Automated Head Impact Detection in Youth Ice Hockey: A Two-Stage Deep Learning Approach

by

Rowan Hussein

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the

Master of Computer Science, Concentration in Artificial Intelligence(MCS)

in

School of Electrical Engineering and Computer Science

University of Ottawa
Ottawa, Canada

Declaration

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University of Ottawa regulations concerning plagiarism, including those regarding consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use

Rowan Hussein
2025

Acknowledgements

The completion of this thesis represents not merely an academic milestone, but the culmination of a transformative journey that began in Egypt and led me to the vibrant research community at the University of Ottawa. This work would not have been possible without the guidance, support, and collaboration of numerous individuals who have enriched both my academic pursuits and personal growth.

First and foremost, I extend my deepest gratitude to my supervisor, Professor Robert Laganière, whose expertise in computer vision and unwavering commitment to research excellence have been instrumental in shaping this work. His patient guidance through the complexities of video analysis and machine learning, combined with his dedication to addressing real-world safety challenges, exemplifies the ideal of socially conscious technological innovation. Professor Laganière's leadership of the VIVA Research Lab has created an environment where rigorous academic inquiry seamlessly integrates with practical applications, fostering research that truly matters to communities.

I extend my sincere gratitude to Professor Blaine Hoshizaki, Director of the Neurotrauma Impact Science Laboratory at the University of Ottawa, and his research team for their invaluable expertise in head impact biomechanics and annotation methodology. Professor Hoshizaki's internationally recognized work in head and brain injury research, spanning over 25 years with 190 refereed publications, provided the foundational knowledge that informed our annotation protocols and impact classification framework. The Neurotrauma lab's specialized expertise in hockey-specific head trauma was instrumental in developing the rigorous annotation standards that ensure the scientific validity of our dataset.

I am particularly grateful to Amir, whose contributions to the player-centric view methodology proved invaluable. His insights into detection and tracking algorithms, coupled with his collaborative spirit, significantly enhanced the technical sophistication of our approach. The many hours spent debugging tracking algorithms and optimizing performance parameters exemplified the collaborative nature of modern research.

To my colleagues at the VIVA Research Lab, I express sincere appreciation for creating an intellectually stimulating environment where ideas flourish through con-

structive dialogue. The lab's commitment to pushing the boundaries of computer vision while maintaining focus on societal impact has profoundly influenced my approach to research.

My journey from Egypt to Canada has been supported by family whose encouragement transcended geographical boundaries. Their belief in the value of education and their sacrifices to enable my academic pursuits provide constant motivation. This cross-continental experience has enriched my perspective, allowing me to approach technical challenges with a global mindset while appreciating the universal importance of youth safety in sports.

I acknowledge the youth hockey organizations and families who permitted the use of game footage for this research. Their trust in our work and commitment to player safety underscore the real-world significance of academic research. The coaches, parents, and young athletes who participate in community hockey represent the ultimate beneficiaries of this work, and their welfare has been the driving force throughout this project.

Finally, I thank the examination committee members for their time and expertise in evaluating this thesis. Their feedback and insights will undoubtedly strengthen both this work and my future research endeavors.

This thesis stands as a testament to the power of international collaboration, interdisciplinary research, and the application of advanced technology to fundamental human needs. As I continue my academic journey, the lessons learned and relationships forged during this work will remain invaluable assets.

Abstract

Detecting head impact events in youth ice hockey is a critical problem at the intersection of computer vision, sports safety, and public health. Despite increasing awareness of the long-term neurological risks of repetitive head trauma, most youth leagues lack systematic monitoring. This thesis addresses that gap by developing and evaluating an automated detection system tailored to the resource constraints of community hockey. The contributions are threefold. First, we introduce the first publicly available annotated dataset for this task, featuring a hierarchical labeling scheme for both general and head impact events. Second, we design a hierarchical two-stage pipeline for rare-event discovery: Stage 1 detects general impact events to prune routine play, and Stage 2 applies specialized classification to determine head involvement. Third, we provide a comprehensive empirical comparison between a player-centric approach (Model A) and a full-frame multi-modal fusion approach (Model B).

Experiments show that the optimal deployment strategy uses a player-centric architecture for Stage 1 and a full-frame architecture for Stage 2. The best Stage 1 model (**Model A: TSM + Motion**) achieves **75% recall** and **25% precision** (F1-score 0.375) for general impact detection, successfully reducing the portion of video requiring human review from 100% to just 6.3%. For the subsequent head impact classification task, the best Stage 2 model (**Model B: RGB + Flow + Pose**) achieves **80% recall** and **67.6% precision** (F1-score 0.733). This results in an end-to-end head impact event detection recall of **60%**. Technical analyses reveal several key findings: motion (optical flow) is the dominant signal for impact detection; Temporal Shift Modules (TSM) consistently outperform Inflated 3D Convnets (I3D) in both accuracy and computational efficiency; and pose estimation features provide limited value for this task due to low resolution and equipment occlusion. This work demonstrates the feasibility of automated safety monitoring in youth hockey and provides a foundational dataset and methodology for future research.

Table of contents

List of figures

List of tables

List of Acronyms

AI	Artificial Intelligence
CNN	Convolutional Neural Network
COCO	Common Objects in Context (dataset)
EMA	Exponential Moving Average
FP16	16-bit Floating Point (mixed precision training)
FPS	Frames Per Second
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
I3D	Inflated 3D Convolutional Networks
IIHF	International Ice Hockey Federation
IoU	Intersection over Union
IRB	Institutional Review Board
LSTM	Long Short-Term Memory
mAP	Mean Average Precision
MCC	Matthews Correlation Coefficient
MLP	Multi-Layer Perceptron
MOT	Multiple Object Tracking
MOTA	Multiple Object Tracking Accuracy
NHL	National Hockey League
NMS	Non-Maximum Suppression
RAFT	Recurrent All-Pairs Field Transforms
RGB	Red, Green, Blue (color channels)
ROI	Region of Interest
SVM	Support Vector Machine
TACDEC	TACkle DEtection Challenge (dataset)
TSM	Temporal Shift Module
TSN	Temporal Segment Networks
YOLO	You Only Look Once (object detection)

Chapter 1

Introduction

1.1 The Critical Gap in Youth Hockey Safety Monitoring

Ice hockey stands as one of the most physically demanding youth sports, combining high-speed skating, frequent body contact, and hard surfaces that amplify impact forces. Recent epidemiological studies reveal alarming statistics: youth hockey players aged 11-18 experience an average of 223 head impacts per player per season, with peak accelerations often exceeding those recorded in professional leagues [14]. These impacts occur during a critical neurodevelopmental period when the adolescent brain undergoes substantial structural and functional changes, making young athletes particularly vulnerable to both immediate and long-term consequences of repetitive head trauma [42]. Despite this heightened vulnerability, youth hockey leagues operate with minimal systematic safety monitoring, relying primarily on volunteer coaches who must simultaneously manage game strategy, player development, and injury detection. This excessive cognitive load inevitably compromises player safety.

The difference between professional and youth hockey safety infrastructure represents a profound inequity in sports medicine. Professional leagues deploy comprehensive monitoring systems as seen in Figure 1.1: the National Hockey League utilizes synchronized multi-camera arrays capturing every angle of play, instrumented equipment transmitting real-time biomechanical data, and dedicated spotters whose sole responsibility is injury detection [43]. Each professional game generates thousands of high-resolution video frames analyzed by both human experts and automated systems. In contrast, community youth leagues typically operate with a single camera, no specialized safety equipment beyond basic helmets, and coaches who receive minimal

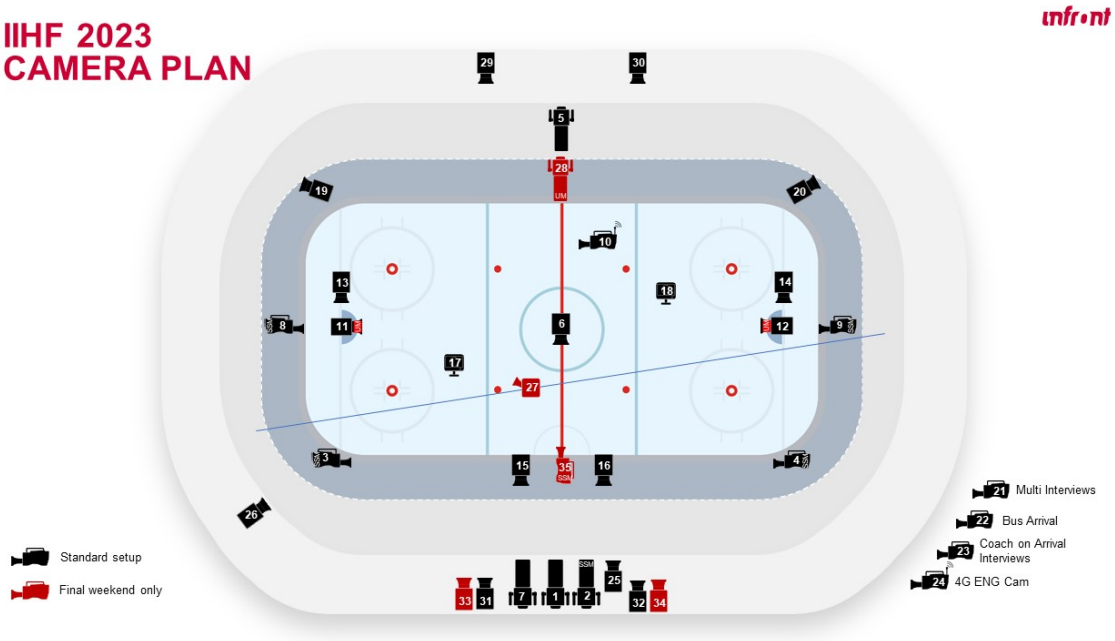


Fig. 1.1 Professional broadcast setup (e.g., IIHF World Championship 2023) deploys ~31 cameras per venue, rising to ~35 for final games, with specialty angles such as ref-cams and cable/robotic units, whereas youth games are often filmed by fixed or automated cameras and a much lower number[6]

training in concussion recognition. This resource gap means that the athletes most susceptible to long-term cognitive impairment from head injuries (those whose brains are still developing) receive the least systematic protection.

1.2 The Dataset Void: A Fundamental Barrier to Progress

The development of automated safety monitoring systems requires annotated training data that captures the specific characteristics of the target domain. While other contact sports have recognized this need and created comprehensive datasets (soccer researchers released TACDEC with 836 annotated tackle events [30], American football's NFL 1st & Future challenge provides 1,089 multi-angle helmet impact clips [44], and rugby's sensor-validated dataset contains 250 collision events [12]), ice hockey remains conspicuously absent from this research infrastructure. This absence is particularly glaring for youth hockey, where no publicly available dataset exists to support algorithm development or comparative evaluation. This data gap creates significant barriers: without training

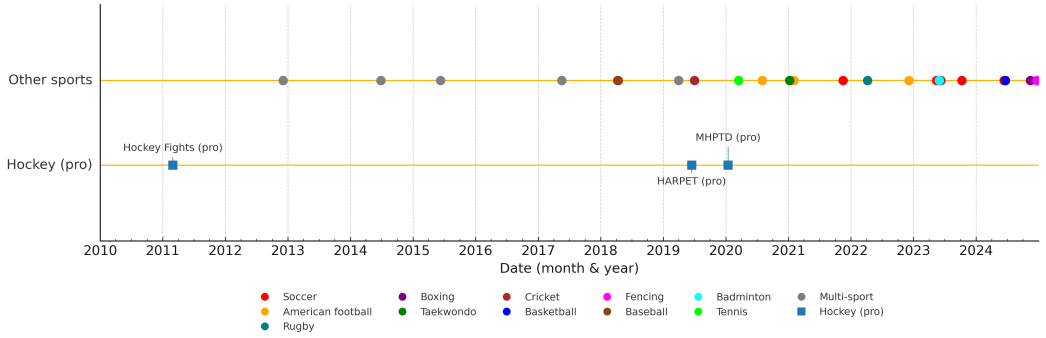


Fig. 1.2 Timeline showing the release of major sports impact detection datasets. While other sports have established benchmarks over the past decade, youth hockey has remained without any public dataset until our contribution in 2024. This gap has prevented the development of automated safety systems for the most vulnerable athlete population.

data, researchers cannot develop detection algorithms; without algorithms, safety systems cannot be validated; without validation, protective technologies cannot be deployed to safeguard young athletes.

Figure 1.2 illustrates how this dataset gap has persisted despite growing awareness of youth sports safety. The unique challenges of creating a youth hockey dataset compound this problem. Unlike professional sports with standardized venues and broadcasting equipment, youth hockey occurs in diverse community facilities with varying ice dimensions, lighting conditions, and camera placements. Obtaining ethical approval for research involving minors requires extensive protocols and parental consent processes that deter many research groups.

Furthermore, the relative rarity of head impacts (our analysis shows they constitute only 8.5% of all general impact events) means that researchers must process hundreds of hours of footage to obtain statistically meaningful samples. To address this gap, we contribute (to our knowledge) the first annotated dataset of youth hockey head impacts: a growing collection of complete games with expert-reviewed impact annotations, including verified head-contact events. By releasing this resource publicly, we enable the broader research community to develop, test, and compare automated safety-monitoring approaches.

1.3 Technical Challenges in Resource-Constrained Environments

The technical realities of youth hockey environments present tough challenges for computer vision systems, as demonstrated in Figure 1.3. Community ice rinks typically employ a single camera mounted at center ice, often a consumer-grade device operated by parent volunteers. This single-camera constraint reflects the practical realities of youth sports infrastructure: community rinks lack the financial resources for multi-camera installations (which require synchronized hardware, professional-grade networking infrastructure, and ongoing technical maintenance), and parent volunteers cannot manage the complexity of multi-viewpoint systems. While a multi-camera setup would theoretically enable 3D reconstruction of player positions, reduce occlusion, and provide complementary viewing angles, implementing such a system would require camera calibration across viewpoints, temporal synchronization of video streams, geometric rectification to account for lens distortion, and sophisticated fusion algorithms to merge information from multiple sources [25, 53]. The cost-benefit tradeoff heavily favors single-camera solutions in resource-constrained youth settings, where a functional monitoring system with modest performance improvements far outweighs an ideal but financially infeasible multi-camera deployment.

This lone viewpoint must capture action across a 200-by-85-foot surface, resulting in player images where crucial anatomical features become nearly indistinguishable. At typical mounting distances, a youth player's head occupies merely 20–30 pixels in diameter, pushing against the fundamental limits of modern object detection architectures that typically require minimum feature sizes of 32×32 pixels for reliable recognition [61]. The small pixel footprint combines with motion blur from rapid skating speeds (often exceeding 20 mph) to create detection scenarios that challenge even state-of-the-art algorithms designed for high-resolution professional broadcasts.

