalman filters [41] to track players and balls. Najafzadeh *et al.* [66] estimate multiple player trajectories in soccer using Kalman filters with fixed number of player targets at any given time. Yu *et al.* [115] track soccer balls utilizing a localized Kalman filter-based matching procedure to extend predicted trajectories during temporary occlusions. The drawback with using Kalman filters though is their *linear motion model assumption* which assumes trajectory paths to be linear, unlike commonly observed in sports. Theagarajan *et al.* [98] use the DeepSORT algorithm [111] to additionally include player reID features for soccer tracking. They use a YOLO-v2 [74] detector to localize players and focus their approach on tactical analysis and soccer ball possession statistics. However, YOLO-v2 [74] has been surpassed by many recent developments [97] while the quality of features from DeepSORT [111] is sensitive to dynamic scene (lighting, camera movements, scale) changes in broadcast. Iwase *et al.* [40] and Xu *et al.* [112] propose tracking players using multiple static cameras by warping them onto a common soccer field template using homography, to handle player occlusions. But, this is only applicable in scenarios where multiple camera feeds are available.

In basketball, initial attempts by Acuna *et al.* [1] use SORT [7], which utilizes YOLO-v2 [74] detections combined Kalman filters [41] to track player trajectories. Sanguesa *et al.* [3] combine convolutional and geometric (pose) features from a VGG network [85] to track players on the court but conclude that their method’s performance is negatively impacted when players are occluded. Lu *et al.* [57] track players using Kalman filters [41] with the underlying assumption that the relationship between time and players’ locations is approximately linear in a short time interval i.e., players always move in a straight line within a short interval of time, which does not apply for fast-paced sports. To overcome these limitations, Zhang *et al.* [117] propose to track occluded players in a multi-camera

setting, which is not readily available in all team sports. Kipf *et al.* [43] predict future player and ball trajectories during ‘Pick and Roll’ instances (common offensive tactic in National Basketball Association (NBA) games) using GNNs, with a CNN-based encoder and a recurrent neural network (RNN)-based decoder. However, their approach directly utilize the location information of players and the ball for each frame, which is unavailable for televised broadcast videos.

Mauthner *et al.* [63] use particle filters to track beach volleyball players from a single fixed camera and use player color cues to declutter occlusion scenarios. Buric *et al.* [10] evaluate SORT [7], DeepSORT [111], and Hungarian algorithm [45] to track handball players captured in indoor environments specifically with artificial illumination.

2.6 Tracking for Ice Hockey

Unlike other team sports, ice hockey presents significant and unique challenges for MOT, as discussed in Section 1.2. Early attempts at tracking ice hockey players were pursued using an ensemble of handcrafted methods. Okuma *et al.* [68] use Adaboost [104] detections with mixed particle filters (MPF) [103] for tracking players from television videos. The MPF framework [103] extends traditional particle filtering by representing posterior state distributions of n objects as n component mixtures, enabling multi-modal (here, it refers to multiple possible states) tracking capabilities. However, MPFs are susceptible to identity switches as the particles tend to merge and split in due process. Cai *et al.* [11] improve upon Okuma *et al.* [68] by embedding the mean-shift algorithm into particle filters to stabilize player trajectories during occlusion and utilize rink coordinates (homography) to compensate camera motion from broadcast. But, the algorithm is not trained or tested on broadcast hockey videos, but performs tracking directly in the rink coordinate system derived from manual homography estimation.

Adopting recent advancements in deep learning, Vats *et al.* [101] present the first *private* benchmark for ice hockey by fine-tuning five different tracking algorithms for hockey MOT. They also extend this work and propose a graph message passing networks (MPN)-based hockey tracking model [101, 102] (henceforth, we shall call this model HockeyMOT for simplicity) by fine-tuning the MOT neural solver [9] architecture on hockey tracking dataset - 1 (HTD-1) (dataset described in Section 3.2). HockeyMOT obtains 94.5% multi-object tracking accuracy (MOTA), incurring 431 ID switches and 62.9% IDF1 score on the test-set from HTD-1, when evaluated using Faster RCNN-based [76] player detections. However, this approach has two key limitations: first, their usage of bounding box attributes for encoding graph edges leads to identity switches and uncertain tracklets, especially during

heavy player occlusions in broadcast feeds where overlapping bounding boxes (intersection over union (IoU) $\uparrow$) cause player misassociations (IDsw $\uparrow$) and lost tracks. Second, they rely solely on appearance features to encode graph nodes which is problematic in hockey, since player identification is complicated by uniform physical builds, protective equipment, identical team uniforms, motion blur, and variable rink illumination.

2.7 Discussion

Consolidating the above review, three main conclusions can be drawn:

1. Unlike soccer and basketball where there exists standard datasets [19,20], benchmarks [62], and multiple tracking approaches, ice hockey remains relatively less explored.
2. The limitations in the current tracking baseline for ice hockey [101] provides further scope for reducing identity switch errors (IDsw) and improving the identify F1 score (IDF1) score.
3. The TBD paradigm is a popular choice among current MOT methods especially in team sports, since each component can be optimized individually for enhanced predictions.

Based on these observations, our first objective is to establish a *reproducible* reference benchmark using the *open-source* HTD-2 to enable broader comparisons among hockey MOT methods. We select six different tracking algorithms including early handcrafted methods [7,12], transformer-based [64] and graph-based [101] architectures, and present a detailed study of their merits and drawbacks on ice hockey in Chapter 4.

Following the TBD paradigm due to its proven suitability for team sports, in Chapter 5, we experiment with adapting the YOLOX detector [122] and OSNet reID architecture [124] for hockey. We also discuss the benefits of fine-tuning them on our hockey datasets.

To the best of our knowledge, HockeyMOT [101] is the current state-of-the-art (SOTA) approach for MOT in ice hockey. This demonstrates that graph-based formulations offer a reliable way to model dynamic player movements and their relationships on the rink. But, the current baseline incurs large IDsw errors which leads to inconsistent tracking. Motivated by works using homography to transform multi-camera feeds in soccer [40], we apply homography [83] on single-camera broadcast feeds to extract view-invariant player positional coordinates and build a novel graph-based MPN (Chapter 6) to minimize player identity switches and predict identity-preserved tracklets.

Chapter 3

Datasets and Evaluation Metrics

3.1 Introduction

Access to high-quality data is a fundamental requirement for any machine learning (ML) algorithm. We derive the prediction $\hat{y} = f(x; \theta)$ by approximating function f based on the available data x (conditioned on parameter θ). The quality and reliability of datasets, therefore, steers the caliber and accuracy of models. In this chapter, we present an overview of the existing hockey tracking dataset - 1 (HTD-1) and introduce hockey tracking dataset - 2 (HTD-2) – an *open-source* extension to the existing dataset for multiplayer tracking in ice hockey. We also discuss different standard evaluation metrics [6] followed by current MOT literature, and analyze their relevance for our application.

3.2 Hockey Tracking Dataset - 1 (HTD-1)

The hockey tracking dataset - 1 (HTD-1) was annotated from scratch using the computer vision annotation tool (CVAT)¹ tool by ice hockey domain experts from Stathletes Inc², specifically for MOT. It contains 26 clips sampled from five different NHL games, with an average duration of approximately 35 seconds per clip. The clips have a standard definition resolution of 1280×720 p, sampled uniformly at 30 Hz from broadcast videos. The train-test-validation split is 12:13:1, carefully partitioned to avoid any game-level bias.

¹Found at: <https://github.com/cvat-ai/cvat>

²<https://www.stathletes.com/>

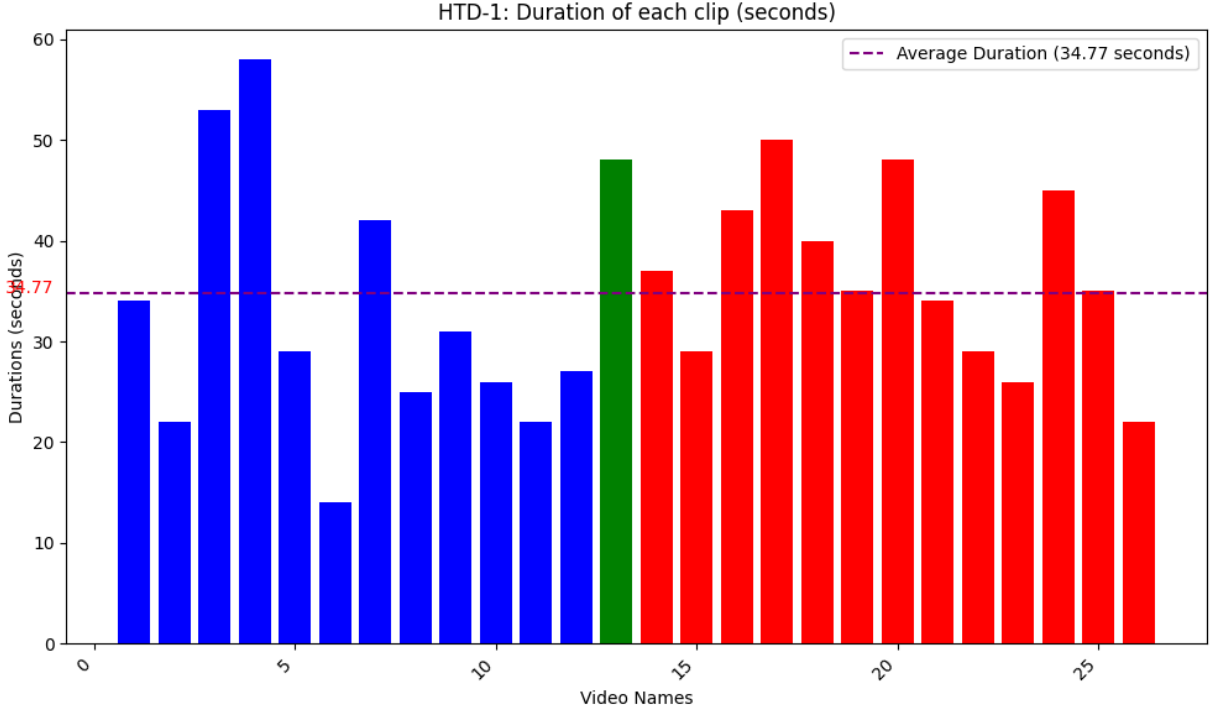


Figure 3.1: Duration of each clip in HTD-1. The average duration of all clips, $\mu(\text{clips}) = 35$ seconds (approximately).

This dataset is annotated in the MOT-challenge [47] format containing frame ID, tracking ID, bounding box coordinates, confidence score, category (foreground/background), and occlusion score. All annotations in this dataset were obtained with permission from Stathletes Inc.

3.3 Hockey Tracking Dataset - 2 (HTD-2)

We introduce hockey tracking dataset - 2 (HTD-2) (also called VIP - hockey tracking dataset (VIP-HTD)) – an open-source, *curated* MOT dataset for ice hockey as a complement to HTD-1, to enable public comparisons and benchmarking. To the best of our knowledge, this is the first *open-source* MOT dataset for ice hockey available with a reference benchmark and remains publicly accessible for research³.

³<https://github.com/harshap-ai/VIP-HTD/>



Figure 3.2: Inaccurate annotations in [MHPTD](#). (L-R) erroneous annotation, missing player detections & offset bounding boxes, respectively.

This dataset is curated from the McGill hockey player tracking dataset ([MHPTD](#)) [114], which remained unusable for [MOT](#) evaluations due to several limitations: (i) erroneous or missing annotations, (ii) redundant clips, (iii) incorrect metadata, and (iv) lack of properly sampled frames (Fig. 3.2). While creating a new dataset from scratch through manual annotations would be an alternative, it is prohibitively time-intensive, requiring approximately 10.45 minutes to annotate a one-minute clip sampled at 30 Hz, and 20 minutes for a 60 Hz clip [101]. Therefore, given the existing [MHPTD](#), we opted for *curation* over creation, focusing on addressing the aforementioned errors through extensive preprocessing, data wrangling, and manual verifications. We refined the [MHPTD](#) dataset by removing redundant videos, correcting erroneous annotations, extracting parsed frames for each clip, and conducting thorough metadata analysis (Table 3.1) to establish hockey tracking dataset - 2 ([HTD-2](#)). Our work complies with all required permissions and licenses⁴.

3.3.1 Metadata

[HTD-2](#) consists of 22 broadcast clips from eight different [NHL](#) games, containing side-of-the-rink views with a standard definition resolution of 1280×720 p. The average duration of each clip, $\mu = 73.5$ seconds (increases twice compared to [HTD-1](#)) with frames sampled at either 30 Hz or 60 Hz. The refined annotations follow the standard [MOT](#) annotation format, with confidence score set to one, as they are manually verified annotations. The train-test-validation split we follow is 14:7:1, chosen with the test set containing mutually exclusive clips from the train set to avoid data leakage and game-level bias.

⁴MHPTD dataset was open-sourced using [GPL-3.0](#). Following all terms and conditions, we inherited the same licensing to open-source our work.

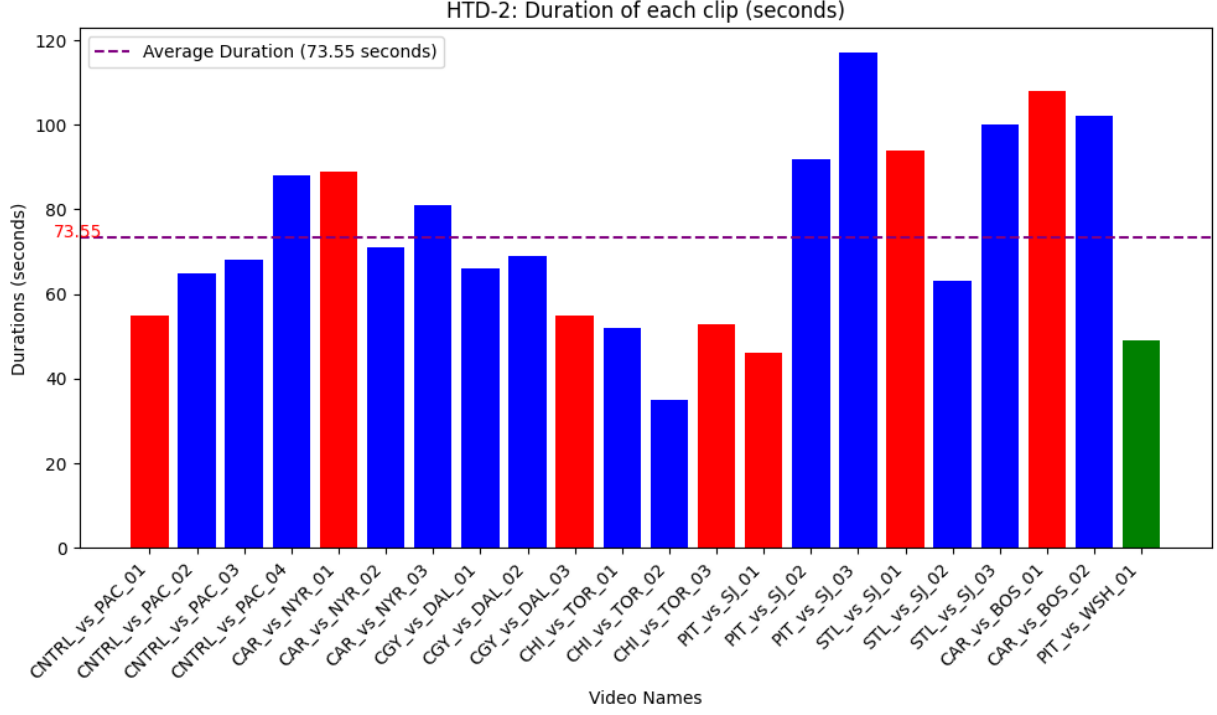


Figure 3.3: Duration of each clip in HTD-2. Blue plots indicate train clips, red plots indicate test clips, and green plots indicate validation clip. The average duration of all clips, $\mu(\text{clips}) = 73.5$ seconds.

Table 3.1 below shows the duration of each clip, number of frames in each sequence, number of player instances annotated (bounding boxes), and their sampling rates. In total, HTD-2 contains 1,560 seconds of annotated clips with 73,889 frames and 576,016 player instances. Figure 3.4 below represents two relationships – games vs. number of player instances, and games vs. number of frames, respectively. The key takeaway from this pie chart is the balanced distribution of games *without* any single game dominating more than 25% of this dataset.

Games	Clips	Duration (sec)	No. of frames	No. of bounding boxes	FPS (Hz)
Central vs. Pacific	Clip #1	55	1,671	11,052	30
	Clip #2	65	1,965	12,572	30
	Clip #3	68	2,066	2,066	30
	Clip #4	88	2,646	15,769	30
Blackhawks vs. Maple Leafs	Clip #1	52	1,585	14,154	30
	Clip #2	35	1,067	7,698	30
	Clip #3	53	1,618	13,912	30
Sharks vs. Penguins	Clip #1	46	1,391	10,685	30
	Clip #2	92	2,762	21,420	30
	Clip #3	117	3,525	27,797	30
Hurricanes vs. Bruins	Clip #1	108	6,517	51,436	60
	Clip #2	44	2,644	22,814	60
Blues vs. Sharks	Clip #1	94	5,647	40,740	60
	Clip #2	63	3,809	29,194	60
	Clip #3	100	6,027	46,323	60
Hurricanes vs. Rangers	Clip #1	89	5,342	44,854	60
	Clip #2	71	4,283	37,919	60
	Clip #3	81	4,879	42,250	60
Stars vs. Flames	Clip #1	66	3,969	32,904	60
	Clip #2	69	4,160	36,367	60
	Clip #3	81	4,876	27,615	60
Penguins vs. Capitals	Clip #1	49	2,981	25,471	60
Total		1,560	73,889	576,016	

Table 3.1: Metadata for each curated clip from [HTD-2](#). Yellow highlighted fields indicate test split, cyan highlighted fields indicate validation split, and the rest indicate training split.