Environmental variability in community facilities further compounds these detection challenges. Unlike professional arenas with broadcast-quality LED lighting systems, youth rinks feature diverse and often suboptimal illumination: aging metal halide fixtures create harsh shadows and temporal flickering, plexiglass barriers generate unpredictable reflections that confuse optical flow algorithms, and varying ice surface conditions (from fresh zamboni passes to late-game deterioration) alter the visual backdrop against which players must be detected. Figure 1.4 showcases these environmental challenges across different facilities. The combination of low resolution, variable lighting, and single-viewpoint constraints creates multiple simultaneous challenges.

Fig. 1.3 **Example from youth hockey video illustrating typical technical challenges. At this camera distance, players appear extremely small with heads occupying fewer than 30 pixels, making precise impact detection difficult**, particularly under motion blur and partial occlusion near the boards.

Algorithms must simultaneously handle occlusions when players cluster near boards, maintain tracking through dramatic illumination changes as players move between differently lit zones, and distinguish genuine impacts from routine physical contact, all while operating on visual information orders of magnitude poorer than that available in professional settings.

1.4 Research Objectives and Methodological Innovation

1.4.1 Why Detecting Youth Hockey Head Impacts Matters

Head-impact detection in youth hockey matters for three reasons. First, concussive and subconcussive events are frequently missed or under-reported when no trained medical professional is present; studies show many suspected concussions in youth sports go unrecognized or unreported [41, 66].

Fig. 1.4 Example of environmental variability in youth hockey facilities. This venue features a modern LED-lit rink with good visibility, but conditions such as lighting type, glare from plexiglass, and camera placement vary substantially across facilities, requiring algorithms that generalize beyond controlled environments.

Second, prompt identification enables removal-from-play and timely clinical evaluation in line with international consensus guidance, reducing the risk of secondary injury and premature return to play [42].

Third, a reliable event log provides an objective record for follow-up care, communication with parents and coaches, and data-driven prevention strategies (e.g., rule modifications, training emphasis, equipment assessment). In budget-limited settings where multi-camera systems or dedicated spotters are not feasible, automated video analysis can continuously monitor games and flag clips for human review, improving safety without additional staff.

1.5 Research Questions and Scope

This thesis addresses the practical question of whether computer-vision techniques developed for high-resource professional sports can be adapted to effectively monitor player safety in resource-constrained youth hockey environments. The challenge spans from extracting signals in low-resolution, motion-blurred footage to handling extreme class imbalance (head impacts constitute roughly 0.54% of frames) without sacrificing performance. To overcome the "needle-in-a-haystack" nature of this problem, we

developed a hierarchical two-stage detection strategy that first identifies general impact events and then applies a more focused analysis to determine head involvement.

To guide this investigation, we address three specific research questions:

RQ1: Can player-centric preprocessing improve head impact detection accuracy compared to full-frame analysis in resource-constrained youth hockey environments?

RQ2: What is the optimal combination of visual modalities (RGB, optical flow, pose) for detecting rare impact events with extreme class imbalance?

RQ3: Can automated systems achieve detection performance that meaningfully improves upon current human-based monitoring capabilities in youth hockey environments?

1.5.1 Scope and Limitations

This research focuses exclusively on vision-based detection using single-camera setups typical of youth hockey venues. We do not address the following:

- Multi-camera fusion strategies
- Real-time streaming architectures
- Clinical validation of detected impacts
- Injury prevention interventions

1.6 Research Contributions and Scientific Innovation

This thesis makes three synergistic contributions to the field of automated sports safety monitoring:

First, we created the first publicly available annotated dataset for youth hockey head impacts, addressing a critical gap in sports safety research. Built from hundreds of hours of game footage across multiple age groups (U11–U18) and competitive levels, its hierarchical annotation scheme allows algorithms to distinguish not just *if* an impact occurred, but also its specific type (e.g., player-to-player vs. player-to-boards).

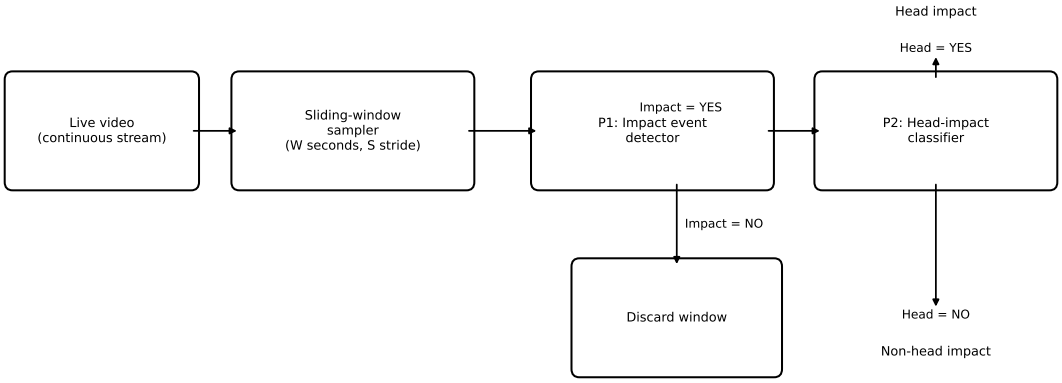


Fig. 1.5 Two-stage pipeline operated on live video. The stream is split into overlapping windows. P1 (impact event detector) labels each window as Impact vs. No-Impact. Only windows with Impact = YES are sent to P2, which classifies Head-Impact vs. Non-Head-Impact.

The dataset creation involved four professional video analysts, each receiving 1-2 months of specialized training in hockey biomechanics and impact identification. Annotating a single hour of video footage required 3-4 hours of careful expert review, reflecting the difficulty of identifying impacts from single-camera viewpoints where occlusion and limited depth perception challenge even trained observers. The annotation team achieved a Cohen's Kappa of 0.68, indicating moderate inter-rater agreement and highlighting the inherent subjectivity of impact classification from monocular video. Figure 1.6 shows representative frames illustrating the dataset's diversity.

Second, we dissect the trade-offs between two competing architectural paradigms. The first, a **player-centric** approach, isolates individual athletes to analyze their motion patterns, which provides clear outputs for diagnosing system failures. The second, a **full-frame** approach, processes the entire scene at once to learn from contextual cues, but at the cost of higher computational demands. By testing multiple architectures (TSM, I3D) and modalities (RGB, optical flow, pose), our analysis reveals a crucial trade-off: the higher precision of player-centric models versus the superior recall of full-frame models, indicating a hybrid strategy is likely optimal.

Third, we establish the real-world viability of our system as a practical monitoring tool. The two-stage pipeline architecture addresses the extreme class imbalance inherent in continuous video monitoring: Stage 1 filters out 93.7% of routine gameplay,

flagging only 6.3% of video clips as containing potential impact events for further analysis. Stage 2 then examines these flagged clips to determine head involvement. This cascade structure reduces the final review burden to a manageable subset of video requiring human verification.

To contextualize these results, we consider the baseline capabilities of human observers. Research on concussion detection consistently demonstrates that identifying head impacts through observation is fundamentally difficult. A recent systematic review of 29 studies encompassing 3,281 sport-related concussions found that expert video reviewers using visible signs achieved sensitivity below 50% across most studies, with specificity above 90% [64]. This finding indicates that even trained professionals conducting frame-by-frame post-game analysis miss more than half of concussive events when relying on observable indicators. A study of National Football League games further demonstrated that approximately 26% of medically diagnosed concussions exhibited no visible signs whatsoever, representing events that video review alone cannot detect [21].

These detection challenges are compounded by underreporting. Studies of high school football players found that 47% of athletes who experienced concussion symptoms did not report them, often because they did not recognize the injury as serious or did not want to leave the game [41, 67]. This creates a substantial gap between actual head impact occurrence and recognized events.

In youth hockey settings, detection conditions are considerably worse than those examined in professional sport research. Volunteer parent-coaches operating cameras must simultaneously manage recording equipment, track game action, and monitor player safety. Unlike the dedicated spotters employed in professional leagues who focus exclusively on injury detection with access to multiple camera angles and instant replay [40], youth hockey volunteers have divided attention and typically access only a single fixed camera view. Given that expert reviewers with optimal conditions achieve below 50% sensitivity, volunteer detection during live gameplay would be substantially lower.

Our expert annotation team of four trained video analysts, working with multiple review passes and master annotator oversight, achieved a Cohen's Kappa of 0.68 for head impact identification. This level of agreement, classified as "moderate" on standard interpretation scales, reflects the inherent difficulty of distinguishing head contact from body collisions using single-camera footage. The disagreement stems from ambiguous cases involving glancing impacts, partial occlusions, and simultaneous multi-point contacts where even careful review cannot establish definitive ground truth.

Fig. 1.6 Representative examples from our youth hockey head impact dataset showing the diversity of impact scenarios. Each line is separated between the two sides of the impact.

The Stage 2 classifier achieves 80% recall for head impacts, which exceeds the sensitivity levels reported for expert video review in professional sports settings. Combined with Stage 1 filtering that reduces the review burden from complete gameplay to 6.3% of clips, the system creates a practical workflow where automated pre-screening flags potential events for human verification. This approach acknowledges that both human and automated systems face limitations imposed by single-camera video quality, while providing systematic coverage that volunteer-based monitoring cannot achieve.

This provides a level of systematic, tireless monitoring that is impossible for volunteer coaches to replicate during live games. Our analysis shows the system's primary failure modes (crowded scenes and player occlusions) are the same scenarios where human observers struggle, highlighting the immediate potential for human-AI collaborative systems to flag events for closer review.

1.7 Significance and Transformative Potential

The immediate significance of this research lies in its potential to democratize safety monitoring across the youth sports ecosystem. Current professional monitoring systems cost upwards of $50,000 per venue and require dedicated technical staff, resources

that community leagues cannot mobilize [24, 1, 2, 48, 49]. By demonstrating that consumer-grade hardware (a single GPU workstation under $2,000) combined with existing camera infrastructure can support meaningful impact detection, we provide a technologically and economically feasible pathway for widespread adoption.

The broader transformative potential extends beyond immediate safety applications to catalyze fundamental changes in how youth sports organizations approach injury prevention. Systematic data collection enabled by automated monitoring can transform coaching practices from intuition-based to evidence-based approaches, identify high-risk game situations that warrant rule modifications, and provide objective metrics for return-to-play decisions following injuries. Long-term deployment could generate population-level insights into the cumulative effects of subconcussive impacts, informing age-appropriate contact policies and equipment standards. Furthermore, the technical innovations developed for this challenging domain (hierarchical detection strategies, extreme class imbalance handling, and multi-modal fusion under resource constraints) provide design patterns applicable to other safety-critical vision applications where rare but important events must be detected in continuous streams. As society increasingly recognizes the public health implications of youth sports injuries, automated monitoring systems will transition from novel research projects to essential infrastructure protecting the next generation of athletes.

1.8 Thesis Organization

The remainder of this thesis systematically presents the scientific foundations, technical contributions, and empirical validation of our automated impact detection system.

Chapter 2 provides a comprehensive survey of related work, tracing the evolution of sports safety monitoring from early sensor-based approaches to modern computer vision systems. We examine the specific challenges that ice hockey presents compared to other sports, review relevant advances in video action recognition and temporal modeling, and identify the gaps our work addresses.

Chapter 3 details our methodology, beginning with the careful process of dataset creation and annotation protocols, followed by the technical design of our two-stage detection pipeline. We provide sufficient implementation detail to ensure reproducibility while highlighting key design decisions and their rationales.

Chapter 4 presents our experimental results through systematic empirical evaluation. We begin with ablation studies that isolate the contribution of individual components,

proceed through comparative analysis of architectural choices, and conclude with end-to-end system performance on held-out test games.

Chapter 5 concludes the thesis by interpreting these results in the context of practical deployment. We analyze failure modes, discuss computational-accuracy tradeoffs, and outline pathways from research prototype to an operational system. This chapter also synthesizes our contributions, acknowledges the study's limitations, and charts future research directions.

Supplementary materials in the appendices provide implementation details and additional experimental information to support reproducibility.

Chapter 2

RELATED WORK

2.1 Sports-impact detection: problem setting & metrics

Collision risk across team sports. Group sports such as basketball and soccer are celebrated for their emphasis on teamwork and skill, yet epidemiological monitoring shows they carry substantial concussion risk whenever athletes collide with one another or the playing surface. U.S. high-school surveillance attributes almost two-thirds of sport-related concussions to athlete–athlete or athlete–surface contact, ranking basketball, football, soccer, and ice hockey among the ten most concussion-prone activities [10]. Professional trends echo this concern: the National Basketball Association logged 189 diagnosed concussions between 1999 and 2018, an average of 9.7 per season, with incidence nearly tripling after mandatory reporting rules were introduced [51]. Ice hockey presents an even harsher biomechanical milieu: low-friction skating elevates velocity, rink boards are unforgiving, and the ice itself amplifies fall severity. Collegiate telemetry confirms that direct player–player contact is the most frequent mechanism, yet player–ice impacts yield the highest head-acceleration magnitudes [68]. Pilot mouth-piece studies on Bantam-level skaters report frequent high-magnitude head impacts during routine play [15], underscoring the vulnerability of youth athletes who lack the anticipatory skills of professionals.

Public datasets that enable impact research in other sports. Safety researchers have responded by releasing curated video corpora that capture collision events with frame-accurate labels, allowing reproducible machine-learning studies. In soccer, *TACDEC* offers 836 expertly annotated broadcast clips of tackles [30], while *DeepImpact*

Fig. 2.1 Example frames for the four tackle labels defined in TACDEC. Reprinted from Kassab *et al.* [30].

pairs five full-match recordings with more than 3000 labelled headers to quantify sub-concussive exposure [56]. American football followed suit through the NFL *1st & Future* challenge, providing over a thousand multi-view helmet-collision segments and standard micro-F1 metrics for benchmarking [45]. Rugby-union researchers released a matched video-and-microsensor dataset of 250 rucks and tackles that now underpins algorithmic evaluations of collision load [11]. These open resources have accelerated impact-detection research within their respective sports contexts, inspiring studies that achieve broadcast-scale tackle recognition, season-long header-load estimation, and real-time helmet-collision alerts, all without proprietary footage.

Missing benchmarks and refined taxonomy for ice hockey. Ice hockey, despite its uniquely hazardous combination of high skating speed, rigid boards, and unforgiving ice, still lacks a public benchmark for routine collision detection. Existing vision datasets target other tasks: *HARPET* focuses on player trajectories and puck tracking [28], and the well-known *Hockey Fight* clips distinguish only fight versus non-fight sequences [47]. No open corpus labels head or body impacts, and published vision studies analyse at most 150 private NHL clips, precluding community comparison. Addressing this gap is a primary contribution of the work presented in this thesis, which introduces a new dataset with a refined taxonomy for three clinically salient collision modes: (i) *player–player* impacts, (ii) *player–glass* impacts, and (iii) *player–ice* impacts.

2.2 Sensor-based head/body-impact monitoring

Instrumented mouthguards and behind-ear IMUs dominate current field studies. Wearable kinematic devices have become a cornerstone of head-impact research because they attach directly to the athlete's skull or torso and can record thousands of events per season. Modern *instrumented mouthguards* encapsulate tri-axial accelerometers and gyroscopes inside a custom-moulded dental shell, locking the sensor to the maxilla and thus capturing true skull motion with minimal artefact [7]. Large NCAA-football deployments have logged more than 3500 video-verified impacts from just 21 athletes across two seasons, and World Rugby now integrates smart-mouthguard alerts into its Head-Injury-Assessment protocol [70]. For non-helmeted codes such as rugby league and Australian football, adhesive *X-Patch* IMUs placed behind the ear offer a portable alternative; video-verified match studies report that these devices capture over 90% of high-g collisions and tolerate sweat and lateral shear during play [8]. Bench tests show under 10% error in peak linear acceleration, and field data reveal that mouthguard-derived rotational kinematics correlate strongly ($\rho > 0.90$) with high-speed-video estimates, evidence that wearable sensing can serve as a first-line tool for quantifying collision exposure in contact sports.

Fig. 2.2 Commercial instrumented mouthguard (Jones *et al.*, 2023). The on-board tri-axial accelerometer, gyroscope, and Bluetooth radio are embedded in a rigid dental shell, yielding high-fidelity skull kinematics.

Video verification reveals systematic false positives and misses. Raw logs from wearables are vulnerable to *false positives* (chewing, stick taps, rapid head turns) and *missed events* when g-thresholds are set too high. Synchronised video therefore

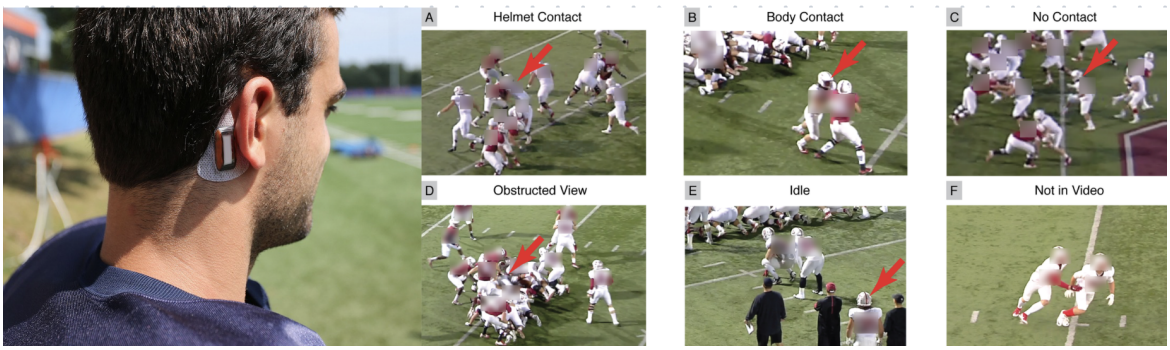


Fig. 2.3 Synchronised timeline of X-Patch triggers (top) and video-verified impacts (bottom) reproduced from Kuo *et al.* 2018. Chewing artefacts (red) inflate counts, whereas many low-magnitude direct hits (yellow) slip below the 20 g threshold.

remains essential for ground truth. In elite youth rugby league, only about 90% of X-Patch triggers above 20 g were confirmed on video, and most unverified events were non-contact motions [8]. A comparison of eight commercial devices in U.S. high-school football showed that helmet- and skin-mounted units over-counted impacts by factors of two to four, whereas mouthguards offered the best precision–recall balance yet still missed roughly 10% of low-amplitude hits [32]. Three recurring failure points explain these discrepancies: insecure mechanical coupling, poorly tuned trigger logic, and duplicate attribution when a single collision activates multiple nearby sensors. Hybrid pipelines that couple sensor detections with video confirmation now represent best practice for accurate epidemiology.

Youth ice-hockey amplifies every limitation of sensor-only monitoring. Community-level hockey introduces extra barriers: helmets shift on smaller heads, and custom dental work is discouraged until permanent teeth erupt. Among 18 under-15 skaters equipped with validated mouthguards, only 58% of the 1540 recorded triggers could be confirmed on multi-camera review. The remainder were dominated by skate vibration and stick taps [62]. A systematic review across youth sports further shows that when linear-g thresholds exceed 10–15 g, wearable systems miss up to 40% of true concussive hits, precisely the low-amplitude falls and slides common in youth hockey [50]. Cost ($300 USD per unit) and the logistics of dental impressions also limit uptake in grassroots programmes. These constraints motivate a shift toward vision-based collision detection, which leverages existing camera infrastructure without per-player instrumentation costs, provides verifiable ground truth through video review, and scales naturally to monitor entire teams simultaneously. Unlike wearable sensors that struggle with hockey-specific noise (skate vibrations, stick impacts), vision-based

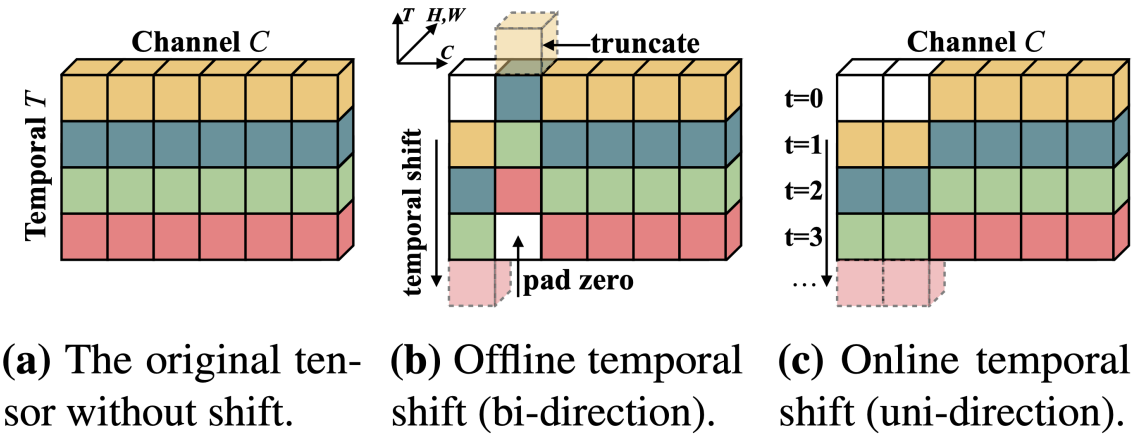
systems can learn to distinguish genuine collisions from routine gameplay through supervised training on annotated video [57, 17].

2.3 Vision-based collision detection frameworks

Frame-level motion cues with STIP and Bag–of–Words. Frame-level motion cues using dense optical flow together with Space–Time Interest Points (STIPs) and a Bag–of–Words (BoW) classifier established one of the first automated collision detection baselines in sports video. The method begins by computing dense optical flow (e.g., Farnebäck) to capture per-pixel motion between consecutive frames. Next, a Harris–in–time detector locates STIPs, points where the flow field exhibits a local maximum in its spatio–temporal gradient [33]. For each STIP, compact descriptors such as HOG, HOF, and MBH are extracted and vector-quantised via k-means into a visual vocabulary. A BoW histogram of codeword frequencies then feeds a linear SVM that labels fixed 1 s clips as *collision* or *non-collision*. These BoW baselines achieved over 85% precision on the TACDEC soccer tackle dataset [30] and 0.78 micro-F1 on the NFL 1st&Future helmet-impact challenge [45], demonstrating their generality across field sports. To date, however, no STIP/BoW pipeline has been evaluated on hockey collisions.

Player-centric clip classifiers and Temporal Shift Modules. Despite its conceptual simplicity, classical STIP/BoW suffers from two major limitations. First, it treats the entire frame uniformly, so camera pans, scoreboard inserts, or off-screen action dilute the collision signature. Second, its temporal support is rigid: BoW histograms integrate over a fixed window (commonly 32–64 frames), causing brief body checks to be underrepresented and extended scrums to generate false positives. Moreover, hand-tuned STIP thresholds often fail in youth-rink environments, where lower contrast and uneven lighting reduce reliable interest-point detection. These weaknesses motivated a shift toward player-centric approaches that first isolate each athlete before deciding whether their local motion constitutes a dangerous impact.

Building on the need to isolate individual players, detector-centric pipelines first localize each skater in every frame using an object detector (e.g., YOLOv8), then crop a fixed margin around those detections to form short tracklets for classification. Early designs inflated a 2D convolutional network into 3D or appended recurrent layers, improving accuracy but at the cost of high latency. The Temporal Shift Module (TSM) introduces temporal context by cyclically shifting a small fraction of feature channels



(a) The original tensor without shift. **(b)** Offline temporal shift (bi-direction). **(c)** Online temporal shift (uni-direction).

Fig. 2.4 Channel-shift operation in the Temporal Shift Module.

one frame forward and backward before each 2D convolution, embedding motion cues without extra parameters or multiplies [34]. Applied to per-player crops, TSM-based classifiers preserve the precision of full 3D models, run in real time, and focus on critical impact indicators (such as rapid deceleration and limb extension) while suppressing distracting background activity.

Multi-view and audio–visual fusion strategies. Multi-view and audio–visual fusion combines video from centre-ice and end-zone cameras with rink-side audio to mitigate occlusions and capture hidden checks. A hard hit unseen by the main camera may still register in the secondary angle or manifest as a sharp spike in the crowd microphone. On the NFL "1st & Future" dataset, graph-attention fusion of sideline and end-zone clips improved micro-F1 by 12% over single-view baselines [46]. In rugby, combining Mel-spectrogram peaks (whistle blasts, crowd reactions) with SIFT motion histograms in an SVM raised tackle recall from 0.71 to 0.84 [58]. In practical deployments, fusion remains efficient: high-energy audio frames gate which visual proposals undergo deep processing, and detections from each view are merged via spatio-temporal non-maximum suppression. This lightweight scheme preserves real-time throughput (essential for rink-side safety dashboards) while ensuring reliable collision detection even when one modality fails.

Windowing and adaptive clip generation. Achieving precise localisation also depends on how video is sliced into analysable units of time. The community default is a fixed sliding window (often 32 frames with 50% overlap) because this neatly fills mini-batches and lets a CNN "see" one second of play at a time. Fixed windows,

however, can break down in scenarios such as prolonged board scrums, which dilute the impact signal across multiple windows, or with lightning-fast hits that are too brief to register. Boundary-Sensitive Networks (BSN) replace this rigidity by predicting a start and an end probability for every video frame and then stitching high-confidence boundaries into variable-length proposals; on the ActivityNet benchmark, this approach lifts average recall by 12 points [36].

Table 3.1 summarizes context window lengths reported across multiple impact detection studies. Window sizes range from as few as 5 frames (Nonaka et al., rugby tackles) to 125 frames (Hicks et al., soccer multi-event detection), with most collision-focused systems converging on 9–50 frames. Short windows (5–15 frames) target rapid, discrete events where the impact itself lasts under half a second, while longer windows (50–125 frames) accommodate contextual lead-up and follow-through phases. For ice hockey, manual inspection of annotated impacts revealed typical event durations of approximately 1 second, suggesting an optimal window in the 25–35 frame range at 30 fps. This empirical finding aligns with established practice in contact-sport video analysis and informed the 30-frame window choice adopted in this thesis.

Channel Attention via Squeeze-and-Excitation. Squeeze-and-Excitation (SE) blocks recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels [29]. The mechanism first squeezes spatial dimensions through global average pooling, producing a channel descriptor that captures global spatial information. An excitation operation (two fully-connected layers with ReLU activation) then learns channel-specific scaling weights that emphasize informative features while suppressing irrelevant ones. For multi-modal fusion, SE blocks can adaptively weight contributions from RGB, flow, and pose streams based on their instantaneous reliability, automatically down-weighting corrupted modalities during occlusions or motion blur. This dynamic recalibration improves fusion robustness without manual weight tuning, reportedly contributing 3–5% accuracy gains in multi-stream architectures.

Training Strategies for Temporal Models

Optimization and Learning Rate Scheduling. AdamW decouples weight decay from gradient-based optimization, addressing the inconsistency in standard Adam where L2 regularization is improperly scaled by adaptive learning rates [38]. This correction is particularly important for video models where batch sizes are constrained by GPU memory, causing gradient noise that standard Adam can amplify. Cosine annealing schedules complement AdamW by gradually reducing learning rates following

a half-cosine curve, allowing aggressive initial exploration before smooth convergence [37]. For hockey collision detection, this combination can stabilize training despite severe class imbalance, as AdamW's corrected weight decay may prevent overfitting to majority classes while cosine annealing ensures stable refinement of decision boundaries.

2.4 Player detection, tracking & identity maintenance

YOLO object detection fundamentals. YOLO (You Only Look Once) introduced a paradigm shift in object detection by casting the task as a single, unified regression problem rather than a multi-stage pipeline. The image is partitioned into an $S \times S$ grid, and for each cell the network predicts bounding-box coordinates, objectness confidence, and class probabilities in one forward pass [54]. This design eliminated the need for separate region proposal and classification stages, enabling inference at video frame rates. Subsequent YOLO versions incorporated multi-scale feature maps to detect objects of varying sizes, introduced anchor boxes to provide geometric priors for different aspect ratios, and decoupled classification and localisation into separate head branches to reduce task interference. Throughout its evolution, the YOLO family has consistently emphasized a balance of simplicity, speed, and accuracy, making it a go-to detector for applications from autonomous driving to live sports analytics.

YOLOv8 refines this lineage with an anchor-free detection head that directly regresses box width and height, removing the need for predefined priors and simplifying training. It uses a CSPDarknet backbone to efficiently extract rich spatial features and a Path Aggregation Network (PANet) to merge information across scales, ensuring that both small skaters and large goal frames are detected reliably [13]. Classification and localisation are handled by separate head branches, each optimised with a focal loss for class imbalance and a Generalised IoU loss for precise boundary alignment. Data-augmentation strategies such as mosaic mixing and random perspective transformations further increase robustness to varied rink lighting and camera angles. Together, these architectural choices deliver a detector that is compact, fast, and well-suited for real-time collision surveillance in youth-hockey broadcasts.

Anchor boxes and non-maximum suppression. Anchor boxes are predefined rectangular priors that enable dense detectors to match objects across scales and aspect ratios without sliding-window enumeration. Each anchor is characterised by width, height, and centre offset relative to its grid cell; typical heads assign 3–9 anchors

per cell so that at least one anchor closely approximates each object [55]. During training, anchors whose Intersection-over-Union (IoU) with a ground-truth box exceeds 0.5 become positive samples, and the network learns small offsets to refine these anchors into tight detections. At inference, hundreds of overlapping proposals arise; non-maximum suppression (NMS) prunes this set by ranking boxes by confidence and discarding any whose IoU with a higher-scoring box exceeds 0.5. Although many modern detectors are anchor-free, an understanding of these concepts remains essential, as they are fundamental to the operation of many tracking and data association algorithms used in robust collision analysis.