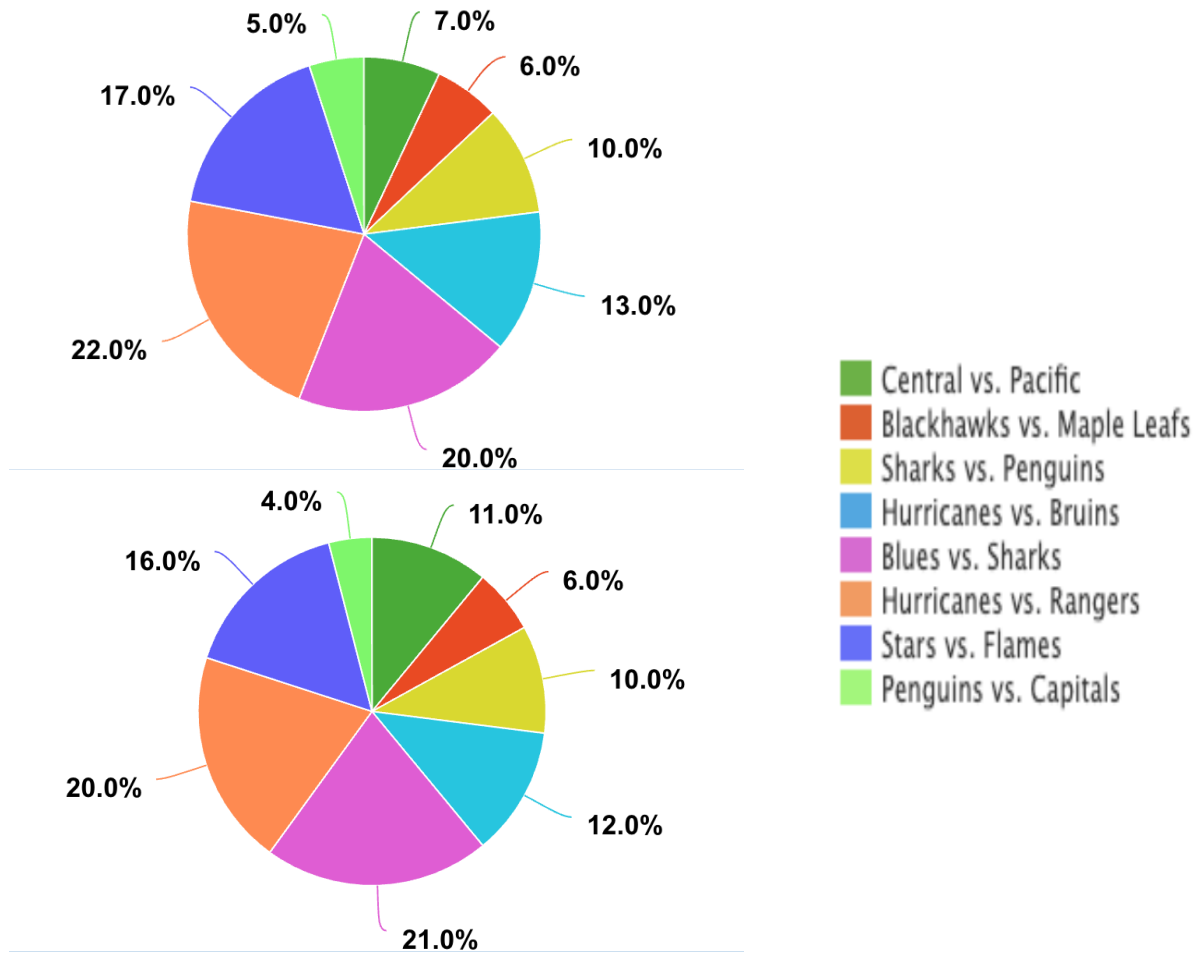


Figure 3.4: **(left-right)** Percentage of player instances by games and percentage of frames by games, respectively, from [HTD-2](#). All values are in percentage (%).

3.3.2 Frame extraction methodology

MHPTD did not contain any extracted frames from clips, which is necessary for **MOT** evaluations and visualizations. Therefore, we parsed all clips into frames using the open-source **CVAT** tool at their native sampling rates. Interestingly, using external parser libraries such as OpenCV⁵ or moviepy⁶ generated excessive frames compared to **CVAT** for any given clip. Our investigation revealed that this deviation stemmed from differences in video synchronization approaches across tools. Specifically, **CVAT** uses FFMPEG⁷ as a backend API for parsing videos, which passes timestamps from the demultiplexer to the multiplexer (enable option: `-vsync 0`), ensuring synchronized frame parsing at the clip’s native rate. Without this synchronization, we observed an offset effect that gradually shifts player bounding box locations towards the latter frames of the sequence. Note that this effect is separate from the original offset errors found in the annotations of **MHPTD**. To prevent such issues during future use, we have included the properly synchronized frames in the dataset repository.

3.3.3 Identity assignment and annotation strategy

Conventionally, in pedestrian-based **MOT** datasets [22, 47, 65], objects (pedestrians) are assigned identities at the *tracklet* level i.e., when an object exits at time t_i and re-enters at time t_j s.t. $j > i$, it receives a *new* ID. This approach is justified as pedestrians typically move unidirectionally in natural environments and rarely re-enter the camera **FoV**. Conversely in sports, object (player) identities are assigned at the *personnel* level i.e., when a player exits at time t_i and re-enters at time t_j s.t. $j > i$, they maintain their previously assigned unique ID. This distinction arises because team sports have a fixed number of players who frequently reappear within the camera **FoV** during the game. Both our hockey datasets follow this *personnel* level annotation scheme, as they are annotated by ice hockey domain experts who can reliably distinguish players using their expertise and multiple available visual cues.

From a computer vision standpoint, the *personnel* annotation scheme addresses the broader goal of unique player identification, assigning each player a single identity throughout the entire game, regardless of temporary occlusions or exits from the camera **FoV**. However, our monocular tracking approach relaxes this constraint, focusing instead on main-

⁵<https://opencv.org/>

⁶<https://pypi.org/project/moviepy/>

⁷<https://www.ffmpeg.org/>

taining unique identities only while players are *detectable* and *visible* within the **FoV**. We perform *frame-to-frame* associations to preserve player identities within predicted tracklets.

To elaborate formally, during our evaluations each predicted hypothesis ID $\mathcal{H}_m$ is matched to the ground truth player ID $\mathcal{O}_n$ based on a pre-defined distance threshold (**IoU** ≥ 0.5 by default), yielding the output pair $\{\mathcal{H}_m, \mathcal{O}_n\}$, where m and n are tracked and ground truth player IDs, respectively. When a player re-enters after an interval Δt , either following occlusions or after exiting and re-entering the camera **FoV**, they receive a new hypothesis ID $\mathcal{H}_k$, creating a new output pair $\{\mathcal{H}_k, \mathcal{O}_n \mid k \neq m\}$ that differs from their previous match. This change increases the **IDsw** score according to standard **MOT** metrics [6]. Subsequently, $\mathcal{H}_k$ becomes the new default hypothesis for ground truth $\mathcal{O}_n$ until another switch occurs, preventing the propagation of the same switch throughout the tracklet.

To address this situation, we translate the *personnel* level annotation scheme to *tracklet* level annotations, aligning with our definitions of consistent tracking and identity-preserved tracklets. For **HTD-2**, both annotation formats are available in the open-source github repository.

3.4 Evaluation Metrics

Clear **MOT** [6] establishes the standard evaluation metrics widely adopted by current **SOTA** tracking methods. It introduces two primary metrics: the multi-object tracking accuracy (**MOTA**) score [6], combining false positives (**FPS**), false negatives (**FNS**), and identity switch score (**IDsw**), and the identify F1 score (**IDF1**) [78]. In the context of ice hockey, these can be defined as follows:

- False positives (**FP**): Number of players detected incorrectly per clip.
- False negatives (**FN**): Number of true players missed per clip.
- Identity switch score (**IDsw**): Number of identity swaps made for players when their trajectories intersect (during occlusions) at the frame-level.
- Identity F1 score (**IDF1**): A measure of tracking consistency that evaluates how well player identities are preserved in predicted tracklets compared to ground truth trajectories at the tracklet-level, expressed as a percentage (%).

The above metrics can be mathematically described as follows:

- **False positives (FPs)** measures the frequency of incorrect detections where a predicted hypothesis $\mathcal{H}_i$ has no matching ground truth GT_i within a specified **IoU** threshold. This is computed as:

$$\text{FP} = \frac{\sum_{i=1}^n \text{FP}_i}{\sum_{i=1}^n \text{GT}_i} \quad (3.1)$$

where n denotes the total number of frames.

- **False negatives (FNs)** measures the frequency of missed detections where a ground truth GT_i has no corresponding hypothesis $\mathcal{H}_i$ within a specified **IoU** threshold. This is computed as:

$$\text{FN} = \frac{\sum_{i=1}^n \text{FN}_i}{\sum_{i=1}^n \text{GT}_i} \quad (3.2)$$

where n denotes the total number of frames.

- **ID switch (IDsw)** errors occur when hypothesis-object assignments swap between frames, where initial pairs $\{\mathcal{H}_1, \mathcal{O}_1\}$ and $\{\mathcal{H}_2, \mathcal{O}_2\}$ become $\{\mathcal{H}_1, \mathcal{O}_2\}$ and $\{\mathcal{H}_2, \mathcal{O}_1\}$. This is computed as:

$$\text{IDsw} = \frac{\sum_{i=1}^n \text{IDsw}_i}{\sum_{i=1}^n \text{GT}_i} \quad (3.3)$$

where n denotes the total number of frames.

- **Multi-object tracking accuracy:** consolidates false negatives, false positives, and identity switches into a single metric that measures overall tracking performance. It is computed as:

$$\text{MOTA} = 1 - \frac{\sum_{i=1}^n \text{FN}_i + \text{FP}_i + \text{IDsw}_i}{\sum_{i=1}^n \text{GT}_i} \quad (3.4)$$

The range of **MOTA** $\in (-\infty, 100]$ and is measured as a percentage (%).

- **IDF1 score** measures tracking consistency by quantifying the ratio of correctly identified player tracklets to the average number of ground truth and predicted identities. Mathematically, it is the harmonic mean of identity precision and identity recall:

$$\text{IDF1} = 2 \times \frac{\text{TP}^{\text{id}}}{(2 \times \text{TP}^{\text{id}}) + \text{FP}^{\text{id}} + \text{FN}^{\text{id}}} \quad (3.5)$$

where, $\text{TP}^{\text{id}}, \text{FP}^{\text{id}}, \text{FN}^{\text{id}}$ represent correctly identified, falsely identified, and missed player identities in the predicted set of tracklets $\mathcal{T}$, respectively.

Discussion

The **FPs** and **FNs** used to calculate **MOTA** are predominantly influenced by detector performance [6]. In the **MOTA** formulation (Eqn. 3.4), detection errors (**FP** and **FN**) carry twice the weight of tracking errors (**IDsw**), making it a poor indicator of tracking consistency even when player associations are accurate.

For evaluating tracking performance in ice hockey, we therefore focus on two more relevant metrics: **IDsw** and **IDF1** scores. The **IDsw** directly measures tracking consistency by counting identity inconsistencies at the frame level, while **IDF1** provides a balanced assessment of both detection and tracking quality at the tracklet level. Consequently, we optimize our approach to minimize **IDsw** errors ($\downarrow$) and maximize **IDF1** score ($\uparrow$) to achieve robust identity-preserved player tracking

3.5 Conclusion

This chapter presented two datasets used in this work: the hockey tracking dataset - 1 (**HTD-1**) and the open-source VIP - Hockey Tracking Dataset (**HTD-2**), curated for reference benchmarking and open-source access. We provided a comprehensive metadata analysis of **HTD-2**, discussing curation challenges, frame extraction methodology, and identity assignment strategies. Finally, we described standard **MOT** evaluation metrics from current literature and identified those most relevant for ice hockey tracking. The following chapter establishes a public benchmark using **HTD-2** and analyzes the performance of six popular tracking methods when fine-tuned for ice hockey.

Chapter 4

Public Tracking Benchmark

4.1 Introduction

Benchmarks in computer vision are integral for identifying drawbacks with existing methods and advancing the current state-of-the-art (SOTA). Following the introduction of hockey datasets in chapter 3, in this chapter we establish a *public* reference benchmark for ice hockey. We evaluate six tracking methods fine-tuned on HTD-2, selected to maximize methodological diversity for a comprehensive comparison.

4.2 Tracking Benchmark: Motivation

Following the MOT benchmark established by Vats *et al.* [101] on HTD-1, we present a new *public* benchmark utilizing contemporary tracking approaches: SORT [7], OC-SORT [12], Tracktor [5], ByteTrack [119], HockeyMOT [101], and Trackformer [64]. To ensure *reproducibility*, all methods are fine-tuned on HTD-2 while maintaining their original optimal hyperparameter configurations. The selected methods represent diverse tracking paradigms: SORT and OC-SORT exemplify the TBD approach, employing classical techniques such as Kalman filters [41] and Hungarian matching [45] for prediction and association. Tracktor and ByteTrack implement the JDT paradigm, performing autoregressive frame-to-frame player associations. HockeyMOT, which serves as our baseline, adopts the TBD framework utilizing a graph-based MPN, while Trackformer introduces an innovative *tracking-by-attention* paradigm leveraging the popular transformer architecture [2].

4.2.1 Simple online and realtime tracking (SORT)

SORT [7] established a significant tracking baseline by combining a CNN-based detection step with Kalman filters [41] for tracking object states (per frame) and Hungarian algorithm [45] for object association (across frames). The algorithm introduced two key innovations: (i) The integration of CNNs for object localization, which achieved substantial improvements in tracking accuracy by 18.9% ($\uparrow$), and (ii) Real time (online) tracking capabilities with a 260 Hz inference speed for tracker state updates. Originally, SORT implemented the Faster R-CNN [76] detector pre-trained on the Pascal VOC [25] dataset without any fine-tuning. In our benchmark experiments, we fine-tune Faster R-CNN on ice hockey for better domain adaptability.

4.2.2 Observation-centric SORT (OC-SORT)

OC-SORT [12] extends the classical SORT framework [7] by addressing the limitations inherent in linear motion model assumptions through the integration of object observations from the YOLOX detector [122] for computing virtual trajectories during occlusion periods. The method specifically targets the mitigation of temporal error accumulation in Kalman filter [41] implementations, where state estimation errors typically propagate through posteriori updates. This enhancement, termed the *re-update* stage by the authors, is activated upon the re-association of lost tracks with observations, utilizing virtual trajectories to bridge tracking discontinuities. The primary methodological contribution lies in its shift from traditional estimation-centric tracking, which relies primarily on Kalman filter predictions, to an *observation-centric* approach that prioritizes actual detector measurements for trajectory updates. In our experiments, we adopt the YOLOX detector from the original work and fine-tune it on hockey.

4.2.3 Tracktor

Tracking without bells and whistles, known as Tracktor [5], reformulates tracking as a bounding box regression problem based exclusively on Faster R-CNN [76] detections. This approach effectively transforms a detector into a tracker by substituting the traditional association step with a regression step, eliminating the need for explicit tracking dataset optimization. The model performs temporal bounding box regression from position b_{t-1}^k at frame $t - 1$ to position b_t^k at frame t , employing RoI pooling on current frame features using the predicted coordinates from the previous frame. This can be formally expressed

as:

$$b_t^k = \text{RoIpool}(b_{t-1}^k, \mathcal{F}^t)$$

where $\text{RoIpool}(\cdot)$ performs regression by RoI pooling on the features of the current frame using previous frame’s bounding box coordinates. The authors enhance this base model with Tracktor++, which incorporates two additional components: a *motion model* to address challenges posed by camera movement and low frame rates, and a Siamese CNN-based reID [37] framework for appearance-based short-term associations. For our benchmark evaluation, we implement Tracktor++ with its detector fine-tuned on HTD-2, owing to its superior performance characteristics.

4.2.4 ByteTrack

ByteTrack [119] introduces an innovative association strategy that diverges from conventional trackers by considering both high and low-confidence detection scores, rather than relying solely on detections above a predetermined confidence threshold. This approach is motivated by the observation that low-confidence detections may contain non-negligible true positiveness (TPs), whose exclusion can result in object misses and track fragmentation. The methodology comprises two distinct association phases: (i) High-confidence association, which adopts the SORT [7] framework utilizing Kalman filters for track prediction and the Hungarian algorithm [45] for association based on either IoU scores or reID feature similarities, and (ii) Low-confidence association, which matches previously unmatched tracks with low-score detections exclusively through IoU-based similarity. The original implementation employs the YOLOX detector [122], pretrained on common objects in context (COCO) [53] and subsequently fine-tuned on a composite pedestrian dataset comprising MOT17 [65], CrowdHuman [84], Cityperson [118], and ETHZ [24]. For consistent evaluation, a detection score threshold $\tau = 0.6$ is maintained. In our experimental setup, we preserve all default procedures while adapting YOLOX through fine-tuning on ice hockey data.

4.2.5 HockeyMOT

HockeyMOT [101], drawing inspiration from the MOT Neural Solver [9] architecture, implements tracking as a fully differentiable framework utilizing MPN [4] in a graph-based structure. The framework encodes player features as node and edge embeddings within the graph, associating similar players through minimum cost flow formulation [116] principles. The model architecture comprises four distinct stages (illustrated in Fig. 4.1):

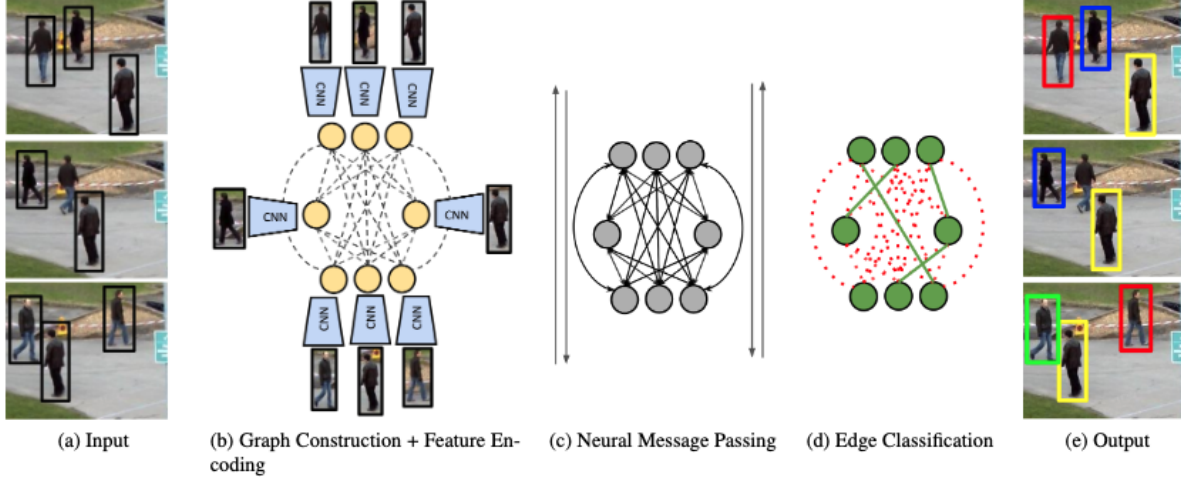


Figure 4.1: HockeyMOT [101] architecture inspired by MOT Neural Solver [9]

- **Graph Construction:** Given a set of object detections, the model constructs one graph per frame where nodes encode player features and edges represent their interconnections.
- **Feature Encoding:** The appearance embeddings for nodes are computed via a CNN, while the geometric embeddings for the edges are derived through an MLP using the relative size of bounding box coordinates, position and temporal (frame number) inputs.
- **Neural Message Passing:** The core component comprises m message passing iterations (hyperparameter), facilitating bidirectional information exchange: nodes share appearance features with connecting edges, while edges communicate geometric information to incident nodes.
- **Training:** The network is trained using a cross-entropy loss function [121] to perform binary classification on the edge embeddings after the final message passing step. During inference, a threshold $\tau = 0.5$ binarizes output probabilities, where 0 and 1 indicate absence and presence of player associations, respectively.

Following the TBD paradigm, HockeyMOT employs a Faster R-CNN detector [76] fine-tuned on HTD-1. It utilizes a Resnet50 [35] architecture as the reID network backbone, pre-trained on standard re-identification datasets: Market1501 [123], CUHK03 [51], and

DukeMTMC [79] adhering to established MOT practices. The implementation adopts $m = 12$ message passing iterations based on empirical evaluation, with tracking MLPs trained on HTD-1. In our experimental framework, while maintaining the pretrained reID network, we fine-tune both detector and tracker components on the HTD-2 dataset for fair comparisons.

4.2.6 Trackformer

Trackformer [64] presents an end-to-end trainable transformer architecture that conceptualizes tracking as a frame-to-frame set prediction task. The framework introduces a novel *tracking-by-attention* paradigm, leveraging both self- and encoder-decoder attention mechanisms for track evolution. The architecture integrates a Resnet50 [35] CNN for bounding box feature extraction, a transformer encoder for feature encoding, and a transformer decoder that generates target output embeddings containing bounding box and class information. The framework’s operational pipeline comprises four sequential stages:

- *Step 1:* Frame-level feature extraction utilizing a Resnet50 [35] CNN.
- *Step 2:* Feature encoding through self-attention mechanisms via a transformer encoder, initially pre-trained on the COCO [53] dataset and subsequently fine-tuned on MOT17 [65].
- *Step 3:* Query decoding in the transformer decoder through dual attention mechanisms: self-attention and encoder-decoder attention.
- *Step 4:* Translation of decoded queries to bounding box coordinates and class predictions via MLPs.

The encoder-decoder architecture (Figure 4.2), inspired by the DETR framework [127], incorporates multi-head self attention (MHSA) layers, layer normalization [48], and fully-connected feed forward networks (FFNs). The decoder head implements dual FFNs for class categorization (person versus background) and bounding box coordinate prediction. Let $\pi : i \rightarrow j$ be a mapping function, where $j = \pi(i)$ assigns predictions $\hat{y}_j \in \hat{\mathbf{Y}}$ to ground truth objects $y_i \in \mathbf{Y}$, established either through track identity or through similarity metrics based on object class and bounding box correspondence scores.

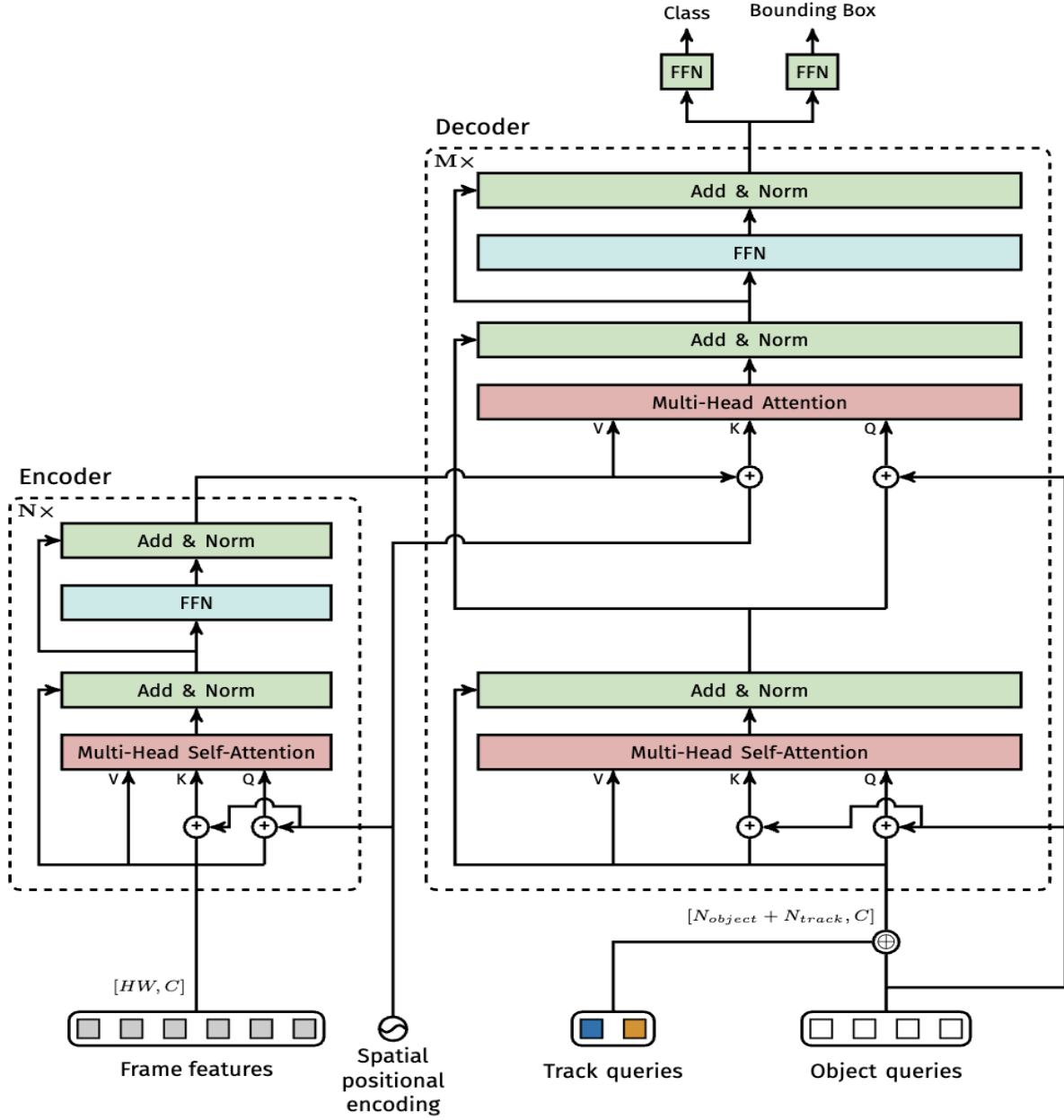


Figure 4.2: The encoder-decoder architecture blocks in Trackformer [64].

4.3 Comparison Study

4.3.1 Implementation details

All benchmark methods, except Trackformer [64], employ either Faster R-CNN [76] or YOLOX [122] for object detection. These detectors are fine-tuned on HTD-2 to address the distribution shift between hockey player detection and conventional pedestrian or crowd scenarios (Section 5.2). The models were trained on 14 training clips and evaluated on seven test clips using an NVIDIA Geforce RTX 2080Ti GPU with 12 GB RAM and 1.3GHz clock speed.

For all the benchmarked methods, we utilized the default hyperparameters since they are optimized for best performance. For SORT [7], we followed the default hyperparameters by setting the IoU threshold at 0.3. For OC-SORT [12], we used a batch size of 64 and set the IoU threshold at 0.3. For Tracktor [5], the batch size was set at 2, learning rate at $1e^{-5}$, and fine-tuned for 27 epochs. For ByteTrack [119], we set the batch size = 48, learning rate = $1e^{-3}$, and fine-tuned the model for 80 epochs. In HockeyMOT [101], we set the batch size = 8, learning rate = $1e^{-2}$, and fine-tune it for 25 epochs, as followed in the baseline. For Trackformer, we set batch size = 2, learning rate = $2e^{-3}$, and fine-tune the attention network for 50 epochs.

4.3.2 Quantitative results

Quantitative results using standard MOT metrics [6] are presented in Table 4.1. For detailed performance analysis, Tables 4.2 and 4.3 provide clip-wise IDsw and IDF1 scores across the seven evaluation clips—two critical metrics that effectively capture tracking performance as discussed in Chapter 3 (subsection 3.4).

From Table 4.1, ByteTrack [119] demonstrates superior performance with a **27.4%** reduction in IDsw and a **3.8%** improvement in IDF1 scores compared to the next best performing methods. This enhanced performance can be attributed to its novel association strategy that leverages both high and low-confidence detections to handle occlusions and visual uncertainties. Conversely, HockeyMOT [101] achieves **73%** fewer FPs and **90%** fewer FNs compared to ByteTrack, owing to its hybrid approach that combines its native tracking with Tracktor-based trajectory imputation during post-processing. However, this introduces additional computational overhead of $O(\text{HockeyMOT} + \text{Tracktor})$ and increased model complexity in terms of floating-point operations per seconds (FLOPs).

Method	MOTA % ($\uparrow$)	FP ($\downarrow$)	FN ($\downarrow$)	IDsw ($\downarrow$)	IDF1 % ($\uparrow$)
SORT [7]	70.24	28,475	33,135	685	61.61
OC-SORT [12]	72.5	24,735	31,076	<u>541</u>	73.81
Tracktor [5]	71.28	30,692	30,169	1,100	59.6
ByteTrack [119]	82.97	<u>13,804</u>	<u>14,637</u>	393	83.5
HockeyMOT [101]	<u>76.02</u>	3,721	1,417	787	<u>80.2</u>
Trackformer [†] [64]	<0	367,253	32,356	858	32.6

Table 4.1: Public reference benchmark for ice hockey. All benchmarked trackers were evaluated on the seven test clips from HTD-2. Best results are in **red**, second best results are underlined. [†] indicates no fine-tuning on ice hockey. Note that **MOTA** $\in (-\infty, 100]$.

Test clips	CAR vs BOS	CAR vs NYR	CGY vs DAL	CHI vs TOR	CNTRL vs PAC	PIT vs SJ	STL vs SJ
SORT [7]	109	128	64	91	114	59	120
OC-SORT [12]	130	80	51	47	110	26	97
Tracktor [5]	141	225	73	152	191	140	178
HockeyMOT [101]	115	71	41	152	136	164	108
ByteTrack [119]	67	46	39	33	81	31	96
Trackformer [†] [64]	166	128	86	147	97	92	142

Table 4.2: Video-wise **IDsw** benchmark on HTD-2. Lower ($\downarrow$) is better. ([†]) indicates no fine-tuning on ice hockey. Best results are highlighted in **red**.

Test clips	CAR vs BOS	CAR vs NYR	CGY vs DAL	CHI vs TOR	CNTRL vs PAC	PIT vs SJ	STL vs SJ
SORT [7]	57.0%	59.5%	56.5%	50.2%	64.6%	63.3%	80.2%
OC-SORT [12]	62.2%	86.3%	72.7%	65.7%	72.2%	73.2%	84.4%
Tracktor [5]	55.8%	57.6%	56.8%	47.0%	61.4%	61.8%	77.2%
HockeyMOT [101]	85.4%	81.1%	77.4%	60.3%	71.7%	70.1%	86.2%
ByteTrack [119]	84.5%	86.6%	88.9%	78.9%	80.3%	80.0%	85.9%
Trackformer [†] [64]	38.5%	27.1%	33.2%	34.7%	37.5%	42.6%	31.0%

Table 4.3: Video-wise [IDF1](#) benchmark on [HTD-2](#). Higher ($\uparrow$) is better. ($\dagger$) indicates no fine-tuning on ice hockey. Best results are highlighted in [red](#).

Tracktor [5] underperforms compared to other [JDT](#) methods due to its assumption of minimal frame-to-frame object displacement—a constraint that requires high frame sampling rates. While it implements a motion model to address this limitation, the underlying *constant velocity* assumption proves inadequate for capturing the rapid, non-linear player movements characteristic of ice hockey. Similarly, SORT [7] and OC-SORT [12] exhibit performance degradation due to their reliance on Kalman filters [41] with *linear velocity* assumptions. Trackformer [64] presents an unexpected phenomenon: its fine-tuned version underperforms compared to its pre-trained model. This counterintuitive result stems from transformers’ lack of inductive bias [46], requiring extensive training data for effective domain adaptation. The superior performance of the pre-trained model, trained on pedestrian MOT [65] and CrowdHuman [84] datasets, suggests that transformers derive greater benefit from general tracking features learned from large-scale, diverse datasets than from domain-specific fine-tuning on limited hockey data. Consequently, we report only the pre-trained Trackformer results (denoted by $\dagger$) in our benchmark tables, as the fine-tuned variant’s diminished performance precludes meaningful comparative analysis.

4.4 Conclusion

We established a comprehensive, public benchmark for MOT in ice hockey, evaluating six diverse tracking methodologies using standard MOT metrics [6]. This provides a reference standard for evaluating future tracking methods in hockey player tracking. Analysis of the various approaches yields two critical insights: (i) JDT methods demonstrate limited potential for further advancement due to the inherent constraints of joint detection-tracking optimization, and (ii) graph-based approaches, while demonstrating competitive performance, offer substantial opportunities for fundamental improvements. Furthermore, graph structures inherently provide a more natural representation for modeling human motion and dynamics. Motivated by these findings, our subsequent chapters focuses on addressing the current limitations of HockeyMOT—specifically its susceptibility to ID switches and tracklet identity inconsistencies—through the development of a novel graphical message passing networks (MPN) for SOTA player tracking.

Chapter 5

Part I: Multi-object Tracking Pipeline

5.1 Introduction

Having established the contextual ground for this thesis, we now present the central framework of this thesis: an end-to-end multi-player tracking pipeline for ice hockey. Our methodology adopts the established TBD paradigm, which traditionally comprises of three core components: (i) detection, (ii) reID, and (iii) tracking. We extend this framework for hockey, by introducing a novel fourth component—homography—that warps camera view-variant broadcast frames onto a common view-invariant overhead rink template, providing a uniform top-down perspective to mitigate positional ambiguities from perspective distortions and occlusions (Section 5.4). This helps our objective by enhancing the robustness of tracking especially during player occlusions. Each component in our proposed pipeline is independently optimized and end-to-end trainable, with no shared parameters across modules. The selection of algorithms was guided by their capacity to effectively model the characteristics specific to hockey’s independent and identical distribution (i.i.d).

The remainder of this chapter details these foundational components: detection, reID, and homography. We discuss the implementation and performance of each module, establishing the technical groundwork for our tracking framework. Chapter 6 then demonstrates how these components integrate within our graph-based tracking module, followed by experimental analysis and ablation studies.