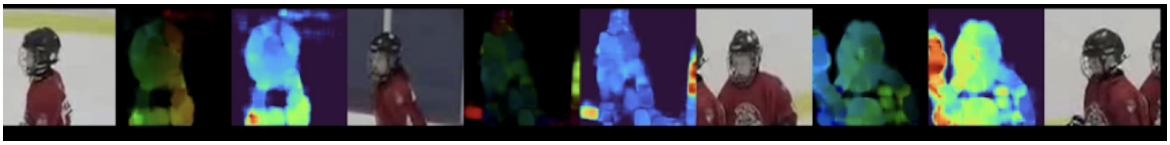


Fig. 2.5 Farnebäck optical-flow visualisation.

Optical flow for motion masking. Optical flow captures the apparent motion of each pixel between two consecutive frames, encoding both speed and direction as a dense field. In sports-video research, flow is routinely used as a pre-filter that isolates genuinely moving regions (players, puck, falling equipment) while suppressing static background such as boards and spectators. The Farnebäck algorithm is a popular choice because it estimates flow in a single pass by fitting a quadratic polynomial to each neighbourhood, making it fast enough for broadcast-rate footage [22]. Researchers threshold the resulting magnitude map to keep only pixels whose motion exceeds a scene-specific cutoff and then group contiguous pixels into blobs. Small blobs (e.g., <200 px) that correspond to noise, camera grain, or scoreboard animation are discarded, leaving clean motion masks that guide subsequent detectors to focus on high-activity zones and skip empty ice. Figure 2.5 shows a colour-coded Farnebäck flow overlay: hue represents direction, while saturation denotes velocity. Such masking has been reported to cut the number of candidate regions by more than half, reducing computational load without sacrificing recall for collision events [22].

RAFT: Recurrent All-Pairs Field Transforms. Recent advances in learning-based optical flow have produced RAFT (Recurrent All-Pairs Field Transforms), which employs iterative refinement through a recurrent unit to achieve state-of-the-art accuracy [63]. RAFT constructs a 4D correlation volume encoding all pairwise similarities between pixels, then iteratively updates flow predictions through a convolutional GRU

that mimics traditional optimization. Unlike classical methods, RAFT handles large displacements and maintains sharp motion boundaries, critical for capturing the rapid accelerations in hockey collisions. The method's 12-iteration refinement process trades computational cost for superior motion estimation, making it highly suitable for offline analysis where accuracy outweighs speed.

Lucas-Kanade Sparse Optical Flow. The Lucas-Kanade method remains a cornerstone of sparse optical flow estimation, tracking select feature points rather than computing dense correspondence [39]. The algorithm assumes constant flow within small neighborhoods and solves for displacement through least-squares optimization. Its pyramidal implementation enables tracking across scale spaces, capturing both fine details and large motions. Its computational efficiency has established the Lucas-Kanade method as a standard choice for real-time tracking systems where a balance between performance and speed is required.

Multi-object tracking under heavy occlusion. Object detectors produce frame-wise boxes; a tracker stitches those boxes into coherent trajectories so that each skater retains a unique identity across time, even when players overlap or leave and re-enter the field of view. A widely adopted baseline is *DeepSORT*, which augments the SORT Kalman-filter motion model with a 128-dimensional appearance embedding learned by a lightweight CNN [69]. At every frame, the tracker solves a bipartite assignment between predicted tracks and new detections, using a combined cost of Mahalanobis motion distance and cosine appearance distance. *StrongSORT* refines this pipeline by adding camera-motion compensation, re-identification via feature averaging, and a Generalised IoU (GIoU) term in the cost matrix; these changes raise identity-F1 (IDF1) by roughly six points on crowded MOT benchmarks [19]. Figure 2.6 shows a typical hockey sequence where coloured boxes illustrate how appearance cues rescue identity when two skaters cross paths and occlude each other. Because appearance varies little within a period, such embedding-based trackers remain the standard in vision-only sports analytics.

Trajectory smoothing and gap filling. Raw tracker outputs often contain short gaps (missed detections for a few frames) or jitter caused by noisy box regression. A common post-processing strategy is to first fill short gaps (typically under 15 frames) using linear interpolation of bounding box coordinates. Longer gaps can be matched to the nearest unassigned trajectory if their intersection-over-union (IoU) exceeds a set

Fig. 2.6 Appearance-assisted tracking preserves identity (colour-coded boxes) as two skaters collide and separate.

threshold, a strategy that helps recover tracks after extended occlusions. Once gaps are resolved, an exponential moving-average (EMA) filter is often applied to the coordinate sequences to remove high-frequency jitter while retaining genuine acceleration profiles. Finally, each smoothed box is enlarged to a fixed crop size to include context before being passed to a classifier. This multi-step refinement process is a widely adopted practice in sports-video pipelines because it reduces identity switches, stabilises motion features, and ensures uniform input size for downstream neural networks [16].

2.5 Temporal action & event-recognition networks

3D CNNs and two-stream architectures. Three-dimensional convolutional neural networks (3D CNNs) extend the familiar 2D filter into the temporal dimension, allowing each kernel to capture both appearance and motion within a short clip. Intuitively, a 3×3 mask detects edges in a still image, whereas a $3 \times 3 \times 3$ volume uncovers how those edges evolve across successive frames, revealing moving limbs, skate blades carving ice, or torsos rotating into impact. Tran et al. formalized this concept with their C3D network, which applies shared 3D kernels over fixed-length windows to learn generic spatiotemporal primitives [65]. Carreira and Zisserman later introduced “inflated” 3D networks (I3D), which adapt pre-trained 2D image models for video, dramatically improving sample efficiency on large benchmarks [9].

Despite their representational power, dense 3D convolutions incur cubic growth in computation as clip length increases, limiting their use on edge devices. The two-stream model proposed by Simonyan and Zisserman offers a computationally efficient alternative: one branch processes raw RGB frames to capture appearance, while a parallel branch ingests stacked optical-flow fields to encode motion [60]. By fusing only their high-level predictions, this architecture avoids early conflicts between features while still yielding a single event label. The two-stream design reuses mature image classifiers, permits sparse flow sampling to reduce FLOPs, and can even highlight anomalous frames when the streams disagree, providing an implicit confidence signal. This approach demonstrates that rich temporal context can be harvested with balanced compute, paving the way for lighter modules that retrofit temporal awareness into standard 2D backbones.

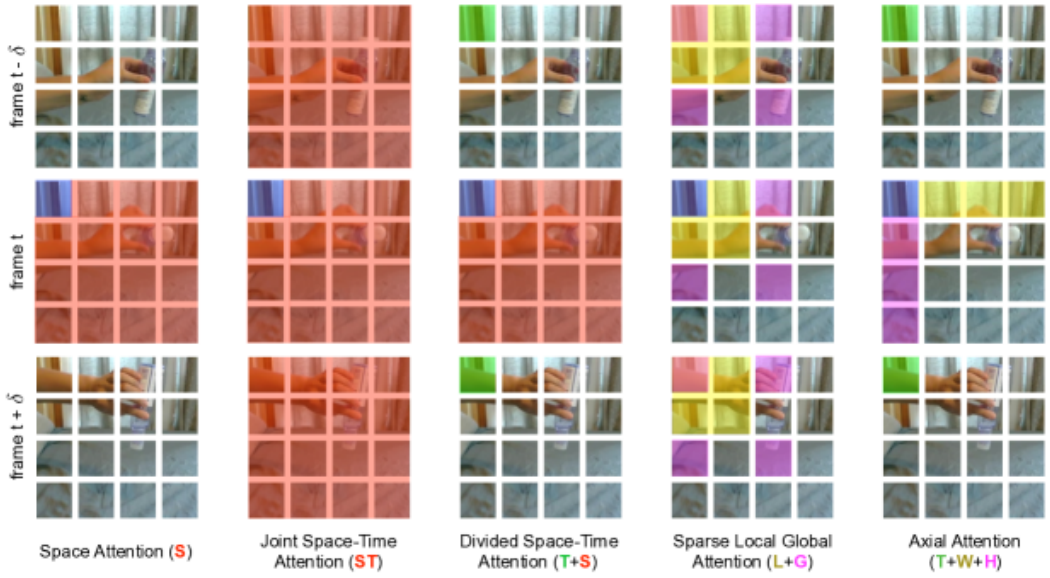


Fig. 2.7 Five attention schemes studied in TimeSformer (Bertasius *et al.*, 2021 [5]).

Efficient Temporal Modules (TSM, TIN). Early sequence models for video (recurrent neural networks (RNNs) and long short-term memory (LSTM) units) treat each frame as a timestep in a chain, allowing information to flow through hidden states over time [27]. Donahue et al. demonstrated that combining convolutional feature extractors with LSTMs (Long-term Recurrent Convolutional Networks) yields strong results on trimmed action clips, but the sequential nature of RNNs limits parallelism and incurs substantial latency when applied to high-frame-rate sports footage [18].

Moreover, vanishing gradients and the need to backpropagate through many timesteps can make training on long video segments unstable.

To address these challenges, recent work has devised lightweight modules that retrofit temporal reasoning into standard 2D backbones without recurrent loops. The Temporal Shift Module (TSM) cyclically shifts a small fraction of feature channels backward and forward along the time axis before each 2D convolution, effectively granting kernels access to adjacent frames while introducing zero additional parameters or floating-point operations [34]. Because shifting is a data permutation, BatchNorm and other pre-trained layers remain unchanged, enabling engineers to inject temporal sensitivity into any image classifier in a single code modification. Empirical studies on Kinetics and other action benchmarks report that TSM-augmented ResNet-50 matches or exceeds 3D CNN accuracy while running twice as fast on edge GPUs.

Building on TSM's premise, the Temporal Interlacing Network (TIN) learns a depth-wise gating mask that adaptively mixes features across time, interpolating between static shifts and full 3D convolutions [3]. By applying interlacing at low spatial resolution, TIN extends the effective temporal receptive field to entire plays without fragmenting videos into overlapping windows, all while keeping computational overhead under 10% compared to vanilla 2D networks.

The choice of temporal architecture involves fundamental trade-offs between accuracy, computational cost, and deployment feasibility. TSM achieves near-3D-CNN performance while maintaining 2D-network efficiency through its zero-parameter channel-shifting mechanism, making it particularly suitable for resource-constrained edge deployment where inference latency directly impacts system responsiveness. I3D and other full 3D convolution networks offer superior spatiotemporal modeling but incur approximately 5× computational overhead compared to TSM, rendering them impractical for continuous monitoring scenarios where hundreds of clips require classification per game. LSTM-based architectures introduce additional challenges: their sequential processing prevents efficient batch parallelization, and gradient flow through long temporal sequences often proves unstable during training. Transformer-based models such as TimeSformer provide global temporal context through self-attention but exhibit quadratic memory complexity with respect to sequence length, limiting practical clip durations and requiring substantial GPU memory even for modest batch sizes. For youth hockey monitoring, where computational resources are limited and processing throughput must match game pace, TSM represents an optimal balance, preserving temporal modeling capability while maintaining the efficiency necessary for practical deployment [34, 9, 5].

Transformer- and graph-based models. Attention mechanisms challenge the locality assumption built into convolution by allowing every spatiotemporal token to weigh the relevance of every other token, no matter their geometric separation. In practice, naively applying self-attention to full-resolution sports video is infeasible because computational complexity grows quadratically with the number of tokens. The TimeSformer resolves this by factorising the operation: attention is first computed along the temporal dimension at each spatial location and only then across spatial locations within each frame [5]. By exploiting this separability, the model can process minute-long clips while staying within desktop-GPU memory budgets. Such global context is particularly relevant to collision studies, where subtle precursors (skate edge changes or torso rotations) may occur several seconds before the actual impact.

Complementary to pixel-space transformers, graph neural networks (GNNs) operate on skeleton topologies extracted by pose detectors. In a Spatial-Temporal Graph Convolutional Network (ST-GCN), joints become nodes, physical connections become edges, and temporal links tie each joint to its past observations [72]. Graph convolutions then diffuse information both within a frame (capturing limb synchrony) and across frames (capturing trajectory smoothness). Because the node count remains constant at roughly one or two dozen joints, complexity scales linearly with clip length, making GNNs attractive for extended sequences where pixel transformers may still saturate memory. Hybrid formulations can tokenize both appearance patches and skeleton nodes, feeding them into a unified transformer that learns when to privilege raw texture (e.g., jersey numbers) and when to privilege kinematics (e.g., elbow extension during a cross-check) [5].

Transfer Learning from Kinetics. The Kinetics dataset family provides large-scale video understanding benchmarks with 400–700 human action classes and over 650,000 clips, establishing the de-facto pre-training corpus for temporal models [31]. ResNet-50, originally designed for ImageNet classification, serves as the standard backbone when adapted to video tasks through temporal extensions like TSM or I3D [26]. The network's residual connections prevent gradient degradation across its 50 layers, while skip connections preserve fine-grained spatial details crucial for detecting subtle impact indicators. Kinetics-pretrained models provide robust initialization for sports-specific fine-tuning, with learned features transferring particularly well to contact sports where human body dynamics dominate.

2.6 Pose-guided clip classifiers (Detectron-based keypoints, ETA features).

Pose guidance isolates the domain of interest (human articulation) before temporal reasoning begins, thereby suppressing background clutter such as boards, crowd motion, and scoreboard graphics. Detectron2's Keypoint R-CNN architecture forms the foundation of most contemporary pipelines: each region proposal branches into a heatmap head that localises canonical body joints, typically following the 17-point COCO schema [71]. Because the detector is trained on large-scale, diverse imagery, it transfers robustly to rink environments where lighting varies and uniforms occlude limbs. Once keypoints are estimated, their coordinates are normalised by torso length to ensure scale invariance and stacked into a tensor that records x, y, and confidence over time. This representation compresses a full-HD frame into a few hundred floating-point numbers, drastically reducing downstream compute while retaining the biomechanical essence of player movement. Importantly, the detector-classifier decoupling means that any improvement in keypoint accuracy can immediately benefit the temporal model without retraining, and vice-versa.

Temporal modelling of skeleton tensors can then proceed with specialised operators such as PoseC3D, which first rasterises joint trajectories into a pseudo-volume and processes it with depth-wise 3D convolutions before applying Efficient Temporal Aggregation (ETA) pooling [20]. The depth-wise design recognises that each joint often follows an independent motion pattern, so sharing filters across joints could dilute discriminative cues. ETA pooling further captures multi-scale dynamics by aggregating over several temporal dilations, allowing the network to recognise both rapid flicks and extended glides without separate passes. Practitioners favour pose-guided pipelines in safety-critical settings because erroneous appearance features rarely propagate into the skeletal domain. Moreover, skeleton trajectories are naturally anonymised, mitigating privacy concerns when youth athletes are involved. In summary, pose guidance provides a robust, compact, and ethically attractive alternative to pixel-heavy processing.

Fusion strategies across modalities. When a single modality proves ambiguous or noisy, combining complementary cues can resolve the uncertainty. A principled way to structure fusion is to decide *when* the streams should interact: at the raw-input stage (early fusion), within intermediate features (deep fusion), or only at the decision layer (late fusion). Early fusion concatenates RGB channels with pre-computed flow and heat-map representations before the first convolution; however, it demands careful

normalisation and risks overfitting when one modality is dominant. Deep fusion attaches attention or gating blocks between modality-specific backbones, letting each stream develop its own abstraction before selective information exchange. Late fusion, the historical starting point, simply averages or stages predictions from independent experts, favouring architectural simplicity and modular debugging [4].