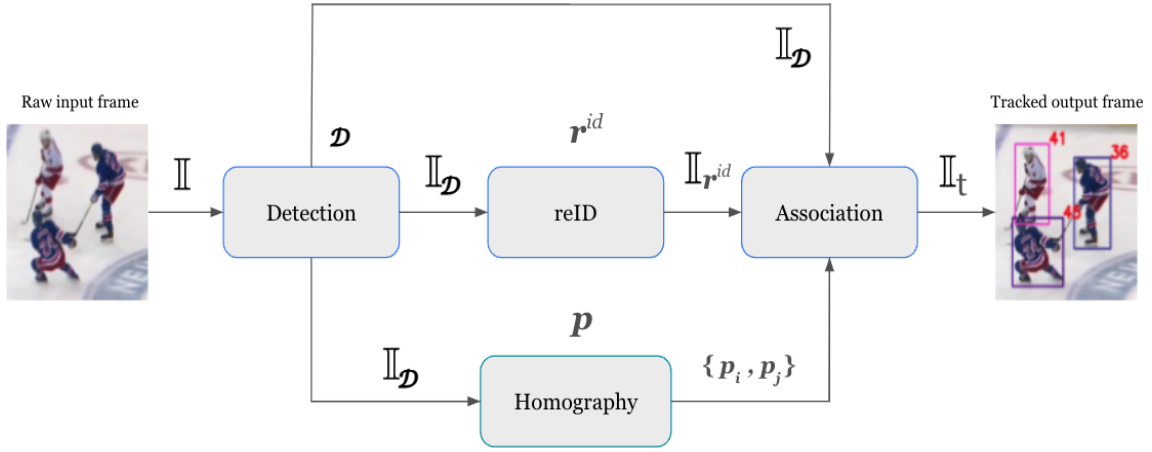


Figure 5.1: *Overall tracking pipeline.* The raw input frame $\mathbb{I}$ is first processed by the detector $\mathcal{D}$, generating bounding box coordinates $\mathbb{I}_{\mathcal{D}}$. These coordinates are input to both the **reID** network r^{id} , which downsamples the image to 256×128 and extracts the 512-dimensional features, and the homography model p , which maps footpoint coordinates to an overhead template $\{p_i, p_j\}$. The outputs from the detection, **reID**, and homography models are then used in the association step to produce the tracked frame $\mathbb{I}_t$. **Blue**-bordered blocks indicate trained or fine-tuned models; **green**-bordered block denotes off-the-shelf homography model. Note that the player track IDs are sequentially initialized and do not correspond to their jersey numbers.

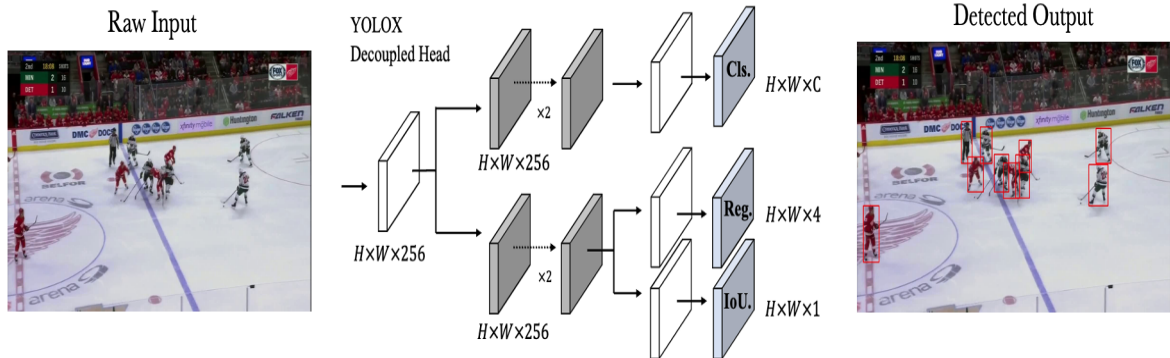


Figure 5.2: YOLOX [122] architecture adapted for ice hockey player detection. Unlike previous YOLO versions (v3-v5), YOLOX employs separate heads for class prediction, IoU prediction, and regression, making it better suited for sports applications.

5.2 Detection

In TBD approaches, detection serves as the primary step in object localization across temporal sequences. Reliable detection reduces tracklet fragmentation. While contemporary tracking methods predominantly utilize either a version of the YOLO family of detectors [8, 27, 73, 74] or RCNN-based architectures [31, 32, 76] for pedestrian tracking scenarios, YOLOX [122], an enhanced version of the YOLO family, has emerged as the preferred choice for sports tracking applications [?, 20, 26, 33, 101]. The advantages of YOLOX are methodological: its anchor-free detection architecture reduces FLOPs and computational latency, while its dual-branch structure separates classification and regression heads, striking a balance between computational efficiency and detection accuracy (Fig. 5.2).

Therefore, for ice hockey implementation, we adopted the largest backbone with 99.1M parameters and 281.9 GFLOPs, pre-trained on the COCO dataset [53], as it demonstrates superior detection accuracy. We fine-tuned the model on the combined training sets from both hockey datasets for 80 epochs using four NVIDIA GeForce RTX 2080 Ti GPUs, each with 12 GB VRAM. The implementation achieved **97.7%** mAP^{val} (@ IoU = 0.50). These results were obtained with a batch size of eight and a learning rate (LR) of 0.000125, evaluated on three randomly selected test videos from each dataset to ensure unbiased validation.

Discussion. Figure 5.3 demonstrates both successful player localizations and failures cases of our model. Detection failures are inherent to monocular broadcast views, where

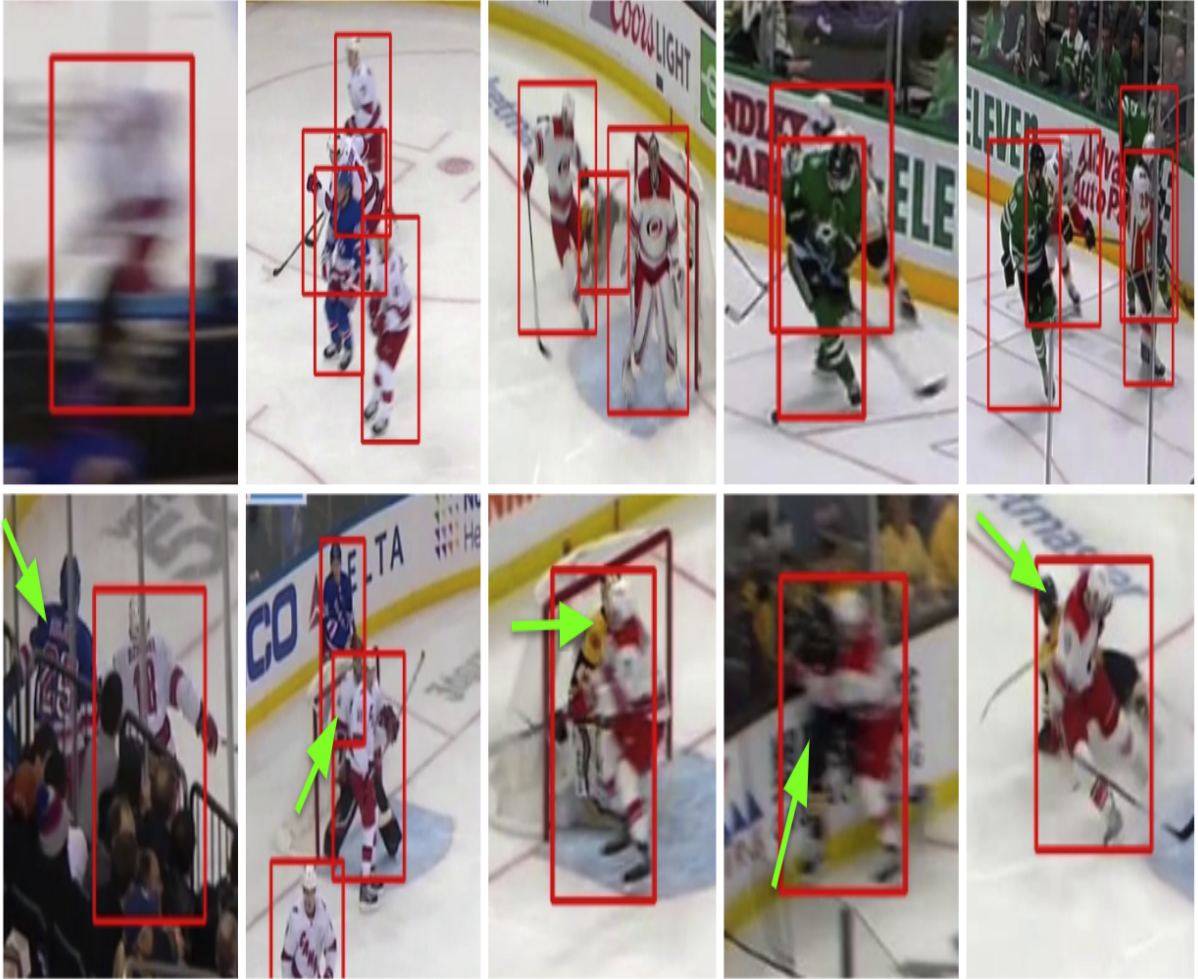


Figure 5.3: Detection results on the [HTD-2](#) test-set. First row shows the robustness of our detections in challenging situations. Second row shows our detector’s limitations, where players are non-localizable ($\tilde{\mathcal{D}} = \{\phi_i\}$, shown using [green](#) arrows).

a single camera perspective cannot reliably detect occluded or severely blurred players. In these cases, multiple players often overlap within the same bounding box, making naive interpolation prone to identity switches during association. We address this by terminating tracklets at the last valid detection and initiating new ones when players become detectable again, preventing tracking errors during such overlaps.

5.3 Re-identification (reID)

In MOT, reID networks generate discriminative feature embeddings that maximize similarity between features of the same object while minimizing similarity between different objects, enabling effective associations. Although primarily developed for multi-camera tracking systems [16, 60, 72], these feature embeddings are equally valuable in monocular tracking for *frame-to-frame* player association. In monocular applications, reID networks function as feature encoders, typically built on convolutional backbones [35] and trained using either cross-entropy [121] or triplet loss [82] to differentiate object features.

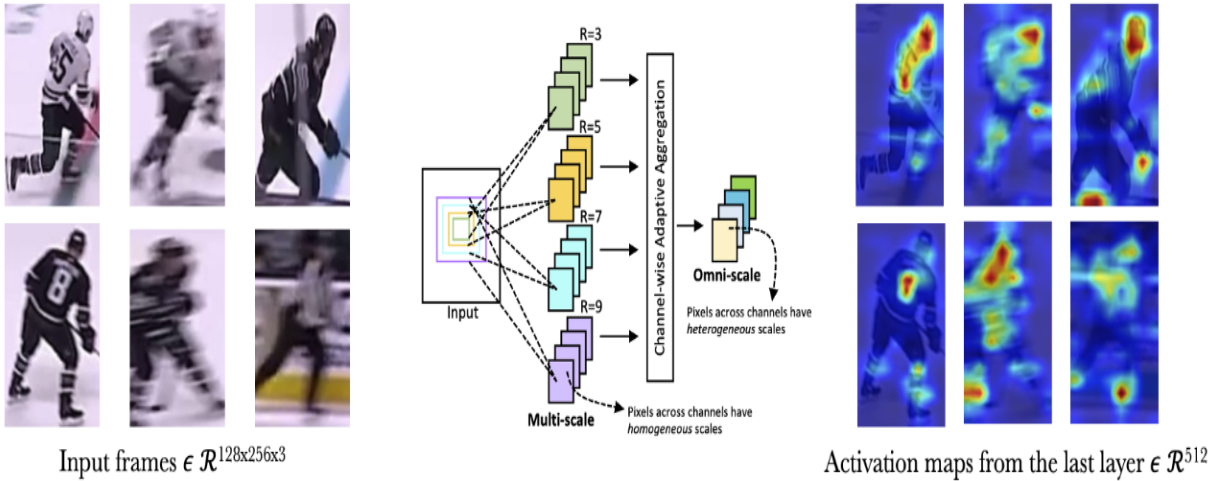


Figure 5.4: Architecture of OSNet [124] adapted for hockey player re-identification. The network processes input images $\mathbb{I} \in \mathbb{R}^{128 \times 256 \times 3}$ through multi-scale convolutions for fine-grained feature extraction, producing 512-dimensional feature vectors. The *omni-scale* feature extraction helps mitigate blur effects (middle column) in player appearances.

The established approach in pedestrian tracking employs reID models pretrained on

external datasets [51, 79, 123] without fine-tuning, as these datasets share similar distributions with pedestrian MOT scenarios. However, this approach is inadequate for hockey applications due to fundamental differences in player movements, visual features, and environmental conditions. This raises a critical question: *Does fine-tuning reID models on ice hockey improve feature quality?* Our experimental results confirm this hypothesis (Fig. 5.6).

For implementation, we selected OSNet [124], which has demonstrated effectiveness in graph-based tracking methods [16, 60] through its multi-scale feature extraction architecture (Fig. 5.4). We adopted the highest-performing backbone pretrained on Imagenet [23] and Market1501 [123], fine-tuning it on our hockey datasets for 25 epochs. We use a multi-step learning rate schedule initialized at $1.5e^{-3}$ with $0.1 \times$ decay every 8 epochs. The objective function combines cross-entropy loss with label smoothening regularization [95] and triplet loss [82], designed to learn distinctive feature representations. Feature matching utilizes cosine distance between query (source) and gallery (target) images, achieving a final accuracy of **68.1%** for ice hockey.

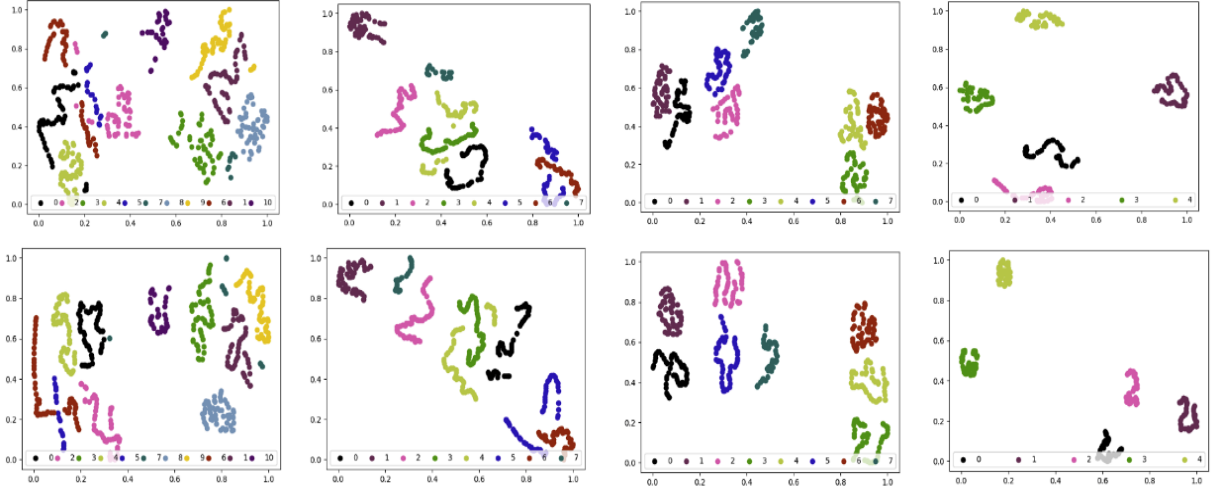


Figure 5.5: t-SNE visualization [99] of temporal reID features encoding player appearance and homography-transformed positional information. Each point (and color) represents a unique player. **Top** row shows feature embeddings from the pre-trained model, where features are less discriminative and closely clustered. **Bottom** row demonstrates improved feature separation after hockey-specific fine-tuning.

Fig. 5.5 demonstrates the improvement in player feature discrimination between pre-

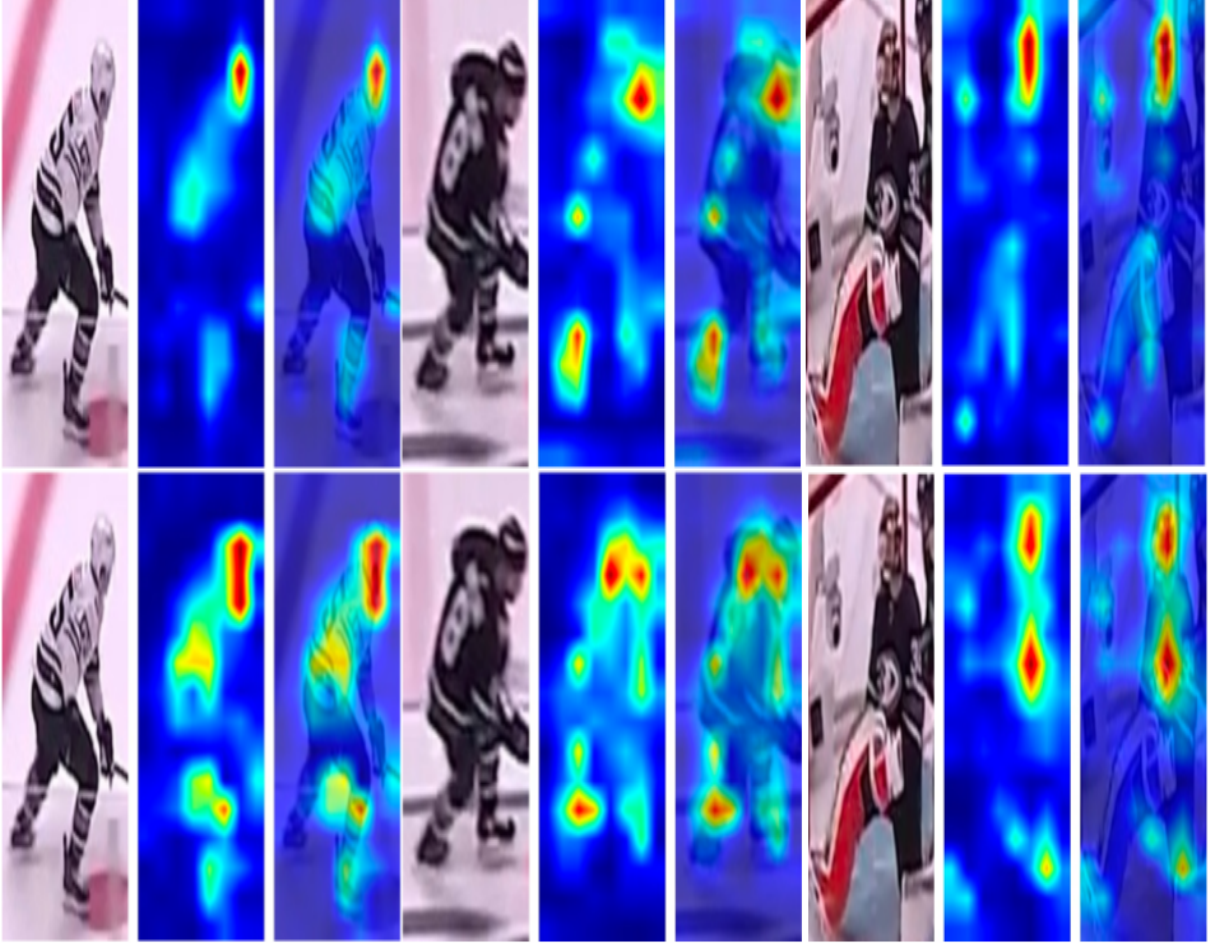


Figure 5.6: Convolutional feature activation maps from the last layer of OSNet [124]. **Top** row illustrates the regions focused by the pre-trained model. **Bottom** row shows the fine-tuned model’s attention areas, exhibiting stronger and more uniform activation across the player form.

trained and fine-tuned OSNet [reID](#) model. Despite the challenge of distinguishing players in hockey due to uniform equipment and full-body gear, our results confirm that effective feature separation is achievable through domain-specific fine-tuning. Figure [5.6](#) presents the feature activation maps from OSNet’s final layer, highlighting regions the model uses for feature discrimination. Notably, both pre-trained and fine-tuned models attend to facial features, an expected behavior given the extensive pre-training on Imagenet [\[23\]](#) with its large collection of human facial images. However, the fine-tuned model extends its attention to players’ back and lower regions and exhibits stronger activation across the entire player form compared to the pre-trained version.

5.4 Homography

Broadcast ice hockey footage, captured from a single camera perspective, presents inherent challenges including frequent player occlusions and camera motion artifacts (blurs, pans, tilts, zooms) as discussed in Chapter [1](#). These limitations affect tracking continuity specifically when players are *localizable* but have intersecting paths, making them susceptible to identity switches even during brief intervals. This is especially pronounced in ice hockey as a fast-paced dynamic sport where players frequently compete together for the puck. This motivates us to explore the key question: ”*Can additional positional cues aid our tracker in maintaining player identities during these scenarios?*”

We find that the answer lies in *homography*, which involves warping camera view-variant broadcast frames onto a common view-invariant overhead rink template (Fig. [5.7](#)). This transformation provides a uniform *top-down* perspective of the rink, reducing positional ambiguities caused by perspective distortions and occlusions. But, since the broadcast camera parameters are unknown, this warping is achieved through an off-the-shelf [SOTA](#) hockey homography estimator [\[83\]](#) that maps each broadcast frame to the overhead template by computing the 3×3 homography matrix, H .

From the warped frames, we need player positional information, which is best represented by their points of contact with the ice surface (footpoints). Following the approach established in ReST [\[16\]](#), we approximate these footpoints at frame F_i using the bottom-center points of players’ bounding boxes, as these points indicate their positions on the rink. The homographic transformation for each player is defined as follows. For player P_i with footpoints (x_i, y_i) in broadcast view:

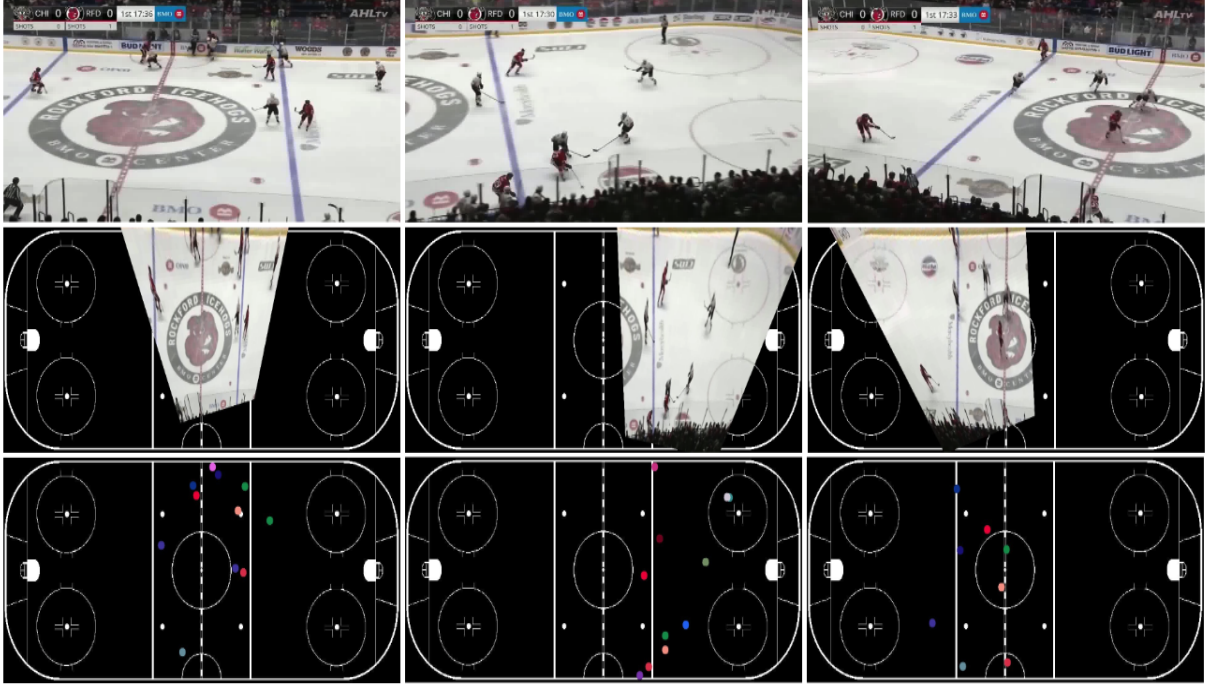


Figure 5.7: Warping broadcast frames onto the overhead rink template using homography. **Top** row shows randomly sampled broadcast frames from three different parts of the rink. **Middle** row shows their warped representations on the view-invariant rink template using homography. **Bottom** row shows the corresponding footpoint visualizations of players as points, where each point portrays a distinct player.

$$p_i = s \begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix} = H \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \quad (5.1)$$

where p_i denotes the homographic projection of player P_i , (x'_i, y'_i) denotes the projected player footpoints in the overhead view, s represents scale factor, and H is the 3×3 homography matrix. This matrix is computed through a two-step process using the homography estimator [83] model: an initial mapping is performed with four points—two from the bottom corners and two from corners at 60% of the image height—followed by a refinement step that adjusts the mapping using the four face-off points in the end zones. The H

matrix helps encode the rotation, shear, translation and perspective distortion information while mapping from the broadcast view to the overhead rink template.

Figure 5.7 shows the warped output from the homography model. We compute the homography matrix, H for each frame from the broadcast view, warp the frames onto the overhead rink template, and derive the projected player footprints (x', y') as input features for the graph tracking framework.

5.5 Conclusion

In this chapter, we introduced the three major building blocks of our multi-player tracking pipeline – detection, [reID](#), and homography. First, in detection, we presented our reasoning behind choosing the YOLOX [122] architecture, its implementation details, success and failure cases in monocular views, and a discussion on how we handle these two cases. Second, in [reID](#), we discussed our choice behind OSNet [124], its implementation details, and the benefits of fine-tuning it on ice hockey. Third, we presented our novel homography integration to derive player positional representations to cue as inputs to our tracking model. In the next chapter, we delve deep into the specifics of our novel tracking network, integrating the above three components to build a [SOTA](#) player tracking framework for ice hockey.

Chapter 6

Part II: Multi-object Tracking Pipeline

6.1 Introduction

Ice hockey differs from most other team sports in posing unique challenges to tracking due to its highly *physical* and *fast-paced* nature. Specifically, MOT in hockey has to address three main challenges: (i) occlusion due to player overlaps in the monocular broadcast PoV; (ii) rapid and non-linear player motion due to dynamic gameplay, and; (iii) frequent camera pan-tilt-zoom effects causing blurs and reduced visibility of players. These limitations increase ID switch errors and give rise to identity-uncertain tracklets (Fig. 6.1).

Recent approaches using deep networks have shown great promise in the accuracy of tracking hockey players from broadcast feeds. Vats *et al.* [101] present a comparison between five different tracking models, fine-tune the MOT neural solver [9] architecture on ice hockey to derive HockeyMOT, and obtain SOTA results. They subsequently use this method to generate tracklets for player identification and team recognition [102]. To the best of our knowledge, this method is currently the SOTA for MOT in ice hockey and therefore, we adopt it as a baseline to evaluate our proposed method. However, there exists two major limitations with this approach: first, their model encodes the *bounding box* attributes of players as graphical edge embeddings which results in identity switches and identity-uncertain tracklets. This is because, when there exists heavy occlusion among players as usually observed in broadcast feeds, their bounding boxes overlap significantly in the monocular view (IoU $\uparrow$), causing misassociation of players (IDsw $\uparrow$) and missing tracks. Second, their node embeddings (Section 6.2.2) are encoded solely based on player

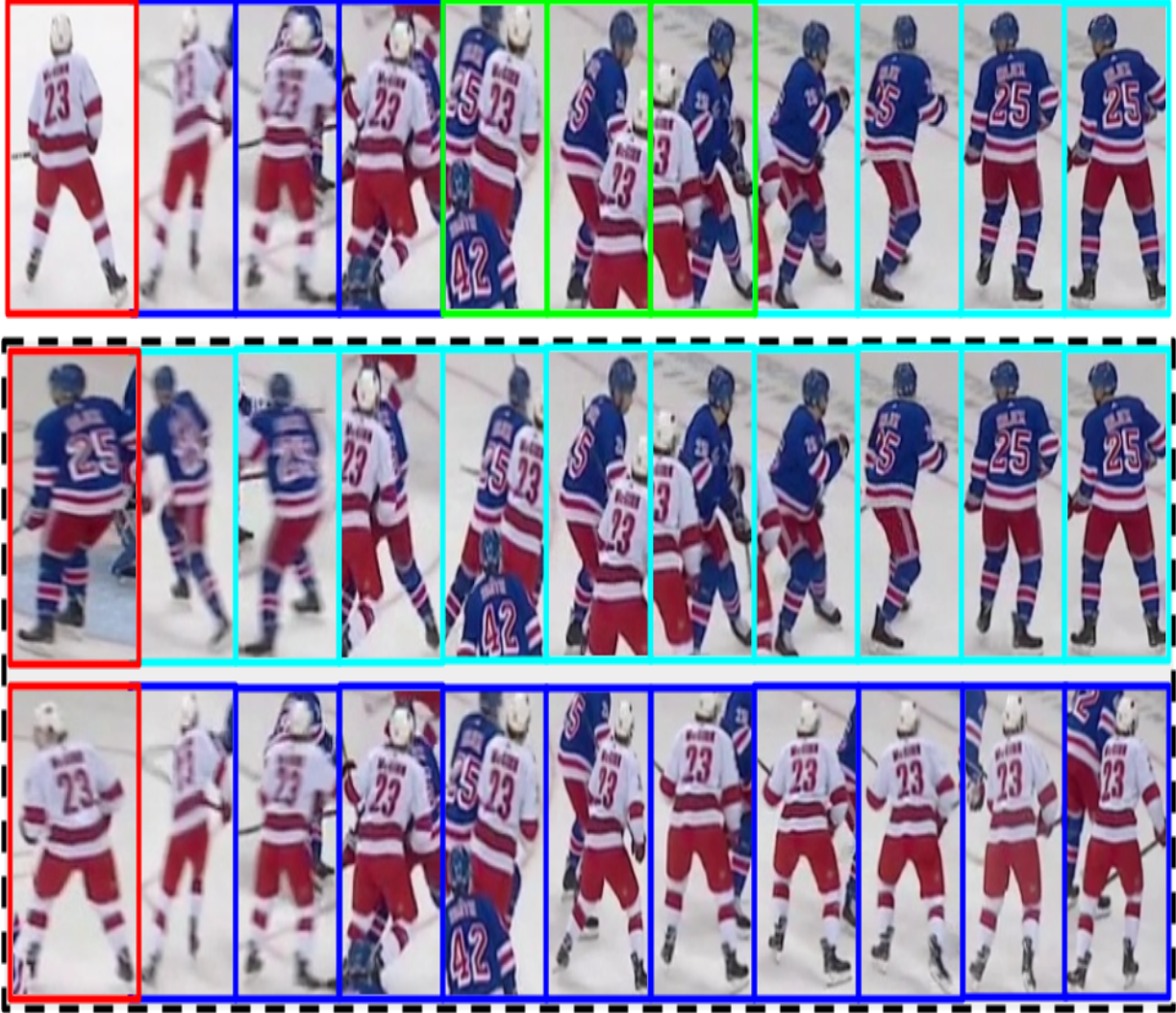


Figure 6.1: Our proposed tracker minimizes identity swaps. **Row 1:** Tracklet predictions from the HockeyMOT baseline [101], where players from the Hurricanes (white jerseys) and Rangers are swapped, leading to identity-inconsistent tracklets. **Rows 2, 3:** Results from our proposed tracker, which fully *removes ID switches* and preserves tracklet identity throughout. The tracklets were sampled at 30 Hz for the first five frames and at 10 Hz thereafter for better illustration.

Frame border color codes:

Red : Starting (reference) frame, **Green** : Players get switched inside a tracklet ($ID_{sw} \uparrow$),
Blue : Hurricanes player tracklet; **Cyan** : Rangers player tracklet;

appearance features, which is ambiguous in hockey due to players’ homogeneity in physical size and build, bulky protective clothing, identical team uniforms and skates, camera blur, and contrastive illumination effects on the rink. With these setbacks present, we ask the question:

”Given a monocular broadcast feed, is it possible to declutter occluded players, minimize identity switches and consistently track players to maintain identity-preserved tracklets?”. Our experimental results demonstrate that this is *indeed achievable*.

We formulate MOT as a link prediction problem in the graph domain as graphs offer a natural way of representing players and their relationships. The *novelty* of our approach lies in utilizing homography to obtain reliable spatial (positional) cues for players on the rink. Specifically, we map each frame from the broadcast feed to an overhead rink template using an off-the-shelf hockey homography estimator [83], since broadcast camera parameters are unknown. This estimates a homography transformation matrix $\{H \in \mathbb{R}^{3 \times 3}\}$ that converts broadcast-view 2D footpoint coordinates $\{(x, y) \in \mathbb{R}^2\}$ to the view-invariant 2D footpoint coordinates $\{(x', y') \in \mathbb{R}^2\}$ in the overhead rink space (Fig. 6.2). We encode these projected positions along with player reID features as graph embeddings, and propagate them throughout the graph network using the message passing networks (MPN) [4, 30] framework. Our model follows the popular tracking-by-detection TBD paradigm; we show evaluations with both ground truth detections and off-the-shelf detector outputs [76] consistent with the norms followed in our baseline. Through this approach, our model is able to utilize the overhead positional information of players as additional cues to track with consistency during occlusions, blurs, and dynamic player movements. Fig. 6.1 illustrates the improved tracking performance of our method compared to the baseline [101].

The contributions of this chapter can be summarized as follows:

- We propose a simple yet effective spatio-temporal graph network utilizing the MPN framework for consistent player tracking from broadcast feeds.
- We introduce a novel homography-based approach that provides robust positional cues to the graph network during challenging tracking scenarios.
- We demonstrate significant improvements in identity switch score (IDsw) and identify F1 score (IDF1) metrics on two broadcast hockey datasets (Section 3).