Temporal Segment Networks (TSN) illustrate how sparse sampling interacts with fusion depth. Instead of scanning every frame, TSN draws evenly spaced snippets across the entire clip and forwards each snippet through RGB, flow, and pose branches whose logits are later aggregated [23]. In practical deployments, this strategy curtails compute during lulls in the action yet still captures long-range dependencies, making it attractive for full-match analytics. Designers can further refine the aggregation by weighting each stream's contribution according to per-snippet entropy, effectively granting more influence to the modality that appears most confident at that moment.

Class-imbalance remedies in prior work. Sports-video datasets are notoriously long-tailed: common actions such as skating or passing overwhelm infrequent head-impact clips, so naively optimising cross-entropy pushes the network toward majority classes. Early work in this area addressed the skew with straightforward *re-weighting*: Sozykin et al. multiplied the loss for rare actions by the inverse of their class frequency and achieved a 7% F_1 gain over an unweighted 3D CNN baseline [61]. Lin et al. generalised this idea into **Focal Loss**, which down-weights easy, well-classified samples and has since become a default loss function in sparse-event detectors [35]. More recent video benchmarks introduce resampling schemes; for instance, the VideoLT study balances head and tail classes by dynamically oversampling frames from under-represented actions during training [73]. Others combine both philosophies, such as using reconstruction and label mixing to regularise decision boundaries [52]. Collectively, these studies show that re-weighting, adaptive resampling, and label-mixing are complementary strategies for handling class imbalance.

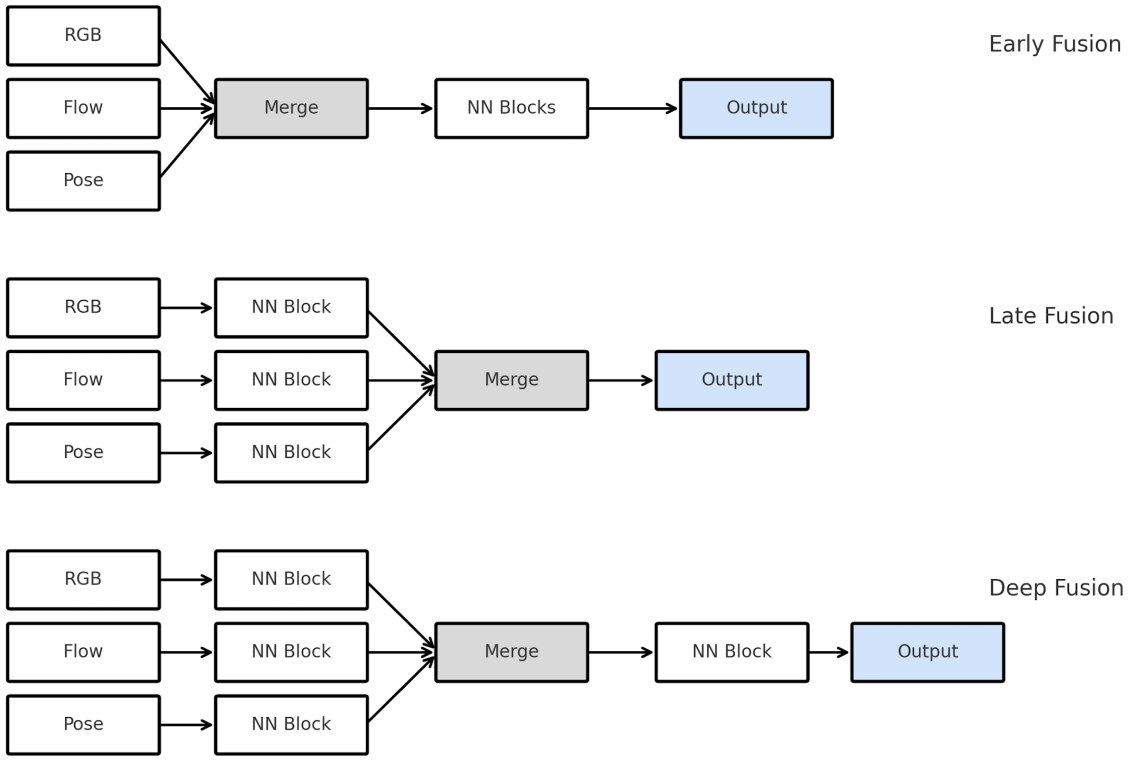


Fig. 2.8 Fusion design strategies for RGB, optical flow, and pose. Top: Early fusion merges input modalities before processing. Middle: Late fusion merges high-level features after independent backbone processing. Bottom: Deep fusion applies staged merging between network blocks.

Chapter 3

Methodology

Existing hockey datasets focus primarily on player movement classification tasks including passing, shooting, and skating maneuvers. The available datasets lack the temporal precision and general impact event annotations required for automated head injury detection systems. The absence of a comprehensive dataset targeting general impact events in youth hockey represents a critical methodological gap that prevents the development of effective detection systems for this population.

Youth hockey environments present distinct challenges compared to professional settings, including single-camera recording systems, variable lighting conditions, and diverse venue characteristics. These environmental differences necessitate domain-specific model development rather than direct adaptation of professional hockey analysis systems.

This chapter presents a comprehensive methodology for automated head impact detection in youth hockey through a two-stage detection pipeline. The first stage performs general impact event detection to identify any general impact event within 30-frame video clips. The second stage performs head impact event classification to determine whether detected general impact events involve head contact. We develop and evaluate two distinct architectural approaches: Model A employs a player-focused pipeline with motion feature extraction from tracked individual players, while Model B implements full-frame multi-modal fusion combining RGB appearance, dense optical flow, and pose keypoint information.

The methodology contains three primary components. Section 3.1 details the creation of a youth-specific hockey dataset with comprehensive annotation of both general impact events and head impact events across 45 games. Section 3.2 outlines the preprocessing pipeline that converts full-game recordings into standardized 30-frame clips suitable for machine learning analysis. Section 3.3 presents the machine learning

architectures, training procedures, and evaluation protocols used to assess detection system performance across both pipeline stages.

3.1 Dataset Creation and Annotation

3.1.1 Video Collection and Characteristics

The dataset comprises official game recordings obtained from youth hockey leagues across four age categories: Under-11 (U11), Under-13 (U13), Under-16 (U16), and Under-18 (U18). The collection includes both male and female leagues across three competitive levels (A, AA, AAA) to ensure complete coverage of youth hockey environments. All participating players represent local Canadian hockey organizations operating under Hockey Canada regulations.

Video recordings were captured at 30 frames per second with 1920×1080 pixel resolution using single fixed-camera configurations positioned at center ice. Youth hockey venues employ single-camera systems rather than the multi-camera arrays used in professional settings due to equipment and operational constraints. The single-camera limitation introduces detection challenges including partial player occlusion, variable player scaling based on distance from camera, and restricted viewing angles that constrain perspective on general impact events.

3.1.2 Impact Classification and Annotation Framework

The annotation framework identifies two distinct event categories to support the two-stage detection pipeline. General impact events encompass any physical contact involving a player, regardless of body region affected. Head impact events constitute a subset of general impact events where player head contact occurs during the general impact event. This hierarchical annotation structure directly supports Stage 1 general impact event detection followed by Stage 2 head impact event classification.

Professional hockey analysts defined four primary impact categories for general impact events:

1. **Player-to-player contact**: Direct contact between two or more players (e.g., body checking, inadvertent contact)

2. **Player-to-glass contact**: Player contact with rink boards or protective glass barriers

3. **Player-to-ice contact**: Player falling and striking the ice surface
4. **Player-to-object contact**: Contact with puck, stick, goal posts, or other equipment

For each general impact event, annotators assign a binary head impact flag indicating whether the player's head made contact during the general impact event, creating the ground truth labels for Stage 2 head impact event classification.

The annotation protocol requires identification of three temporal markers per game:

- Game start frame (excluding pre-game activities and advertisements)
- Game end frame (conclusion of official play time)
- All general impact event frames occurring within the gameplay window

Each annotated general impact event is recorded with four components:

- Exact frame number of the general impact event
- Bounding box coordinates for the involved player's body and head
- Impact category classification (one of four categories)
- Binary head impact flag: 1 if head contact occurred, 0 otherwise

The annotation process employed a team of four professional video analysts whose training was developed in consultation with the Neurotrauma Impact Science Laboratory at the University of Ottawa, directed by Professor Blaine Hoshizaki. The laboratory's 25+ years of research expertise in head impact biomechanics and sports-related brain trauma informed the annotation framework design, particularly in defining clinically relevant impact characteristics and distinguishing head contact from body collisions. Each analyst received 1–2 months of specialized training incorporating biomechanical principles established through the Neurotrauma lab's extensive work in hockey-specific head trauma assessment. This collaboration ensured that our annotation protocols captured impact events with the precision necessary for developing automated detection systems that align with medical and safety standards. The labor-intensive nature of this task is reflected in the time requirements: annotating a single hour of video footage required 3–4 hours of careful expert review. This significant time investment stems from the difficulty of identifying impacts from single-camera viewpoints where player occlusion and limited depth perception challenge even trained observers. The

annotation workflow followed a single-pass protocol, with a designated master annotator making final decisions on ambiguous cases to ensure labeling consistency. Inter-rater reliability analysis yielded a Cohen's Kappa of 0.68, indicating moderate agreement among annotators and highlighting the inherent subjectivity of impact classification from monocular video. This moderate agreement reflects the fundamental challenges of the task: determining head contact from a single viewing angle requires careful frame-by-frame analysis, and even expert human observers disagree on borderline cases where contact regions are partially occluded or viewing angles are suboptimal.

Complete dataset statistics are presented in Section 3.2.3, showing the distribution of general impact events and head impact events used for model training and evaluation.

3.2 Data Preprocessing

3.2.1 Temporal Window Selection

The selection of an appropriate temporal window for general impact event detection requires balancing sufficient context for accurate classification with computational efficiency. Table 3.1 presents context window lengths employed in related impact event detection studies across various sports domains.

Based on the analysis of context window lengths in Table 3.1, studies employing temporal approaches for action detection utilize windows ranging from 5 to 125 frames, with hockey-related studies specifically using 15 to 50 frames. We selected 30 frames to capture complete general impact event dynamics including pre-contact approach, general impact event occurrence, and post-contact recovery phases while maintaining computational efficiency for real-time processing requirements.

3.2.2 Clip Generation

The preprocessing pipeline converts full-game videos into fixed-length clips suitable for machine learning training. The sliding window extraction process follows the formal definition:

$$C_i = F_{i \times s : i \times s + w} \tag{3.1}$$

where C_i represents the i-th extracted clip, F denotes the full video frame sequence, $s = 5$ frames defines the stride length, and $w = 30$ frames specifies the window size.

Table 3.1 Context window lengths reported in impact event detection studies. N is the total number of frames the model (or the evaluation protocol) uses around each labelled general impact event, assuming the authors' native frame rates (30 fps unless otherwise noted).

Paper	**Sport / Goal**	N (frames)
Rezaei & Wu 2022 (DeepImpact)	Soccer – header detection	11
Sozykin *et al.* 2018	Hockey – multi-label actions	15
NFL 1st & Future 2021	American football – helmet impacts (±4-frame tolerance)	9
Gerats *et al.* 2021	Soccer – player & group actions	9 / 13
Nonaka *et al.* 2022	Rugby – high-risk tackles	5
Hicks *et al.* 2020	Soccer – real-time multi-event	125 (5 s)
Rongved *et al.* 2020 (MDPI)	Soccer – audio–visual events	32
Hassner *et al.* 2011 (Hockey-Fight)	Ice hockey – violence vs. play	50
Duan *et al.* 2022 (PoseConv3D [59])	Skeleton-only action recognition	48
Bertasius *et al.* 2021 (TimeSformer-L)	Generic video actions	96

The sliding window configuration produces an overlap of:

$$O = \frac{w - s}{w} \times 100\% = \frac{30 - 5}{30} \times 100\% = 83.3\% \tag{3.2}$$

The high overlap ensures comprehensive temporal coverage while maintaining computational tractability. Each 30-frame clip represents 1.0 second of gameplay at 30 fps, providing sufficient temporal context to capture general impact event dynamics.

The labeling strategy differs between the two pipeline stages. For Stage 1 general impact event detection, clips receive positive labels if any general impact event occurs within the entire 30-frame sequence. For Stage 2 head impact event classification, clips are labeled positive only if head contact occurs within the central temporal region (frames 10-20 of the 30-frame window), ensuring the general impact event appears

within the model's primary attention window while maintaining adequate contextual frames.

3.2.3 Dataset Statistics

Before presenting dataset statistics, we define the two distinct event categories used throughout this work. General impact events encompass any physical contact involving a player, regardless of body region affected, including all four impact categories defined in Section 3.1.2. Head impact events constitute the subset of general impact events where player head contact occurs during the general impact event.

Table 3.2 presents comprehensive statistics for the youth hockey dataset used in this study.

Table 3.2 Youth hockey dataset statistics

Dataset Characteristic	**Value**
Coverage	
Age groups	U11, U13, U16, U18
Competition levels	A, AA, AAA
Gender coverage	Boys and girls
Annotated General Impact Events	
Head impact events	8.5%
Impact Categories[a]	
Player-to-player contact	68.1%
Player-to-glass contact	17.0%
Player-to-ice contact	12.0%
Player-to-object contact[b]	3.0%
Processed Dataset (Post-Sliding Window)	
Non-event clips	93.7%
General impact clips	6.3%
Head impact clips	0.5%

[a] Percentages calculated relative to total general impact events

[b] Includes puck, stick, goal posts, and other equipment contact

Head impact events represent 8.5% of all general impact events, with player-to-player contact constituting the dominant mechanism (68.1%). After sliding window processing, general impact clips comprise 6.3% of the total dataset, while head impact clips represent 0.5%. The 6.3% versus 93.7% class imbalance necessitates specialized

training strategies including weighted sampling and modified loss functions to ensure effective model learning for both Stage 1 general impact event detection and Stage 2 head impact event classification tasks.

3.3 Machine Learning Methodology

3.3.1 Overall Experimental Framework: A Two-Stage Detection Pipeline

Detecting rare head impact events directly from continuous game footage is challenging. Player heads are small, fast-moving targets that are often occluded and captured from distant camera angles, making single-step detection unreliable. To address this, we developed a **two-stage detection pipeline** that decomposes the problem into two manageable steps: first identifying any potential impact, and then classifying only those impacts for head involvement. This hierarchical approach offers significant advantages over end-to-end detection. A single-stage model must simultaneously learn to identify rare head impacts (0.5% of clips) while filtering vast amounts of routine gameplay, forcing the network to balance two competing objectives. The two-stage design separates these concerns: Stage 1 focuses exclusively on broad impact detection with relaxed precision requirements, while Stage 2 specializes in the more nuanced task of head involvement classification. This architectural separation enables each stage to optimize for its specific objective without compromise. Furthermore, the cascade structure provides substantial computational savings Stage 2 processes only the 6.3% of clips flagged by Stage 1, reducing the computational burden of the more complex head-specific analysis by over 93%.