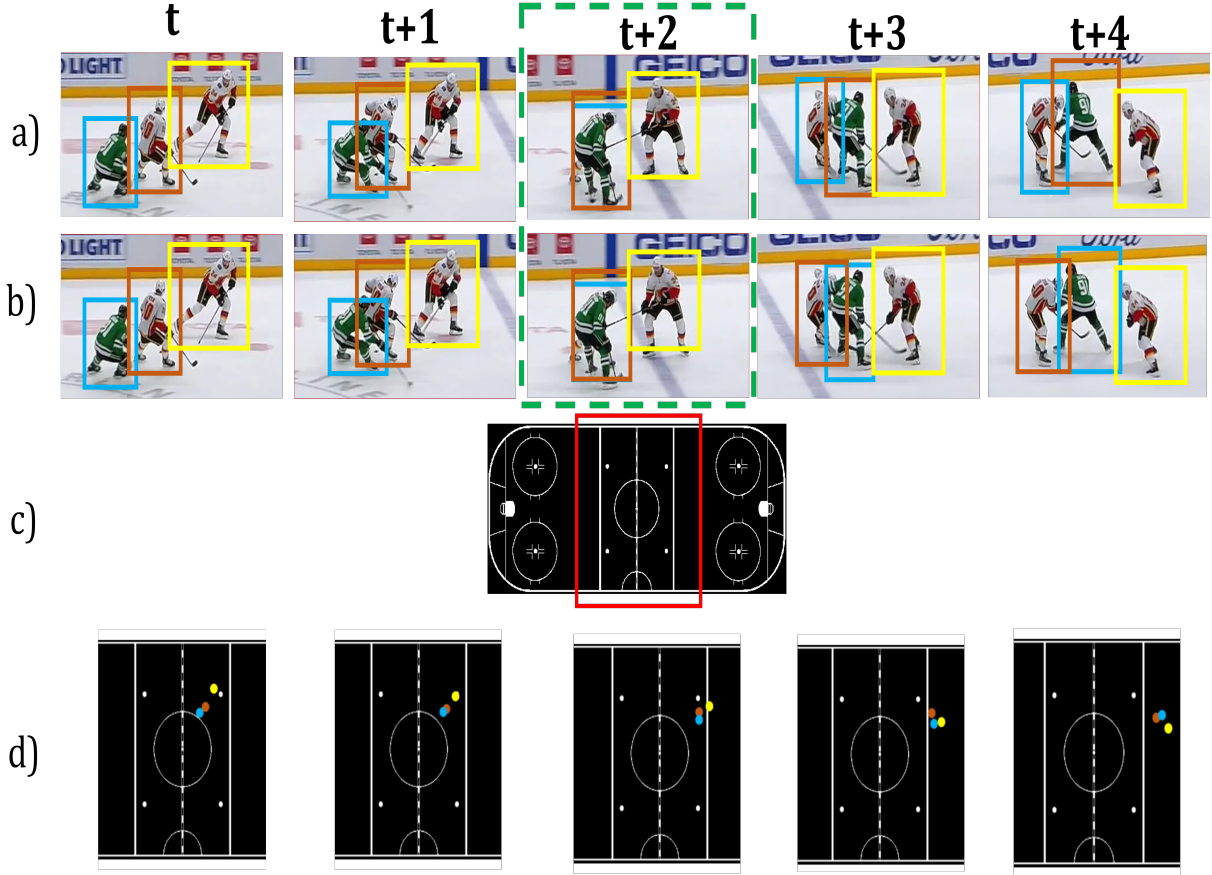


Figure 6.2: *What makes our approach novel?* (a) During a significant occlusion scenario ($t + 2$), other trackers lead to an **IDsw**; (b) Our method consistently tracks players before and after occlusion; (c) Overhead rink template used for homography projection; (d) Player footpoint coordinates mapped to the overhead template. At ($t + 2$), there is a clear distinction between overlapping players from the top view. This information aids our tracker in preserving player identities within tracklets.

6.2 Proposed Method

6.2.1 Overview

In this work, we leverage GNNs to build a *frame-by-frame* spatio-temporal graph tracker, concatenating homography footprint projections with reID embeddings [124] as node features, and their relative distances as edge features. Note that the nodes represent players and their temporal connections form tracklets. We utilize the MPN framework [4, 30] to propagate player features across consecutive frames, and update the node and edge embeddings at each message passing step. The association between two consecutive frames f^i and f^j s.t. $j > i$, is framed as a bipartite graph matching problem, using a binary classifier with a sigmoid final layer to output association probabilities. During inference, we post-process each graph by pruning low-confidence associations and resolving many-to-one violations. Finally, we assign unique IDs to nodes where a node either inherits an ID from its prior connection or receives a new ID otherwise (Fig. 6.3).

6.2.2 Graph formulation

We formulate MOT as a bipartite graph matching problem and deterministically create one graph per frame from the given sequence. Each node in this graph points to one player in that frame and the edges correspond to their relationships with neighboring nodes (players) in other frames. Let us consider the undirected graph $G_t = (V_t, E_t)$ with bidirectional connections, where V_t denotes the vertex set and E_t denotes the edge set at time t . $NodeFeatures(.)$ represents the concatenation of appearance features and homographic projections of players at time t , and $E_t = (V_i \times V_j)$ represents a one-one mapping between vertex sets v_i and v_j s.t. $i \neq j$, meaning associations do not occur between players within the same frame.

Node formulation

Each node $v_i \in V_t$ represents one unique player i and contains: frame ID $f^i \in \mathbb{R}^1$ (timestamp), player ID $t_i^{id} \in \mathbb{R}^1$ (ground truth, only used for training), bounding box coordinates $b^i = \{x_{min}, y_{min}, width, height\} \in \mathbb{R}^4$, reID appearance features $r_i^{id} \in \mathbb{R}^{512}$, and homographic projection coordinates $p_i = \{x'_i, y'_i\} \in \mathbb{R}^2$.

The **reID** features are generated for each node i via an off-the-shelf **reID** network [124] fine-tuned on ice hockey, described by:

$$r_i^{id} = \text{reID}(b^i|_{crop}) \quad (6.1)$$

where $(b^i|_{crop})$ denotes the cropped bounding box for the i^{th} player. For homography, the left footpoint $x_i = (x_i + \frac{\text{width}_i}{2})$ and the right footpoint $y_i = (y_i + \text{height}_i)$ for player i are projected onto the overhead rink template as:

$$p_i = H_i(x_i, y_i) \quad (6.2)$$

with H_i being the 3×3 homography matrix (Eq. 5.1).

Edge Formulation

Each edge $e_{ij} \in E_t$ is represented as the interconnection between two players from two distinct frames. It is encoded as the concatenation of relative appearance Δr_{ij}^{id} and positional Δp_{ij} features between the pair of nodes v_i and v_j , where $e_{ij} = v_i \times v_j, i \neq j$. This can be represented as:

$$\Delta r_{ij}^{id} = [\|r_i^{id} - r_j^{id}\|_1, \text{cosine_similarity}(r_i^{id}, r_j^{id})] \quad (6.3)$$

$$\Delta p_{ij} = [\|p_i - p_j\|_1, \|p_i - p_j\|_2] \quad (6.4)$$

$[\cdot, \cdot]$ denotes the concatenation of Euclidean distance and cosine similarity for Δr_{ij}^{id} and Euclidean and Manhattan distance for Δp_{ij} , respectively. This consideration is inspired from the MC-MOT framework [60] to obtain high-dimensional distinctive features.

6.2.3 Temporal graph association

To facilitate the association of players across consecutive frames, we design a simple temporal graph network to correlate player features. We iteratively correlate the learned graph G_{t-1}^T at time $t-1$ with the adjacent graph G_t at time t to form a new graph G_t^T . Assuming this as the n^{th} iteration, each node $v_{t-1} \in G_{t-1}^T$ contains aggregated embeddings from $n-1$ iterations and is connected with the new nodes $v_t \in G_t$ to form the new temporal graph G_t^T at time t (Algorithm 1). The edge set thus created can be denoted as:

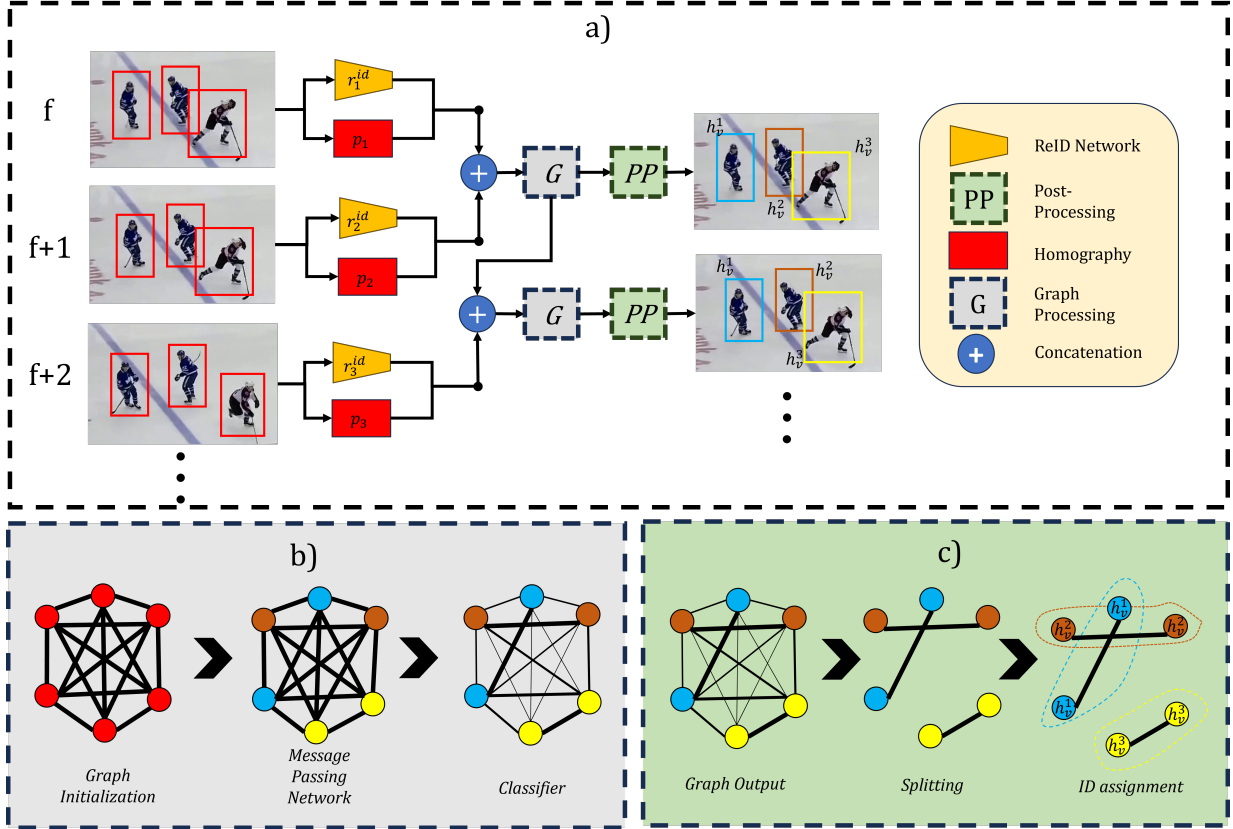


Figure 6.3: Proposed tracking architecture. **(a)** The general pipeline of our spatio-temporal graph. We estimate the bounding boxes (detection), **reID** features, and homographic projections, and pass them as input embeddings to our graph tracker. **(b)** A macroscopic illustration of the graph network G , denoting the three stages of graph initialization, **MPN**, and classification. **(c)** A macroscopic illustration of the post-processing stages (PP), where we prune, solve graph violations, and assign player IDs.

$$e_{t-1,t} = \begin{cases} 1, & \text{if } v_i \times v_j \forall v_i \in G_{t-1}, v_j \in G_t \\ 0, & \text{otherwise} \end{cases}$$

This aggregation of uniterated graphs with learned graphs helps propagate learned features and assign consistent identities for the same player (Fig. 6.3).

Algorithm 1 Model Inference

```

1: Given ground truth  $y_{e_{ij}}$ , players  $P_k \in G_{t-1} \oplus G_t$ ,  $i = 0$  &  $w = 2$ ;
2: for  $t = 0$  to  $(T - 1)$  do
3:   Initialize  $h_{v_i}^0, h_{v_j}^0$  and  $h_{e_{ij}}^0$  for  $G_{t-1} \oplus G_t$ 
4:   if  $i > 0$  then
5:     Replace  $G_{t-1}$  with  $G_t^{learned}$ 
6:   end if

7:    $h_{v_i}^0, h_{v_j}^0 = f_v^{FE}(h_{v_i}^0, h_{v_j}^0)$ 

8:    $h_{e_{ij}}^0 = f_e^{FE}(h_{e_{ij}}^0)$ 

9:   while  $(l \geq L)$  do
                                 $h_{v_i}^L, h_{v_j}^L, h_{e_{ij}}^L = \text{MPN}(G_{t-1} \oplus G_t, h_{v_i}^{l-1}, h_{v_j}^{l-1})$ 
                                 $\hat{y}_{e_{ij}} = f_{cls}^L(h_{e_{ij}}^L)$ 
10:  end while

11:   $G_t^{learned} = \text{Post-Processing}(G_{t-1} \oplus G_t, \hat{y}_{e_{ij}}^L)$ 
12:   $i++$ ;
13: end for

```

6.2.4 Message passing network

We adopt the message passing network [MPN](#) framework introduced by Gilmer *et al.* [30], to propagate the node and edge information across the graph G . Message passing intuitively helps the graphs learn their neighboring features; each edge learns about the projection and appearance feature of its neighboring nodes and each node learns about the geometric

features of its neighboring edges. To begin with, we initialize the node embeddings and edge embeddings as:

$$h_{v_i}^{(0)} = f_v^{FE}([r_{v_i}^{id}, p_{v_i}]) \quad (6.5)$$

$$h_{e_{ij}}^{(0)} = f_e^{FE}([\Delta r_{e_{ij}}^{id}, \Delta p_{e_{ij}}]) \quad (6.6)$$

Note that we add p and Δp to both the node and edge features respectively to propagate player positional information throughout the graph. Given these initial embeddings, we perform L iterations of edge updates and node updates as two separate steps, as followed by well-established methods [4, 9, 16, 60, 101].

Edge update

We utilize a learnable [MLP](#) to perform edge encoding for $l = \{1, \dots, L\}$ steps using the source and destination nodes connected by the edge:

$$(v \rightarrow e) \quad h_{e_{ij}}^{(l)} = f_e^{ME}([h_{v_i}^{(l-1)}, h_{v_j}^{(l-1)}, h_{e_{ij}}^{(l-1)}]) \quad (6.7)$$

where, f_e^{ME} is the edge encoder. This leads to the sharing of appearance and projection embeddings from the neighboring nodes h_{v_i} and h_{v_j} to its connecting edge $h_{e_{ij}}$.

Node update

Similar to Eq. 6.7, we utilize another learnable [MLP](#) to perform node update for $l = \{1, \dots, L\}$ steps, using the aggregated messages coming from its neighboring nodes:

$$(e \rightarrow v) \quad h_{v_i}^{(l)} = \sum_{j \in \mathcal{N}(v_i)} f_v^{ME}([h_{v_j}^{(l-1)}, h_{e_{ij}}^{(l-1)}]) \quad (6.8)$$

where, f_v^{ME} is the node encoder and $\mathcal{N}(v_i)$ denotes the neighboring nodes of v_i . Note that f_v^{ME} and f_e^{ME} are two separate networks with different dimensions, but share the same [MLP](#) architecture (Table 6.1).

In both the above updates, the [MLP](#) encodes all the information into a higher-dimensional feature space. The message passing step L is akin to the receptive field found in [CNNs](#), and a higher value of L corresponds to farther propagation of information in the graph, but at the cost of computational complexity (Section 6.5.2).

Layer Num.	Type	Output Size
Feature Encoders		
Nodes ($\mathcal{N}_v^{enc}$)		
0	Input	515 $\rightarrow$ 128
1	FC + GELU	4 $\rightarrow$ 32
Edges ($\mathcal{N}_e^{enc}$)		
0	Input	4 $\rightarrow$ 8
1	FC + GELU	8 $\rightarrow$ 6
Message Passing Network		
Node Update ($\mathcal{N}_v^{ME}$)		
0	Input	38 $\rightarrow$ 64
1	FC + GELU	64 $\rightarrow$ 32
Edge Update ($\mathcal{N}_e^{ME}$)		
0	Input	70 $\rightarrow$ 32
1	FC + GELU	32 $\rightarrow$ 6
Classifier ($\mathcal{N}_{class}$)		
0	FC + GELU	6 $\rightarrow$ 4
1	FC + Sigmoid	4 $\rightarrow$ 1

Table 6.1: Featuring encoder and MPN layer details.

Classification

We propose to learn the association between nodes as a *link prediction task* by framing player tracking as a graph partition problem. That is, after L iterations, we perform binary classification to predict the edge probabilities $\hat{y}_{e_{ij}}$ connecting nodes v_i and v_j , as:

$$\hat{y}_{e_{ij}} = f^{cls}(h_{e_{ij}}) \quad (6.9)$$

where f^{cls} is a yet another learnable MLP with a sigmoid final layer to output probabilities. During training, the binary ground truth labels $y_{e_{ij}}$ and their corresponding predictions $\hat{y}_{e_{ij}}$ are compared to find the sigmoid focal loss [81]. During inference, the edges with low confidence scores (weak connections) are removed.

6.2.5 Post-processing

We adopt a post-processing step during inference to prune, resolve many-to-one violations, and propagate IDs consistently in our learned graphs.

Pruning

This is the first step in refining the predicted edge confidence score $\hat{y}_{e_{ij}}$ by our classifier. We define a confidence threshold hyperparameter ξ , where:

$$\text{Pruned}(\hat{y}_{e_{ij}}) = \begin{cases} \hat{y}_{e_{ij}}, & \text{if } \hat{y}_{e_{ij}} > \xi \\ 0, & \text{otherwise} \end{cases} \quad (6.10)$$

This helps eliminate low-confidence edges and retains only the most strongest correspondences. For our experiments, we have used $\xi = 0.9$ as followed in ReST [16] to retain only the most strongest connections. Similar to ReST [16], we also remove many-to-one edge violations, wherein, there can only exist at most one connection between two nodes within the connected graphs. This ensures that we represent each player identity through a unique connected component, since no two different players should share the same identity within the same frame.

Assigning IDs

The final graph contains unique connected components where each component represents a unique player tracklet. As we iterate temporally, we assign a new ID when there exists no previous connected components for a node or propagate the same ID otherwise.

$$\text{ID}(h_{v_j}) = \begin{cases} \text{Same ID, if } h_{v_j} = h_{v_i}; \forall j \in G_t, i \in G_{t-1} \\ \text{New ID, otherwise.} \end{cases} \quad (6.11)$$

6.3 Implementation Details

This section contains details about datasets used, training scheme, evaluation metrics and experimental results.

6.3.1 Datasets

We evaluate our approach using both [HTD-1](#) and [HTD-2](#). For fair comparison, we retrained the baseline HockeyMOT [\[101\]](#) model on [HTD-1](#). To demonstrate generalizability, we use the [HTD-2](#) test set for cross-dataset validation without training on it, thus avoiding data leakage. While both datasets are sourced from broadcast [NHL](#) games, they differ in their challenging scenarios – featuring varying levels of occlusions, motion blur, and player movements due to different teams’ gameplay styles.

6.3.2 Training details

As described in Section [5](#), we fine-tune the YOLOX [\[122\]](#) detector and OSNet [\[124\]](#) reID network on both the hockey tracking datasets. Since the homography model [\[83\]](#) is already pre-trained on ice hockey, we directly use it to infer player footpoint projection coordinates. We concatenate the $\mathbb{R}^{512}$ reID feature embeddings thus obtained, with the $\mathbb{R}^3$ homography projections, and encode them into $\mathbb{R}^{32}$ node embeddings (Eq. [6.5](#)) using the node feature encoding MLP ($\mathcal{N}_v^{enc}$). Similarly, we encode the $\mathbb{R}^4$ edge features into 6-D edge embeddings (Eq. [6.6](#)) using the edge feature encoding MLP ($\mathcal{N}_e^{enc}$). We run the MPN for $L = 6$ iterations and pass the final graph output into our binary classifier for predicting edge association probabilities.

During training, we utilize the ground truth player annotations and calculate the prediction losses using focal loss [\[81\]](#).

$$\text{Focal Loss} = \sum_l^L \sum_{e_{ij} \in G_{t-1} \cup G_t} \mathcal{L}(\hat{y}_{e_{ij}}^l, \mathbf{1}(y_{e_{ij}})) \quad (6.12)$$

where, $\hat{y}_{e_{ij}}^l$ is the edge prediction at iteration l and $\mathbf{1}(y_{e_{ij}})$ is the ground truth indicator function. This function is given by:

$$\mathbf{1}(y_{e_{ij}}) = \begin{cases} 1, & \text{if } v_i = v_j \\ 0, & \text{otherwise} \end{cases} \quad (6.13)$$

Algorithm 2 Model Training

```
1: Given g.t  $y_{e_{ij}}$ , players  $P_k \in G_{t-1} \oplus G_t$  &  $w = 2$ ;  
2: for  $t = 1$  to  $(T - 1)$  do  
3:   Initialize  $h_{v_i}^0, h_{v_j}^0$  and  $h_{e_{ij}}^0$  for  $G_{t-1} \oplus G_t$   
  
4:    $(h_{v_i}^0, h_{v_j}^0) = f_v^{FE}(h_{v_i}^0, h_{v_j}^0)$   
  
5:    $h_{e_{ij}}^0 = f_e^{FE}(h_{e_{ij}}^0)$   
  
6:   while  $l \leq L$  do:  
        $h_{v_i}^L, h_{v_j}^L, h_{e_{ij}}^L = MPN(G_{t-1} \oplus G_t, h_{v_i}^{l-1}, h_{v_j}^{l-1})$   
        $\hat{y}_{e_{ij}} = f^{cls}(h_{e_{ij}}^L)$   
7:   end while  
8:    $graph\_loss = sigmoid\_focal\_loss(\hat{y}_{e_{ij}}, y_{e_{ij}})$   
9:    $graph\_loss.backward()$   $\triangleright$  Backpropagation  
10:   $optimizer.step()$   $\triangleright$  Update parameters  
11: end for
```

Method	MOTA % $\uparrow$	FP $\downarrow$	FN $\downarrow$	IDsw $\downarrow$	IDF1 $\uparrow$ %
SORT [7]	92.4	2403	5826	673	53.7
DeepSORT [111]	94.2	1881	4334	528	59.3
FairMOT [120]	91.9	1179	7568	768	61.5
Tracktor [5]	94.4	1706	4216	687	56.5
ByteTrack [†] [119]				<u>298</u>	<u>80.7</u>
Hockey MOT [101]	94.5	1653	4394	431	62.9
Hockey MOT [†] [101]				1056	71.8
MOTH (Ours)	95.4	1924	4323	453	71.3
MOTH (Ours)[†]				151	95.1

Table 6.2: Evaluation results on HTD-1.[†] indicates ground truth annotations used for evaluation. Best results are in red, second best results are underlined.

where, for every player match $y_{e_{ij}} = 1$ else 0. We use the Adam optimizer [42] without weight decays to update our model parameters. The learning rate (LR) is initialized at 0.01, with a gradual warmup for the first 10 epochs and a cosine annealing schedule (min. LR = 0.001) thereafter. We trained our model for 30 epochs with a batch size of 16 on a single NVIDIA GeForce RTX 4090 GPU with 24 GB RAM, 2.5GHz clock speed, and performed validation after every two training epochs.

6.4 Experimental Results

We prefer evaluations directly based on the ground truth annotations to obtain a true picture of our tracking performance; but, to be consistent with the baseline, we also evaluate on detection bounding boxes as followed in HockeyMOT [101] for a fair comparison. Note that the baseline uses the Faster R-CNN [76] detector. Table 6.2 shows our experimental results on HTD-1 compared to the existing methods benchmarked in Vats *et al.* [101] and the HockeyMOT baseline. For ground truth evaluations (notated with [†]), our model outperforms the baseline by roughly $10\times(\downarrow)$ reduction in the IDsw score and a **32.45%** ($\uparrow$) increase in the IDF1 score, respectively. This signifies that our approach satisfies our core objective of minimizing identity switches and maintaining identity-preserved tracklets. We see similar trends with the detection-based results, where our model surpasses the baseline by **13.35%** ($\uparrow$) increase in IDF1, but negligibly decreases in the IDsw score. This

NHL teams (clips)	No.of frames	FPS (Hz)	IDsw ($\downarrow$)	IDF1 (%) ($\uparrow$)
CAR vs. BOS	2606	60	3251	26.3
CAR vs. NYR	2137	60	2508	23.0
CGY vs. DAL	1330	60	1105	23.4
CHI vs. TOR	1618	30	1006	19.7
CNTRL vs. PAC	1671	30	724	17.9
PIT vs. SJ	1391	30	827	26.2
STL vs. SJ	2258	60	2165	27.7
HockeyMOT	13,011		11,707	24.6
CAR vs. BOS	2606	60	11	89.3
CAR vs. NYR	2137	60	7	93.1
CGY vs. DAL	1330	60	6	92.4
CHI vs. TOR	1618	30	6	90.9
CNTRL vs. PAC	1671	30	13	90.5
PIT vs. SJ	1391	30	7	97.4
STL vs. SJ	2258	60	7	96.3
MOTH (Ours)	13,011		60	92.8

Table 6.3: Cross-dataset validation on [HTD-2](#). The results demonstrate our tracker’s generalization on in-the-wild datasets and zero-shot capabilities.

is expected because of an additional post-processing step employed in the baseline for detection-based inputs as discussed in section 4.3.2. The **95.4% MOT** accuracy obtained by our model is higher than all other compared methods, due to the ability of our detector [122] in balancing the trade-off between **FPs** and **FNs**.

Table 6.3 shows the cross-validation results of our model on **HTD-2**. Our model achieves **92.8% IDF1** score and just **60 IDsw**, compared to the baseline. Two critical observations can be made from this result: **(i)** our model generalizes well to *in-the-wild* ice hockey sequences belonging to different games, **(ii)** our model’s performance is not affected by an increase in frame sampling rate (both 60 and 30 Hz sampled frames present in **HTD-2**).

6.5 Ablation Studies

We present empirical results from experimenting with different network choices and our reasoning behind the proposed model architecture. This includes: **(i)** effect of homography inputs, **(ii)** effect of different message passing steps, L , **(iii)** video-wise quantitative results on the 13-test set videos from **HTD-1**, **(iv)** t-SNE [99] visualization of four random clips before and after graph association.

6.5.1 Effect of homography

Our method’s novelty lies in the inclusion of homography input cues to simulate a *bird’s-eye view* of player positions to help with decluttering during occlusions. Table 6.4 shows that only using appearance features underperforms by a noticeable margin in each test sequence, while including homography cues improves **IDF1** score and minimizes **IDsw**. This validates that additional player positional information from a view-invariant space, especially for tracking in monocular **PoV** helps reduce identity switches and improves the overall tracking performance.

Test video #	Appearance	Homography	IDsw ↓	IDF1 ↑
1	✓	✓	31	94.8
	✓		8	95.7
2	✓	✓	37	90.7
	✓		27	94.1
3	✓	✓	19	93.5
	✓		14	93.9
4	✓	✓	13	93.8
	✓		4	98.2
5	✓	✓	53	89.2
	✓		13	97.5
6	✓	✓	25	90.0
	✓		12	94.0
7	✓	✓	28	87.9
	✓		11	88.8
8	✓	✓	35	79.0
	✓		2	97.9
9	✓	✓	16	92.8
	✓		6	97.1
10	✓	✓	25	81.5
	✓		23	90.8
11	✓	✓	25	94.1
	✓		4	97.6
12	✓	✓	23	88.2
	✓		16	92.9
13	✓	✓	11	94.9
	✓		10	97.7

Table 6.4: Effect of including homography cues on IDsw and IDF1 scores. The results are based on the 13 test videos from [HTD-1](#). Best results are **emboldened**.

6.5.2 Number of message passing steps, L

The number of message passing steps L is analogous to the receptive field of CNNs. Our empirical results shows that the best value for $L = 6$ steps. $L < 6$ is inadequate to propagate information across neighboring nodes and edges, leading to lesser capability (as evident from the increase in IDsw $\uparrow$), while $L > 6$ leads to a possible over-fitting scenario on the training data due to redundant message passing. Thus, L should be treated as a hyperparameter, which is best chosen through empirical observation for specific applications.

# of message passing steps	IDsw $\downarrow$	IDF1 $\uparrow$
2	173	85.5
4	51	93.4
6	27	94.1
8	34	92.2
10	42	84.7
12	107	67.4
14	388	43.3

Table 6.5: Effect of message passing steps on IDsw and IDF1 scores. The tested sequence was randomly sampled from HTD-1. Best results are marked in red.

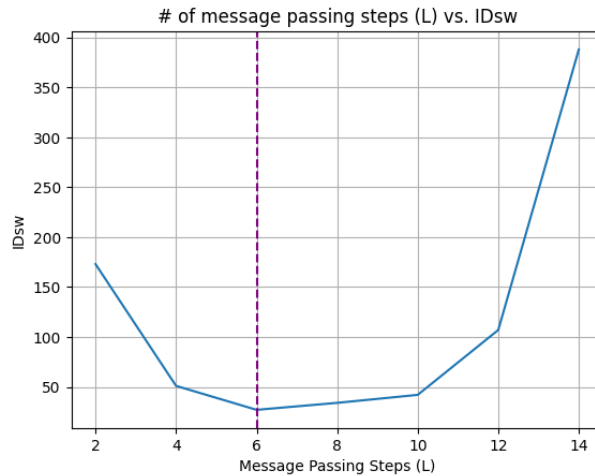


Figure 6.4: A comparison of the number of IDsw incurred vs. the number of message passing steps, L

6.5.3 Video-wise quantitative results

Table 6.6 presents a detailed summary of our test results on the 13 test videos in HTD-1. Our model achieves SOTA results in each individual sequence.

Test Video #	No. of frames	IDsw $\pm\downarrow$	IDF1 $\pm\uparrow$ %	IDsw ($\downarrow$)	IDF1 ($\uparrow$) %
1	1,303	85	80.7	8	95.7
2	1,210	113	75.4	27	94.1
3	878	114	77.0	14	93.9
4	1,033	45	79.7	4	98.2
5	1,071	154	54.9	13	97.5
6	1,441	81	76.1	12	94.0
7	682	58	63.0	11	88.8
8	1,056	50	80.7	2	97.9
9	1,497	87	64.4	6	97.1
10	890	54	55.9	23	90.8
11	1,109	95	80.3	4	97.6
12	1,372	56	72.44	16	92.9
13	795	64	75.15	11	97.7
Overall	14,337	1056	71.98	151	95.1

Table 6.6: Sequence-wise results on HTD-1 test set. This table shows ground truth evaluation results for both the baseline and our methods. $\pm$ denotes the HockeyMOT [101] baseline.

6.5.4 t-SNE visualization

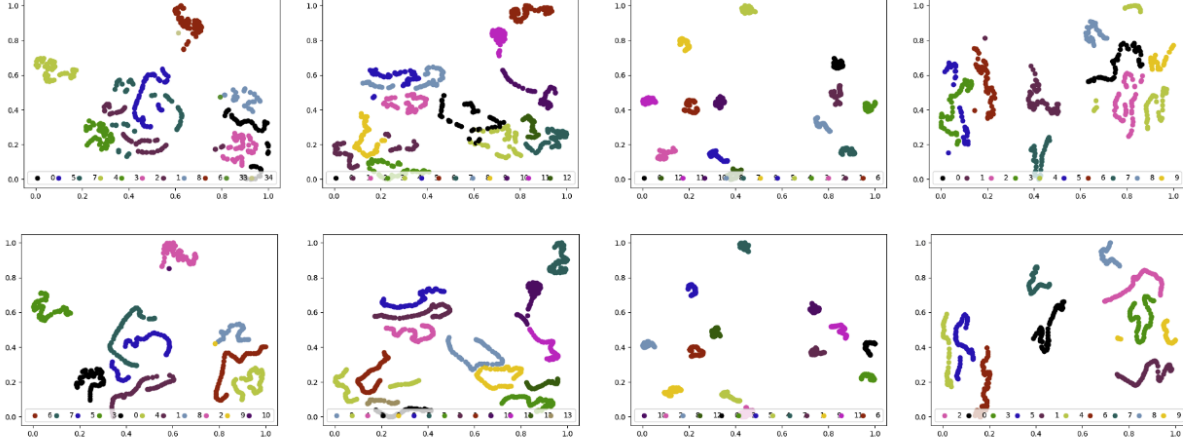


Figure 6.5: t-SNE [99] visualization for the graph association. **(top-bottom)** Before and after the temporal graph association stage. Our model promotes further disentanglement of the spatio-temporal features for better tracking ID assignments.

The t-SNE [99] visualization was performed on 4 randomly chosen sequences from HTD-2 for 150 frames. It can be observed from Fig. 6.5 that similar features (players) are clustered together while dissimilar features are pushed apart. This shows that our model promotes disentanglement of the spatio-temporal features for better tracking ID assignments.

6.6 Summary

In this chapter, we present a *novel* MOT approach for ice hockey, using message-passing GNNs with homography cue inputs. Our model handles occlusions and out-of-FoV scenarios by creating multiple, *identity-preserved* player tracklets for consistent tracking. Experimental results show SOTA performance on HTD-1 and excellent generalization capabilities on HTD-2. The model is capable of disentangling cluttered players in monocular view, resulting in a simple and cost-effective tracking solution. Ablation studies show that homography cues reduce identity switches significantly across all sequences. With this work, we take a major step towards handling occlusions and achieving identity-preserved tracklets.

Chapter 7

Conclusion and Future Research

This dissertation introduces a novel approach for **MOT** in ice hockey, addressing the challenge of player identity switches during occlusions to ensure identity-preserved tracklets. The following sections summarize this thesis’s contributions, suggest future research directions, and discuss the applicability and impact of this thesis.