Within this framework, we systematically implemented and compared two distinct architectural philosophies. **Model A** represents a *player-focused* paradigm, which first isolates and tracks individual players before analyzing their motion. **Model B** represents a *full-frame* paradigm, which processes the entire scene at once using multi-modal fusion to learn from contextual cues. To ensure a comprehensive comparison, we evaluate multiple configurations within each architecture:

- **Model A Configurations**: TSM RGB only, TSM + Motion, and TSM + Tracking.
- **Model B Configurations**: RGB only, Optical Flow only, Pose only, and all fusion combinations (RGB + Flow, RGB + Pose, Flow + Pose, and RGB + Flow + Pose).

This experimental design allows us to isolate the contributions of different architectural choices and input modalities.

The two-stage approach functions as a coarse-to-fine cascade, inspired by similar methods in object detection. Stage 1 acts as a fast and efficient filter designed for high recall, ensuring that very few potential impacts are missed. This initial pass significantly reduces the amount of video that requires further analysis. Stage 2 then performs a more computationally intensive and precise classification, but only on the small subset of clips flagged by Stage 1. This division of labor makes the overall system both accurate and computationally feasible. The process is formalized by the following conditional probability:

$$P_{\text{final}}(\text{head impact}|\text{clip}) = P_{\text{stage2}}(\text{head impact}|\text{impact}) \cdot P_{\text{stage1}}(\text{impact}|\text{clip}) \tag{3.3}$$

Stage 1: General Impact Detection

The first stage is a binary classifier that processes every 30-frame clip to determine if it contains a general impact event. By filtering out routine gameplay, this stage reduces the data requiring intensive analysis to only 6.3% of the original footage. The model's prediction is given by:

$$\hat{y}_{\text{stage1}} = \begin{cases} 1 & \text{if } P_{\text{stage1}}(\text{impact}|\text{clip}) \geq \tau_1 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

where τ_1 is a threshold optimized on a validation set to balance precision and recall.

Stage 2: Head Impact Event Classification

The second stage performs a specialized classification exclusively on the clips that were flagged by Stage 1. This model is trained to distinguish the subtle visual and motion cues that differentiate general body contact from specific head contact events. The final prediction is:

$$\hat{y}_{\text{stage2}} = \begin{cases} 1 & \text{if } P_{\text{stage2}}(\text{head impact}|\text{impact}) \geq \tau_2 \text{ and } \hat{y}_{\text{stage1}} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

where τ_2 is the classification threshold for this stage. By focusing only on pre-qualified impact events, the Stage 2 classifier can learn more discriminative features than a single, end-to-end model.

Temporal Architecture Selection Rationale

The selection of an appropriate temporal modeling architecture involves fundamental trade-offs between representational capacity, computational efficiency, and deployment feasibility. We evaluated four major architectural families before selecting TSM as our primary temporal backbone.

Inflated 3D Convolutional Networks (I3D) extend 2D spatial convolutions into the temporal dimension through 3D kernels, enabling direct spatiotemporal feature learning. While I3D architectures offer strong modeling capacity for complex actions, they incur substantial computational overhead: our empirical evaluation demonstrated that I3D models required approximately 5× longer inference time compared to TSM-based approaches while achieving 20–22% lower F1-scores in Stage 2 head impact classification. The cubic growth in parameters and FLOPs with temporal kernel size renders I3D impractical for continuous monitoring scenarios where hundreds of clips require classification per game.

Recurrent architectures such as LSTM and GRU process video frames sequentially, accumulating temporal information through hidden states. However, this sequential dependency prevents efficient batch parallelization during both training and inference. Additionally, backpropagating gradients through long temporal sequences often proves unstable, requiring careful hyperparameter tuning. The computational overhead of recurrent processing combined with training difficulties led us to favor feedforward temporal modeling approaches.

Transformer-based architectures such as TimeSformer provide global temporal context through self-attention mechanisms, allowing each frame to attend to all other frames in the sequence. While theoretically appealing, transformers exhibit quadratic memory complexity with respect to sequence length, limiting practical clip durations and requiring substantial GPU memory even for modest batch sizes. The attention computation over 30-frame clips at 224×224 resolution creates memory bottlenecks that constrain deployment on resource-limited hardware typical of youth hockey venues.

The Temporal Shift Module (TSM) achieves near-3D-CNN temporal modeling performance while maintaining 2D network efficiency. TSM operates through a zero-parameter channel-shifting mechanism that exchanges information between adjacent frames by shifting feature channels along the temporal dimension before 2D convolution. This design preserves the efficiency of 2D CNNs while introducing temporal awareness, enabling TSM networks to process video at approximately twice the speed of equivalent 3D architectures. The compatibility with standard 2D network components allows direct utilization of ImageNet-pretrained weights, accelerating convergence during

fine-tuning. For youth hockey monitoring, where computational resources are limited and processing throughput must accommodate real-time or near-real-time analysis, TSM represents an optimal balance between temporal modeling capability and practical deployment constraints.

3.3.2 Model A: Player-Focused TSM Classifier

Model A employs a player-focused approach that isolates individual players before classifying their actions. This method hypothesizes that localizing and tracking individual players enables more effective learning of spatiotemporal impact patterns by reducing background motion interference. The methodology consists of four components: player detection, tracking, motion feature extraction, and temporal classification.

We evaluate four distinct Model A configurations across the two stages:

- **TSM RGB only**: A baseline using only visual appearance features from player crops.
- **TSM + Motion**: A multi-modal approach combining RGB features with optical flow vectors. For Stage 1, this uses RAFT. For Stage 2, we evaluate two variants: **TSM + Motion (Lucas-Kanade)** and **TSM + Motion (Farnebäck)**.
- **TSM + Tracking**: The baseline tracking configuration evaluated in Stage 1, which applies the TSM classifier to standard player crops from an IoU-assisted StrongSORT tracker.
- **TSM + Enhanced Tracking**: A Stage 2 variant that incorporates head-specific detection features into the tracking pipeline.

For Stage 1 general impact event detection, the **TSM + Motion** configuration uses Recurrent All-Pairs Field Transforms (RAFT) for dense optical flow. For Stage 2, the same configuration employs either Lucas-Kanade or Farnebäck algorithms as separate variants.

Player Detection and Tracking

Player detection utilizes a domain-adapted YOLOv8[13] model fine-tuned for youth hockey environments. The detection schema employs five classes: Player_Light, Player_Dark, Goalie_Light, Goalie_Dark, and Referee, leveraging consistent jersey color conventions to enhance tracking stability.

TSM + Tracking Configuration. The baseline **TSM + Tracking** configuration, evaluated in Stage 1, uses an IoU-assisted StrongSORT [19] tracker to generate player trajectories. To improve tracking stability where players on the same team have high appearance similarity, we modify the association cost matrix:

$$C_{\text{total}} = \alpha \cdot C_{\text{appearance}} + (1 - \alpha) \cdot C_{\text{IoU}} \tag{3.6}$$

where the parameter α is set to 0.6. This cost matrix is used to associate new YOLOv8 detections with existing player tracks. Additionally, a class-aware Non-Maximum Suppression (NMS) with an IoU threshold of 0.85 is used to prevent the erroneous removal of valid detections when players from different teams are in close proximity.

TSM + Enhanced Tracking Configuration. The **TSM + Enhanced Tracking** configuration, evaluated in Stage 2, builds upon the baseline tracker by incorporating additional head-specific features. For this configuration, a dedicated head detection model is applied to identify helmet regions within each player's bounding box. These localized head regions provide focused input features to the subsequent classification network, designed to help the model distinguish head contact from general body contact.

Motion Feature Extraction

For motion-based configurations, we extract motion descriptors using optical flow algorithms.

RAFT Optical Flow (Stage 1). For Stage 1 general impact detection, we employ RAFT for dense motion field estimation. RAFT computes displacement fields through iterative refinement with 12 iterations and processes frames padded to multiples of 64 pixels.

Lucas-Kanade Optical Flow (Stage 2). The Lucas-Kanade algorithm computes sparse motion fields via iterative optimization:

$$\mathbf{v} = \arg\min_{\mathbf{v}} \sum_{x,y} W(x,y)[I_1(x,y) - I_2(x + v_x, y + v_y)]^2 \tag{3.7}$$

where I_1 and I_2 are consecutive frames, $\mathbf{v} = (v_x, v_y)$ is the displacement vector, and $W(x,y)$ is a spatial weighting function. Corner features are detected using Shi-Tomasi

detection with a maximum of 60 features within the helmet region (upper 30% of the bounding box) and tracked across consecutive frames using a 3-level pyramidal implementation.

Farnebäck Optical Flow (Stage 2). The Farnebäck algorithm employs polynomial expansion for dense flow estimation:

$$f(\mathbf{x}) \approx \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \tag{3.8}$$

where coefficients $\mathbf{A}$, $\mathbf{b}$, and c are estimated via least-squares fitting. The displacement field is computed by analyzing coefficient changes between consecutive frames.

Motion Descriptor Aggregation. Motion features are computed within two regions of interest: the helmet area and the full player region. The aggregation follows:

$$\mathbf{f}_{\text{motion}} = [\text{median}(\Delta x), \text{median}(\Delta y), \text{median}(|\mathbf{v}|)] \tag{3.9}$$

where displacement vectors are aggregated into horizontal, vertical, and magnitude statistics.

Trajectory Post-Processing. Trajectory smoothing addresses tracking artifacts through temporal filtering:

$$\mathbf{p}_t^{\text{smooth}} = \alpha_{\text{smooth}} \cdot \mathbf{p}_t + (1 - \alpha_{\text{smooth}}) \cdot \mathbf{p}_{t-1}^{\text{smooth}} \tag{3.10}$$

where $\alpha_{\text{smooth}} = 0.3$ is the EMA coefficient. Short gaps up to 15 frames are interpolated using cubic splines. Player regions are standardized by cropping and resizing to 224×224 pixels, maintaining aspect ratio through zero-padding.

Network Architecture Modifications

Input Channel Adaptations. The different Model A configurations require distinct network inputs:

- **TSM RGB only**: Standard 3-channel RGB input from player tracklets.
- **TSM + Motion**: 6-channel input combining RGB (3 channels) with motion descriptors (3 channels) that are spatially replicated.

- **TSM + Tracking**: This configuration uses the same 3-channel RGB input as the baseline; its distinction lies in the experimental evaluation context for Stage 1.

Feature Integration and Fusion. For motion-augmented configurations, RGB and motion features are integrated via early fusion:

$$\mathbf{f}_{\text{combined}} = \text{Conv}_{1\times1}([\mathbf{f}_{\text{RGB}} \odot \mathbf{M}_{\text{spatial}}; \mathbf{f}_{\text{motion}}]) \tag{3.11}$$

where $\mathbf{M}_{\text{spatial}}$ is a mask for replicating the motion descriptors to match the RGB tensor's spatial dimensions, and a 1×1 convolution projects the concatenated features to the standard ResNet-50 input dimensions.

Training Methodology

Loss Function Formulations. For Stage 1, we employ a weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{stage1}} = -\frac{1}{N}\sum_{i=1}^{N}[w_{\text{pos}} \cdot y_i \log(\hat{y}_i) + w_{\text{neg}} \cdot (1 - y_i)\log(1 - \hat{y}_i)] \tag{3.12}$$

where $w_{\text{pos}} = 15.87$ and $w_{\text{neg}} = 1.07$ are inverse class frequency weights for the 6.3% positive class.

For Stage 2, we employ Focal Loss to address the severe class imbalance:

$$\mathcal{L}_{\text{stage2}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{3.13}$$

where $\alpha_t = 0.25$, $\gamma = 2.0$, and p_t is the predicted probability for the true class.

Training Configuration. Training employs an AdamW optimizer with a learning rate of 5×10^{-5}, weight decay of 1×10^{-5}, and a cosine annealing schedule. We use a batch size of 16 clips and prevent overfitting with early stopping after 10 epochs of no validation improvement. Feature normalization applies:

$$\mathbf{x}_{\text{norm}} = \frac{\mathbf{x} - \mu}{\sigma + \epsilon} \tag{3.14}$$

where μ and σ are channel-wise mean and standard deviation, and $\epsilon = 1 \times 10^{-5}$.

Temporal Classification with TSM

The classification network uses a Temporal Shift Module (TSM) architecture built upon a ResNet-50 backbone pre-trained on the Kinetics dataset. TSM enables spatiotemporal feature learning via channel shifting:

$$\text{TSM}(X_{t,c}) = \begin{cases} X_{t-1,c} & \text{for } c < C/4 \\ X_{t+1,c} & \text{for } C/4 \leq c < C/2 \\ X_{t,c} & \text{otherwise} \end{cases} \tag{3.15}$$

where $X_{t,c}$ is the feature tensor at time t and channel c. TSM modules are inserted after the first convolutional layer in each ResNet-50 residual block.

The network architecture processes **8 temporal segments** with a **shift division of 8**. For Stage 1, the network outputs a binary probability for general impact presence. For Stage 2, the same architecture processes pre-filtered clips to determine head impact occurrence.

Threshold Selection and Evaluation. Classification thresholds are optimized by maximizing the F1-score on the validation set:

$$\tau^* = \arg\max_{\tau} \text{F1}(\tau) = \arg\max_{\tau} \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)} \tag{3.16}$$

Final predictions follow $\hat{y} = 1$ if $P(y = 1|x) \geq \tau^*$, and $\hat{y} = 0$ otherwise.

3.3.3 Model B: Full-Frame Multi-Modal Fusion Architecture

Model B abandons player-centric preprocessing in favor of full-frame analysis. This approach hypothesizes that processing complete video frames captures contextual information such as the relative positions of multiple players that is lost when focusing solely on individuals. This strategy also eliminates any dependency on the performance of a separate player detection and tracking system, avoiding potential error propagation.

We evaluate seven distinct Model B configurations to analyze individual and combined modality contributions:

- **RGB only**: A baseline using only visual appearance features.
- **Optical Flow only**: Using only dense motion field features.
- **Pose only**: Using only human skeletal keypoint features.

- **RGB + Flow**: A two-stream fusion of appearance and motion.
- **RGB + Pose**: A two-stream fusion of appearance and kinematics.
- **Flow + Pose**: A two-stream fusion of motion and kinematics.
- **RGB + Flow + Pose**: A three-stream fusion of all modalities.

Additionally, we evaluate Inflated 3D ConvNet (I3D) architectures for comparative analysis with TSM approaches across the single-stream (RGB) and two most powerful fusion configurations (RGB + Flow, and RGB + Flow + Pose).

Input Modalities

RGB Stream. The RGB stream processes full 224×224 pixel frames. Pixel values are normalized to the range $[0, 1]$:

$$\mathbf{I}_{\text{norm}} = \frac{\mathbf{I}_{\text{raw}}}{255.0} \tag{3.17}$$

Dense Optical Flow Stream. Dense optical flow is computed using the RAFT algorithm. RAFT estimates displacement fields through iterative refinement with 12 iterations:

$$\mathbf{f}^{k+1} = \mathbf{f}^{k} + \Delta\mathbf{f}^{k} \tag{3.18}$$

Frame preprocessing requires padding to multiples of 64 pixels. The final flow fields result in a tensor of shape $(29, 224, 224, 2)$ for each 30-frame clip.

Pose Keypoint Stream. Pose extraction employs a Keypoint R-CNN to detect 17 COCO-style body joints. In youth hockey video, multiple players are typically visible in each frame, resulting in numerous detected skeletons. To create a single representative pose descriptor for the entire frame, we rank all detected skeletons using a composite score that combines detection confidence with spatial centrality:

$$S_i = \alpha_{\text{pose}} \cdot \frac{1}{17} \sum_{j=1}^{17} c_{i,j} + (1 - \alpha_{\text{pose}}) \cdot \exp(-d_i^2/\sigma^2) \tag{3.19}$$

where $c_{i,j}$ denotes the confidence score for keypoint j of skeleton i, d_i represents the distance from the frame center, and $\alpha_{\text{pose}} = 0.7$ weights confidence over centrality. We select the top 10 highest-scoring skeletons per frame (or fewer if less than 10 are

detected) to capture the most prominent players while filtering unreliable detections. These top-10 skeletons are then aggregated via confidence-weighted averaging:

$$\mathbf{p}_{\text{agg}} = \frac{\sum_{i=1}^{K} w_i \mathbf{p}_i}{\sum_{i=1}^{K} w_i}, \quad w_i = \frac{1}{17} \sum_{j=1}^{17} c_{i,j} \tag{3.20}$$

where $K \leq 10$ is the number of selected skeletons and w_i is the average keypoint confidence for skeleton i. This weighted aggregation produces a single representative pose tensor of shape $(30, 17, 2)$ for each 30-frame clip, emphasizing the most reliably detected players while suppressing noise from low-confidence detections.

Network Architecture Configurations

TSM-Based Architectures. Each modality stream employs a ResNet-50 backbone augmented with Temporal Shift Modules (TSM). The TSM network architecture processes **8 temporal segments** with a **shift division of 8**. The RGB and optical flow streams use identical architectures, except the first convolutional layer of the flow stream is modified for a 2-channel input. Following the backbone, bidirectional GRU layers aggregate temporal information, producing 1024-dimensional embeddings for the RGB and flow streams.

Pose Processing Pipeline. The pose stream uses a Transformer encoder (8 attention heads, 2 layers) to model spatial relationships between the 17 keypoints. Its multi-head self-attention mechanism computes:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3.21}$$

The output is then fed into a bidirectional GRU for temporal aggregation, producing a final 512-dimensional pose embedding.

Multi-Modal Fusion Mechanism. To fuse features, the 1024-dimensional RGB and flow embeddings are projected to 512 dimensions:

$$\mathbf{e}_{\text{RGB}}^{\text{proj}} = \mathbf{W}_{\text{RGB}}\mathbf{e}_{\text{RGB}}, \quad \mathbf{e}_{\text{flow}}^{\text{proj}} = \mathbf{W}_{\text{flow}}\mathbf{e}_{\text{flow}} \tag{3.22}$$

The embeddings are then concatenated and passed through a Squeeze-and-Excitation block which computes attention weights $\mathbf{w}$ to adaptively weigh each modality:

$$\mathbf{w} = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \text{GAP}(\text{Concat}[\mathbf{e}_{\text{RGB}}^{\text{proj}}, \mathbf{e}_{\text{flow}}^{\text{proj}}, \mathbf{e}_{\text{pose}}])))) \tag{3.23}$$

I3D Architecture Comparison. For comparison, we also implement I3D networks, which use 3D convolutions. Our I3D models inflate a pre-trained 2D ResNet-50 by expanding its kernels, initializing the 3D weights from the 2D weights to preserve learned features:

$$\mathbf{W}_{\text{3D}}(i, j, k) = \frac{1}{T}\mathbf{W}_{\text{2D}}(i, j) \tag{3.24}$$

where $T = 3$ is the temporal kernel size.

3.3.4 Training and Evaluation Protocol

Both Model A and Model B employ a consistent training and evaluation protocol. All models are implemented in PyTorch with automatic mixed-precision (FP16) training and a batch size of 16 clips per GPU.

Training Methodology

Loss Function Specifications. For Stage 1 general impact event detection, we employ a weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[w_{\text{pos}} \cdot y_i \log(\hat{y}_i) + w_{\text{neg}} \cdot (1 - y_i) \log(1 - \hat{y}_i)] \tag{3.25}$$

where $w_{\text{pos}} = \frac{1}{0.063} \approx 15.87$ and $w_{\text{neg}} = \frac{1}{0.937} \approx 1.07$ represent inverse class frequency weights addressing the 6.3% positive class imbalance.

For Stage 2 head impact event classification, we employ Focal Loss for severe imbalance handling:

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{3.26}$$

where $\alpha_t = 0.25$, $\gamma = 2.0$, and p_t represents the predicted probability for the true class.

Optimization Configuration. Training employs an AdamW optimizer with a cosine annealing learning rate schedule. The learning rate η_t at epoch t is defined as:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{t\pi}{T}\right)\right) \tag{3.27}$$

where $\eta_{\max} = 5 \times 10^{-5}$, $\eta_{\min} = 1 \times 10^{-7}$, and T is the total number of epochs. We apply a weight decay regularization of $\lambda = 1 \times 10^{-5}$ and use gradient clipping with a maximum L2 norm of 1.0. Early stopping terminates training after 10 epochs without validation F1-score improvement.

Data Sampling Strategy. To further address class imbalance, balanced mini-batch sampling is employed using weighted random sampling, where the probability of selecting a sample x_i is given by:

$$P(x_i) = \frac{w_{y_i}}{\sum_{j=1}^{N} w_{y_j}} \tag{3.28}$$

where w_{y_i} is the inverse class frequency weight for the label y_i of sample x_i.

Evaluation Protocol

Dataset Partitioning. Game-based partitioning prevents data leakage by ensuring test games remain completely unseen during training and validation. The training and validation sets contain clips exclusively from a designated set of training games, while the test set comprises clips only from held-out games.

Threshold Optimization. Classification thresholds are optimized by maximizing the F1-score on the validation set. The optimal threshold τ^* is found by:

$$\tau^* = \arg\max_{\tau \in [0,1]} \text{F1}(\tau) = \arg\max_{\tau \in [0,1]} \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)} \tag{3.29}$$

Performance Metrics. Model performance is evaluated using metrics appropriate for imbalanced classification tasks. The primary metrics are precision, recall, and the F1-score.

Chapter 4

Results

This chapter presents the experimental evaluation of the two-stage head impact detection pipeline described in Chapter 3. All models are assessed on a held-out test set comprising complete games not used during training or validation, ensuring unbiased performance measurement. The evaluation follows the pipeline structure, first examining **Stage 1: General Impact Detection** where models classify 30-frame clips as containing any general impact event or routine gameplay, then **Stage 2: Head Impact Classification** where only Stage 1 positive clips are analyzed to determine head contact presence.

Performance metrics include precision, recall, and F1-score, with classification thresholds optimized for each configuration to maximize validation F1-score. Due to severe class imbalance (general impact events represent 6.3% of clips in Stage 1), F1-score serves as the primary evaluation metric. We compare Model A (player-focused TSM architectures) and Model B (full-frame multi-modal fusion) across all configurations, with configuration names matching those defined in Chapter 3. The chapter also includes computational performance analysis reporting per-clip inference times on RTX 3090 hardware to examine the relationship between detection accuracy and processing speed.

4.1 Stage 1: General Impact Detection

Motion information emerges as the critical factor for distinguishing collision events from routine gameplay, with player-focused motion extraction achieving superior performance over all tested configurations. The systematic evaluation of both architectural approaches demonstrates that visual appearance features alone prove fundamentally insufficient for impact detection, while the integration of motion through different

methodologies reveals significant performance variations that establish clear architectural preferences.

4.1.1 Evaluation Protocol

All models are evaluated using F1-score and recall as primary performance metrics, with classification thresholds optimized to maximize F1-score for each configuration. The evaluation uses clips from held-out games not seen during training, with impact events representing 6.3% of all clips, making F1-score the appropriate balanced measure for this imbalanced classification task.

Table 4.1 Stage 1 General Impact Detection Performance Summary

Model	Configuration	Recall	Precision	F1-Score	Threshold	Architecture
Model A	TSM + Motion	**0.75**	0.25	**0.375**	0.52	Player-Focused
	TSM RGB only	0.52	0.16	0.245	0.72	Player-Focused
	TSM + Tracking	0.49	0.145	0.224	0.81	Player-Focused
Model B	RGB + Flow	0.74	0.245	0.368	0.55	Full-Frame
	Optical Flow only	0.62	0.21	0.313	0.61	Full-Frame
	RGB + Flow + Pose	0.60	0.20	0.30	0.63	Full-Frame
	Flow + Pose	0.61	0.205	0.307	0.63	Full-Frame
	RGB + Pose	0.59	0.195	0.293	0.66	Full-Frame
	RGB only	0.52	0.16	0.245	0.73	Full-Frame
	Pose only	0.50	0.15	0.231	0.78	Full-Frame

4.1.2 Model A: Player-Focused Pipeline Analysis

Motion features extracted from tracked players provide the most effective approach to impact detection among all tested configurations. The TSM + Motion configuration achieves the highest F1-score of 0.375 and recall of 0.75, as shown in Table 4.1. This performance indicates that focused motion analysis captures the essential dynamics of collision events more effectively than alternative feature extraction approaches.

Visual appearance alone proves fundamentally inadequate for impact detection in the player-focused pipeline. The RGB-only baseline achieves an F1-score of only 0.245 with recall of 0.52, as illustrated in Figure 4.1. This limitation persists despite player-centric preprocessing that isolates individual athletes, indicating that static visual features lack the temporal information necessary to distinguish collision events from routine movements.

The integration of motion features transforms detection performance through capture of collision-specific dynamics. The motion-enhanced configuration increases F1-

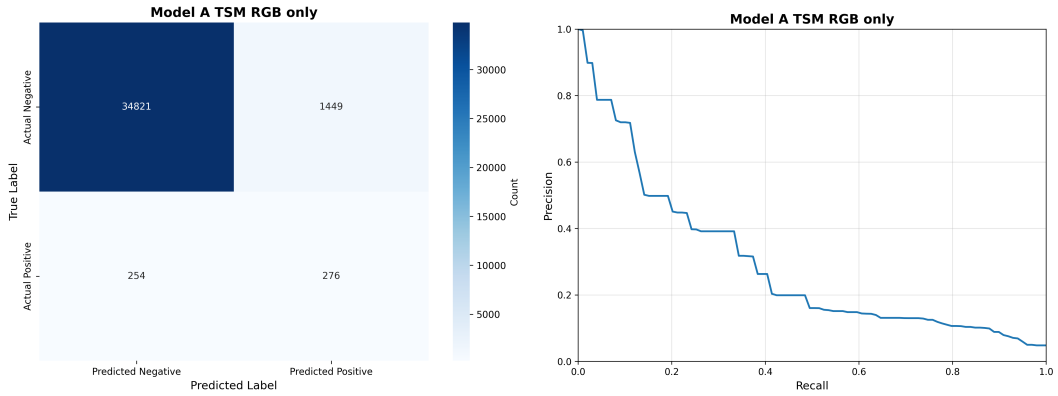


Fig. 4.1 Model A RGB baseline performance showing limited discriminative capability with static visual features.

score by 53% over the RGB baseline, achieving recall of 0.75 that successfully detects three-quarters of impact events. This substantial improvement demonstrates that motion features capture the sudden velocity changes and directional shifts characteristic of collision events, which remain invisible to single-frame analysis approaches.

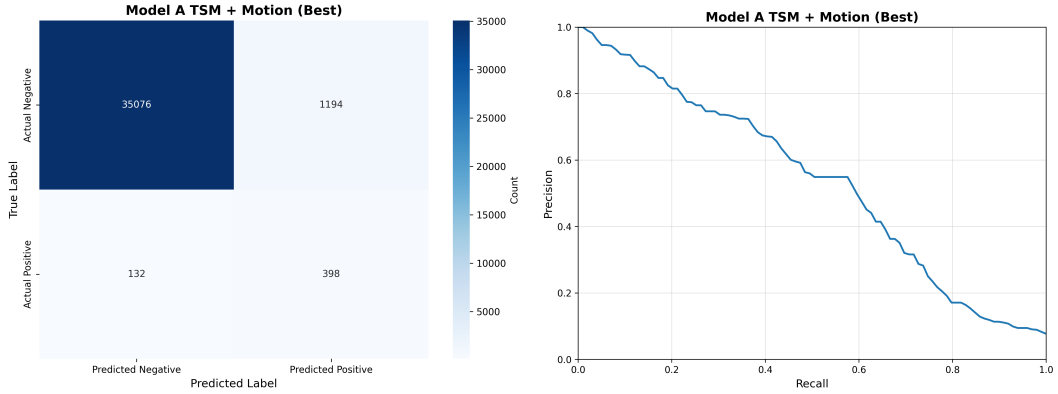


Fig. 4.2 Model A TSM + Motion performance demonstrating superior detection capability with motion features.

Tracking information without motion extraction degrades performance below the RGB baseline. The TSM + Tracking configuration achieves F1-score of only 0.224 and recall of 0.49, as shown in Figure 4.3. This counterintuitive result indicates that tracking consistency alone provides no discriminative value for impact detection, suggesting that tracking artifacts may interfere with visual feature learning while failing to provide the essential temporal dynamics.

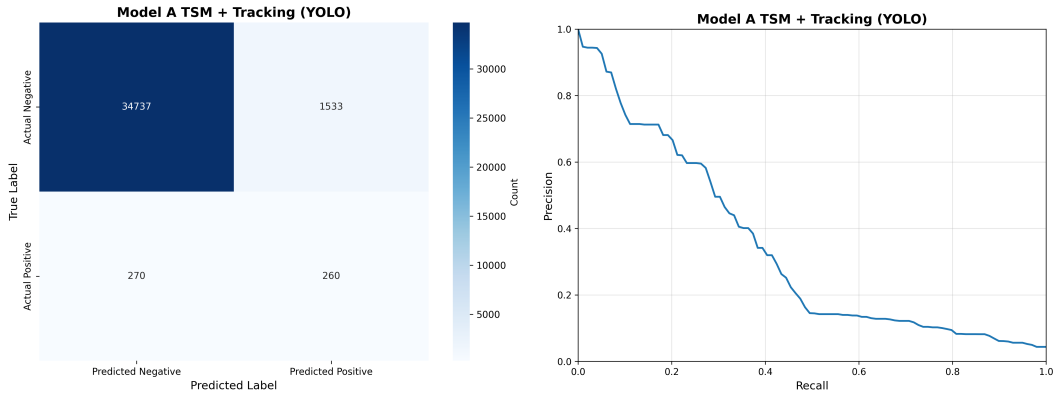


Fig. 4.3 Model A TSM + Tracking performance showing degradation compared to RGB baseline.