7.1 Summary of Contributions

The contributions of this work are as follows:

- **Introduction of open-source **HTD-2** dataset.** This thesis introduces hockey tracking dataset - 2 (**HTD-2**) – an *open-source* multi-player tracking dataset for ice hockey using broadcast **NHL** games. We justify its development, provide comprehensive metadata analysis, and detail the curation strategies. This dataset is open-sourced at: <https://github.com/harshap-ai/VIP-HTD>.
- **Public benchmark for ice hockey tracking.** We present a reference *public* benchmark for **MOT** in ice hockey using six different tracking methods. It is comprehensive with initial approaches using Kalman filters and Hungarian matching (SORT, OC-SORT), advanced **JDT** techniques (Tracktor, ByteTrack), and contemporary methods leveraging graph-based (HockeyMOT) and transformer-based (Trackformer) architectures. We present a detailed analysis on the merits and drawbacks of each approach, and substantiate it with experimental results. This benchmark is intended to serve as a reference standard for evaluating future tracking methods in ice hockey.

- **Novel tracking approach using homography.** The core contribution of this thesis is the novel tracking pipeline with YOLOX [122] for detection, OSNet [124] for [reID](#), and a new graphical [MPN](#) designed specifically for ice hockey for tracking. The novelty of our network lies in its ability to incorporate homography for player positional representation inputs. This is aimed at minimizing player identity switches during occlusions and overlaps, to consistently track players throughout an ice hockey game. Aside from demonstrating better tracking performance, we also prove the generalization and zero-shot learning capabilities of our model. Experimental results on [HTD-1](#) show that our approach outperforms the current [SOTA](#) baseline model by 10× reductions in the [IDsw](#) score and **32.45%** increase in [IDF1](#) score, respectively. Generalization-wise, our approach achieves **92.8%** in [IDF1](#) and **60** in the [IDsw](#) score on 7 test clips from [HTD-2](#). This superior performance establishes our method as the new state-of-the-art ([SOTA](#)) for [MOT](#) in ice hockey.

7.2 Future Research Directions

Tracking is the fundamental first step towards enabling a plethora of downstream tasks. This work forms the foundation for future research in the following directions:

7.2.1 Unique player identification

Unique player identification involves consistently identifying a player throughout the given sequence, irrespective of challenges like occlusion and duration of the sequence. Current literature prefers recognizing players through their jersey numbers; but, this is often unreliable due to blurs, jitters, and frequent occlusions present in monocular broadcast view, exacerbated by other challenges such as two players from opposite teams sharing the same jersey number. An alternate approach can be stitching together multiple identity-preserved tracklets belonging to the same player using a strong [reID](#) model conditioned on less explored features such as appearance, homography, handedness, etc. Our experiments with [reID](#) in section 5 (Fig. 5.5) show that player features can be adequately disentangled by choosing the right set of training instructions and loss functions.

7.2.2 Multi-view tracking

Monocular tracking, as seen throughout this work, presents several inherent challenges due to only a single viewpoint being available to track, especially during occlusions when

only limited visual cues are available. Having more information/features helps – this was validated by our superior results including homography cues into our tracking network. Extending this further, incorporating ≥ 2 cameras (and viewpoints) will greatly augment the tracker’s capacity to handle challenging situations and make better predictions. This is especially relevant for team sports in general, and ice hockey in particular due to their fast-paced and dynamic player movements.

7.2.3 Action recognition

Player action recognition is an important task to identify what each player does at any given point during the game. Tracking provides the necessary temporal context and localization to differentiate different actions like skating, shooting, or passing. With tracking data, models can identify not only what action occurred but also *where* and *when* it took place. This can be important for strategic analysis, such as recognizing the exact moment a player initiates a specific move, or pinpointing where certain actions are more likely to happen. It also helps distinguish between offensive plays, defensive formations, and transitions, which are often indicative of specific actions, like a breakaway or a set play.

7.2.4 Trajectory imputation

Trajectory imputation aims to predict *missing* trajectory coordinates to extrapolate off-screen player behavior. It is extremely useful for *counterfactual reasoning* and use within downstream models that require fully observable data. For ice hockey, this translates to prediction of player trajectories outside the observable [FoV](#) and when they’re unlocalizable due to obstructions. This also enables unique identification throughout the given sequence and shift detection, with additional inputs. Furthermore, several key player statistics like positions, velocity, acceleration, energy, etc. can be easily gauged by imputing trajectories throughout the game.

7.3 Thesis Applicability

Tracking in general, is an ubiquitous computer vision task with major application areas including autonomous vehicles, traffic analysis, security and surveillance, environmental and wildlife monitoring, supply chain and logistics, augmented reality, virtual reality, smart cities and urban planning. In sports, it’s useful for injury prevention, performance analysis,

strategizing, and several downstream tasks including pose estimation, action recognition, game analysis, relationship modeling, etc. Although this thesis focuses on ice hockey, the developed tracking methodology can be extended to other team sports like football, volleyball, and basketball with minimal modifications, provided player location coordinates or homography data is available. The core tracking principles and algorithms remain applicable across these sports. The proposed graphical MPN is generally adaptable beyond sports tracking, with the flexibility to incorporate domain-specific features for different applications.

7.4 Thesis Impact

This thesis advances the field of MOT in ice hockey through a novel graphical method using standard broadcast footage, eliminating the need for specialized sensors or equipments. This helps reduce setup and maintenance costs by simplifying the entire tracking process, and saves several human hours currently spent on reviewing, annotating, and observing each minute of the game for insights and analysis. Our research presents the first *open-source* MOT dataset (HTD-2) and a comprehensive benchmark for ice hockey, establishing a foundation to foster further research and development in hockey player tracking. The proposed homography-based graph tracking approach enables MOT in monocular (constrained) view with high consistencies, producing identity-preserved player tracklets crucial for several downstream tasks.

References

- [1] David Acuna. Towards real-time detection and tracking of basketball players using deep neural networks. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, pages 4–9, 2017.
- [2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.
- [3] Adrià Arbués-Sangüesa, Coloma Ballester, and Gloria Haro. Single-camera basketball tracker through pose and semantic feature fusion. *arXiv preprint arXiv:1906.02042*, 2019.
- [4] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinićius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. arxiv 2018. *arXiv preprint arXiv:1806.01261*, 2018.
- [5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.
- [6] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 01 2008.
- [7] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *CoRR*, abs/1602.00763, 2016.
- [8] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

- [9] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6247–6257, 2020.
- [10] Matija Buric, Marina Ivasic-Kos, and Miran Pobar. Player tracking in sports videos. In *2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 334–340, 2019.
- [11] Yizheng Cai, Nando de Freitas, and James J Little. Robust visual tracking for multiple targets. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*, pages 107–118. Springer, 2006.
- [12] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023.
- [13] Hua-Tsung Chen, Chien-Li Chou, Tsung-Sheng Fu, Suh-Yin Lee, and Bao-Shuh P Lin. Recognizing tactic patterns in broadcast basketball video using player trajectory. *Journal of Visual Communication and Image Representation*, 23(6):932–947, 2012.
- [14] Ping-Yang Chen, Ming-Ching Chang, Jun-Wei Hsieh, and Yong-Sheng Chen. Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *IEEE Transactions on Image Processing*, 30:9099–9111, 2021.
- [15] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [16] Cheng-Che Cheng, Min-Xuan Qiu, Chen-Kuo Chiang, and Shang-Hong Lai. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10051–10060, 2023.
- [17] Nicolas Chenouard, Isabelle Bloch, and Jean-Christophe Olivo-Marin. Multiple hypothesis tracking for cluttered biological image sequences. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2736–3750, 2013.
- [18] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. *2023 IEEE/CVF*

Winter Conference on Applications of Computer Vision (WACV), pages 4859–4869, 2021.

- [19] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3502, 2022.
- [20] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9921–9931, 2023.
- [21] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 1, pages 886–893. Ieee, 2005.
- [22] Patrick Dendorfer, Hamid Rezaatofghi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. 2020.
- [23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [24] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- [25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
- [26] Zimeng Fang, Chao Liang, Xue Zhou, Shuyuan Zhu, and Xi Li. Associate everything detected: Facilitating tracking-by-detection to the unknown. *arXiv preprint arXiv:2409.09293*, 2024.
- [27] Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. In *Computer vision and pattern recognition*, volume 1804, pages 1–6. Springer Berlin/Heidelberg, Germany, 2018.

- [28] P. Figueroa, N. Leite, R.M.L. Barros, I. Cohen, and G. Medioni. Tracking soccer players using the graph representation. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 787–790 Vol.4, 2004.
- [29] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [30] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
- [31] Ross Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.
- [32] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.
- [33] Matias Gran-Henriksen, Hans Andreas Lindgaard, Gabriel Kiss, and Frank Lindseth. Deep hm-sort: Enhancing multi-object tracking in sports with deep features, harmonic mean, and expansion iou. *arXiv preprint arXiv:2406.12081*, 2024.
- [34] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 1735–1742. IEEE, 2006.
- [35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [36] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [37] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [38] Min-Chun Hu, Ming-Hsiu Chang, Ja-Ling Wu, and Lin Chi. Robust camera calibration and player tracking in broadcast basketball video. *IEEE Transactions on Multimedia*, 13(2):266–279, 2011.

- [39] Hsiang-Wei Huang, Cheng-Yen Yang, Jiacheng Sun, Pyong-Kun Kim, Kwang-Ju Kim, Kyoungoh Lee, Chung-I Huang, and Jenq-Neng Hwang. Iterative scale-up expansion iou and deep features association for multi-object tracking in sports. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 163–172, 2024.
- [40] S. Iwase and H. Saito. Parallel tracking of all soccer players by integrating detected positions in multiple view images. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 751–754 Vol.4, 2004.
- [41] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35, 1960.
- [42] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [43] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pages 2688–2697. PMLR, 2018.
- [44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.
- [45] H. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2, 05 2012.
- [46] Itay Lavie, Guy Gur-Ari, and Zohar Ringel. Towards understanding inductive bias in transformers: A view from infinity. *arXiv preprint arXiv:2402.05173*, 2024.
- [47] L Leal-Taixé. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.
- [48] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.
- [49] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

- [50] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.
- [51] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.
- [52] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.
- [53] 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.
- [54] Yang Liu, Luiz G Hafemann, Michael Jamieson, and Mehrsan Javan. Detecting and matching related objects with one proposal multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4520–4527, 2021.
- [55] Zelin Liu, Xinggang Wang, Cheng Wang, Wenyu Liu, and Xiang Bai. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- [56] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.
- [57] Wei-Lwun Lu, Jo-Anne Ting, James J Little, and Kevin P Murphy. Learning to track and identify players from broadcast sports videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1704–1716, 2013.
- [58] Wei-Lwun Lu, Jo-Anne Ting, James J Little, and Kevin P Murphy. Learning to track and identify players from broadcast sports videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1704–1716, 2013.
- [59] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Retinatrack: Online single stage joint detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2020.

- [60] Elena Luna, Juan C SanMiguel, José M Martínez, and Pablo Carballeira. Graph neural networks for cross-camera data association. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):589–601, 2022.
- [61] Yudong Luo, Oliver Schulte, and Pascal Poupart. Inverse reinforcement learning for team sports: Valuing actions and players. In *IJCAI*, pages 3356–3363, 2020.
- [62] Mehrtash Manafifard, Hamid Ebadi, and H Abrishami Moghaddam. A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159:19–46, 2017.
- [63] Thomas Mauthner, Christina Koch, Markus Tilp, and Horst Bischof. Visual tracking of athletes in beach volleyball using a single camera. In *Proceedings of 6th International Symposium on Computer Science in Sport, IACSS*, pages 83–91, 2007.
- [64] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022.
- [65] Anton Milan, Laura Leal-Taixé, Ian Reid, S Roth, and K Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.
- [66] Nima Najafzadeh, Mehran Fotouhi, and Shohreh Kasaei. Multiple soccer players tracking. In *2015 The international symposium on artificial intelligence and signal processing (AISP)*, pages 310–315. IEEE, 2015.
- [67] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking - linking identities using bayesian network inference. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 2187–2194, 2006.
- [68] Kenji Okuma, Ali Taleghani, De Freitas, J.J. Little, and David Lowe. A boosted particle filter: Multitarget detection and tracking. volume 3021, 05 2004.
- [69] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 145–161. Springer, 2020.

- [70] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR 2011*, pages 1201–1208. IEEE, 2011.
- [71] Harish Prakash, Yuhao Chen, Sirisha Rambhatla, David Clausi, and John Zelek. Vip-htd: A public benchmark for multi-player tracking in ice hockey. In *Computer Vision and Intelligent Systems*, 2024.
- [72] Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, and Khoa Luu. Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13784–13793, June 2021.
- [73] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2015.
- [74] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.
- [75] Donald Reid. An algorithm for tracking multiple targets. *IEEE transactions on Automatic Control*, 24(6):843–854, 1979.
- [76] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.
- [77] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3047–3055, 2015.
- [78] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on Computer Vision*, pages 17–35. Springer, 2016.
- [79] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

- [80] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017.
- [81] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017.
- [82] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE, June 2015.
- [83] Jia Cheng Shang, Yuhao Chen, Mohammad Javad Shafiee, and David A Clausi. Rink-agnostic hockey rink registration. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 73–81, 2023.
- [84] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.
- [85] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [86] Statista. Operating income of national hockey league teams 2010. <https://thenhlzone.com/nhl-viewership-continues-to-be-drastically-up/>, n.d. Accessed: 2025-01-10.
- [87] Statista. Operating income of national hockey league teams 2010. <https://www.statista.com/statistics/193744/operating-income-of-national-hockey-league-teams-in-2010/>, n.d. Accessed: 2025-01-10.
- [88] Statista. Operating income of national hockey league teams 2010. <https://shorturl.at/ThE5L/>, n.d. Accessed: 2025-01-10.
- [89] Statista. Operating income of national hockey league teams 2010. <https://uwaterloo.ca/news/media/going-top-shelf-ai-better-track-hockey-data/>, n.d. Accessed: 2025-01-10.
- [90] Statista. Operating income of national hockey league teams 2010. <https://www.nhl.com/news/>

[nhl-edge-launches-website-for-puck-and-player-tracking-data/](#), n.d.
 Accessed: 2025-01-10.

- [91] Statista. Operating income of national hockey league teams 2010. <https://sports.yahoo.com/nhl-player-tracking-tech-faces-cost-obstacles-nhlpa-suspicion-172447605.html/>, n.d. Accessed: 2025-01-10.
- [92] Statista. Operating income of national hockey league teams 2010. https://www.espn.com/nhl/story/_/id/38724195/nhl-brings-advanced-puck-tracking-stats-public/, n.d. Accessed: 2025-01-10.
- [93] Josephine Sullivan and Stefan Carlsson. Tracking and labelling of interacting multiple targets. In *European Conference on Computer Vision*, 2006.
- [94] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [95] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [96] Masaki Takahashi, Kensuke Ikeya, Masanori Kano, Hidehiko Ookubo, and Tomoyuki Mishina. Robust volleyball tracking system using multi-view cameras. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2740–2745. IEEE, 2016.
- [97] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, November 2023.
- [98] Rajkumar Theagarajan and Bir Bhanu. An automated system for generating tactical performance statistics for individual soccer players from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2):632–646, 2021.

- [99] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.
- [100] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pages 791–808. Springer, 2016.
- [101] Kanav Vats, Mehrnaz Fani, David A Clausi, and John S Zelek. Evaluating deep tracking models for player tracking in broadcast ice hockey video. *arXiv preprint arXiv:2205.10949*, 2022.
- [102] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John S Zelek. Player tracking and identification in ice hockey. *Expert systems with applications*, 213:119250, 2023.
- [103] Vermaak, Doucet, and Perez. Maintaining multimodality through mixture tracking. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1110–1116 vol.2, 2003.
- [104] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.
- [105] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.
- [106] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1470–1478, 2018.
- [107] Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint object detection and multi-object tracking with graph neural networks. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 13708–13715. IEEE, 2021.
- [108] Yu-Hsiang Wang, Jun-Wei Hsieh, Ping-Yang Chen, Ming-Ching Chang, Hung-Hin So, and Xin Li. Smiletrack: Similarity learning for occlusion-aware multiple object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5740–5748, 2024.

- [109] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, pages 107–122. Springer, 2020.
- [110] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *CoRR*, abs/1909.12605, 2019.
- [111] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. *CoRR*, abs/1703.07402, 2017.
- [112] Ming Xu, J. Orwell, and G. Jones. Tracking football players with multiple cameras. In *2004 International Conference on Image Processing, 2004. ICIP '04.*, volume 5, pages 2909–2912 Vol. 5, 2004.
- [113] Kefu Yi, Kai Luo, Xiaolei Luo, Jiangui Huang, Hao Wu, Rongdong Hu, and Wei Hao. Ucmtrack: Multi-object tracking with uniform camera motion compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6702–6710, 2024.
- [114] Kua Chen Yingnan Zhao, Zihui Li. A method for tracking hockey players by exploiting multiple detections and omni-scale appearance features. *Project Report*, 2020.
- [115] Xinguo Yu, Changshen Xu, Qi Tian, and Hon Wai Leong. A ball tracking framework for broadcast soccer video. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 2, pages II–273, 2003.
- [116] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- [117] Ruiheng Zhang, Lingxiang Wu, Yukun Yang, Wanneng Wu, Yueqiang Chen, and Min Xu. Multi-camera multi-player tracking with deep player identification in sports video. *Pattern Recogn.*, 102(C), June 2020.
- [118] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.
- [119] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022.

- [120] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11), 2021.
- [121] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.
- [122] Ge Zheng, Liu Songtao, Wang Feng, Li Zeming, Sun Jian, et al. YoloX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [123] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015.
- [124] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. 2019.
- [125] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019.
- [126] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. 2022.
- [127] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.