4.1.3 Model B: Full-Frame Multi-Modal Evaluation

Optical flow emerges as the most informative single modality for impact detection in full-frame analysis. Flow-only configuration achieves F1-score of 0.313, substantially outperforming both RGB (F1=0.245) and pose keypoints (F1=0.231), as detailed in Table 4.1. This hierarchy establishes that motion information provides superior discriminative power compared to appearance or structured kinematic features when analyzing the complete scene.

RGB features demonstrate identical limitations across both architectural approaches. The full-frame RGB-only configuration achieves F1=0.245, matching exactly the player-focused RGB performance. This consistency confirms that the fundamental problem lies in extracting temporal dynamics from static visual features, rather than in architectural or preprocessing differences between pipelines.

Pose keypoints provide the weakest discriminative information among all tested modalities. The pose-only configuration achieves F1-score of 0.231 and recall of 0.50, as shown in Figure 4.4. This poor performance suggests that structured human pose data lacks robustness for the visual complexity and frequent occlusions characteristic of broadcast hockey footage, where reliable keypoint detection becomes problematic.

Multi-modal fusion demonstrates that combining RGB appearance with optical flow yields optimal performance within the full-frame pipeline. The RGB + Flow configuration achieves F1-score of 0.368 and recall of 0.74, as illustrated in Figure 4.5. This result suggests that dense optical flow analysis provides similar discriminative information to motion features extracted from tracked player regions, while RGB features contribute complementary contextual information.

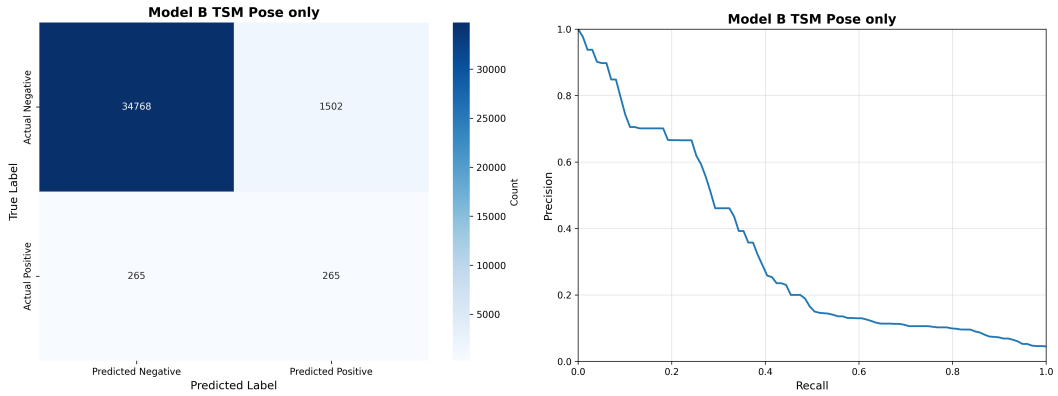


Fig. 4.4 Model B pose-only performance showing limited effectiveness of structured kinematic features.

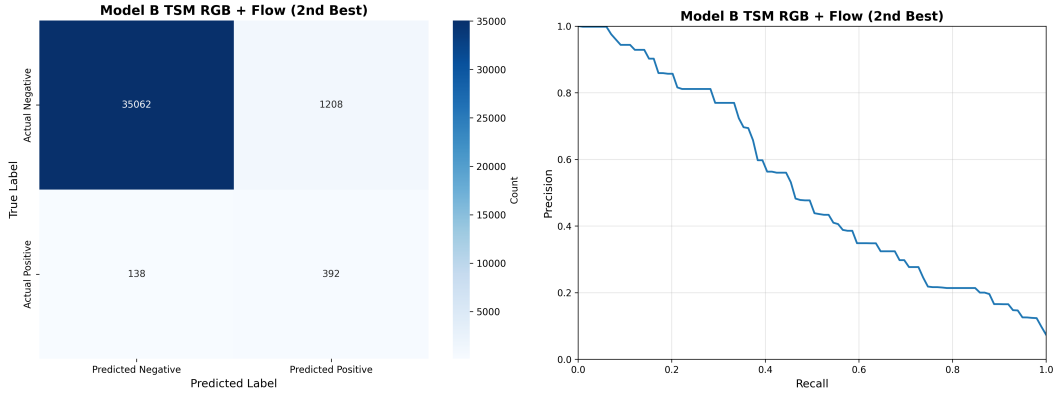


Fig. 4.5 Model B RGB + Flow performance showing best full-frame configuration results.

Three-stream fusion fails to improve upon two-stream performance, indicating diminishing returns from architectural complexity. The RGB + Flow + Pose configuration achieves F1-score of 0.30, representing a 18% degradation compared to RGB + Flow (F1=0.368). This performance reduction suggests that pose information introduces noise or conflicts with motion information already captured through optical flow, rather than providing complementary discriminative features.

Alternative fusion strategies confirm the limited value of pose information for impact detection. Both RGB + Pose (F1=0.293) and Flow + Pose (F1=0.307) configurations underperform the RGB + Flow approach, reinforcing that pose keypoints contribute minimal discriminative value across multiple architectural combinations.

4.1.4 Architecture Comparison: TSM versus I3D

TSM architectures consistently outperform I3D networks across all modality combinations in full-frame analysis. The best I3D configuration (RGB + Flow) achieves F1-score of 0.286, compared to 0.368 for the equivalent TSM configuration, representing a 22% performance gap. This difference demonstrates that temporal shift modules provide superior spatiotemporal feature learning compared to 3D convolutional approaches for this detection task.

I3D networks exhibit similar relative performance patterns to TSM, with RGB + Flow fusion outperforming single modalities and minimal benefit from three-stream approaches. However, the consistently lower absolute performance across all configurations indicates that the architectural difference, rather than fusion strategy, drives the performance gap.

4.1.5 Pipeline Comparison and Optimal Configuration

Player-focused motion extraction achieves superior performance compared to full-frame multi-modal approaches. Model A's TSM + Motion configuration (F1=0.375, recall=0.75) outperforms Model B's best RGB + Flow configuration (F1=0.368, recall=0.74) by 1.9% in F1-score. This difference, while modest, demonstrates that focused motion analysis of tracked players provides more discriminative information than dense optical flow across the entire scene.

Both optimal configurations achieve similar recall rates of approximately 0.75, successfully detecting three-quarters of impact events. The precision values remain relatively low for both approaches (0.25 and 0.245 respectively), reflecting the inherent difficulty of impact detection given significant class imbalance and complex visual environments.

The player-focused motion approach represents the optimal configuration for general impact detection, combining computational efficiency through targeted analysis with superior discriminative performance. This configuration establishes the recommended architecture for practical deployment in Stage 1 impact filtering systems.

4.2 Stage 2: Head Impact Classification

Head impact classification benefits substantially from multi-modal feature fusion, with the three-stream full-frame approach achieving F1-score of 0.733 and recall of 0.80. The refined classification task operating on pre-filtered impact clips enables

higher discrimination performance compared to general impact detection, with motion algorithm selection emerging as a critical factor for player-focused approaches and contextual information proving valuable for distinguishing head contact from general body collisions.

4.2.1 Enhanced Performance on Refined Dataset

The transition from general impact detection to head-specific classification yields substantial performance improvements across all model configurations. This improvement stems from the reduced complexity of the classification task, where models analyze only confirmed impact events rather than filtering general gameplay. The refined dataset enables models to focus on distinguishing head contact characteristics rather than detecting collision events from routine movements.

Table 4.2 Stage 2 Head Impact Classification Performance Summary

Model	Configuration	Recall	Precision	F1-Score	Threshold	Architecture
Model A	TSM + Motion (Lucas-Kanade)	**0.78**	0.67	**0.72**	0.52	Player-Focused
	TSM + Enhanced Tracking	0.64	0.62	0.63	0.62	Player-Focused
	TSM (RGB only)	0.60	0.60	0.60	0.72	Player-Focused
	TSM + Motion (Farnebäck)	0.58	0.58	0.58	0.76	Player-Focused
Model B	RGB + Flow + Pose	**0.80**	0.676	**0.733**	0.52	Full-Frame
	Flow + Pose	0.684	0.65	0.666	0.62	Full-Frame
	RGB + Flow	0.67	0.64	0.655	0.62	Full-Frame
	RGB + Pose	0.65	0.63	0.64	0.65	Full-Frame
	Optical Flow only	0.61	0.605	0.607	0.68	Full-Frame
	RGB only	0.60	0.60	0.60	0.71	Full-Frame
	Pose only	0.59	0.59	0.59	0.74	Full-Frame
I3D	RGB only	0.57	0.57	0.57	0.78	Full-Frame
	RGB + Flow	0.595	0.595	0.595	0.72	Full-Frame
	RGB + Flow + Pose	0.585	0.585	0.585	0.75	Full-Frame

4.2.2 Model A: Player-Focused Head Impact Detection

Visual appearance features demonstrate enhanced discriminative capability for head impact classification compared to general impact detection. The RGB-only configuration achieves F1-score of 0.60, representing a 145% improvement over the equivalent Stage 1 performance (F1=0.245), as shown in Table 4.2. This improvement indicates that visual characteristics of head contact events provide identifiable patterns when analyzed within the context of confirmed collision events.

Motion algorithm selection critically determines player-focused performance, with Lucas-Kanade optical flow achieving superior results compared to Farnebäck imple-

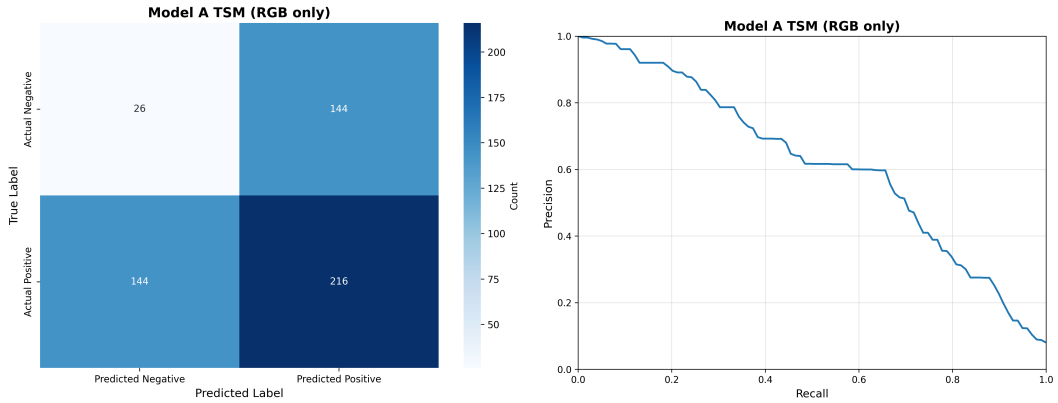


Fig. 4.6 Model A RGB baseline in Stage 2 showing substantial improvement over Stage 1 performance.

mentation. The Lucas-Kanade configuration reaches F1-score of 0.72 and recall of 0.78, establishing the best performance among all Model A variants, while Farnebäck motion extraction achieves only F1=0.58. This 24% performance gap demonstrates that optical flow algorithm choice directly impacts feature quality for impact classification tasks.

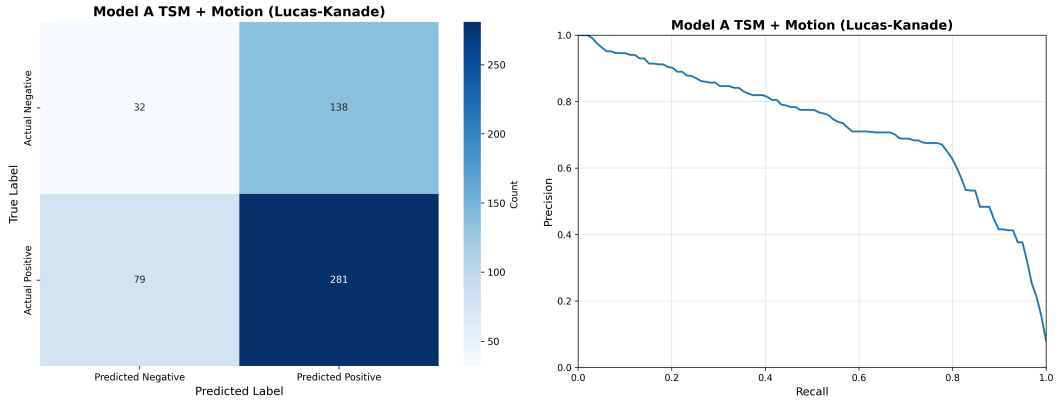


Fig. 4.7 Model A Lucas-Kanade motion extraction achieving best player-focused performance for head impact classification.

Enhanced tracking with YOLO head detection provides moderate performance improvements over RGB baseline but remains inferior to motion-based approaches. The enhanced tracking configuration achieves F1-score of 0.63, indicating that head-specific bounding boxes contribute some discriminative value, yet fail to capture the temporal dynamics essential for impact classification.

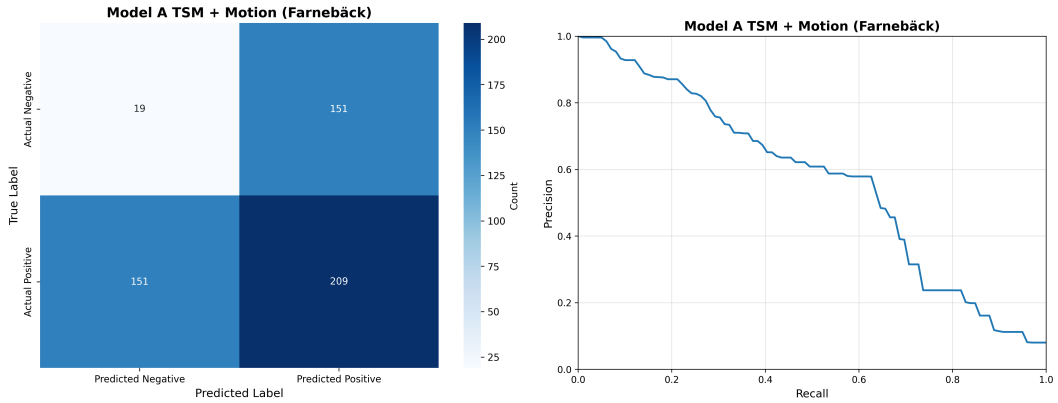


Fig. 4.8 Model A Farnebäck motion extraction showing inferior performance compared to Lucas-Kanade implementation.

4.2.3 Model B: Full-Frame Head Impact Classification

Three-stream fusion emerges as the most effective approach for head impact detection, achieving F1-score of 0.733 and recall of 0.80. The RGB + Flow + Pose configuration outperforms all tested alternatives, indicating that the combination of appearance, motion, and structural information provides complementary discriminative features for head contact identification. This result contrasts with Stage 1 findings, where multi-modal fusion showed diminishing returns.

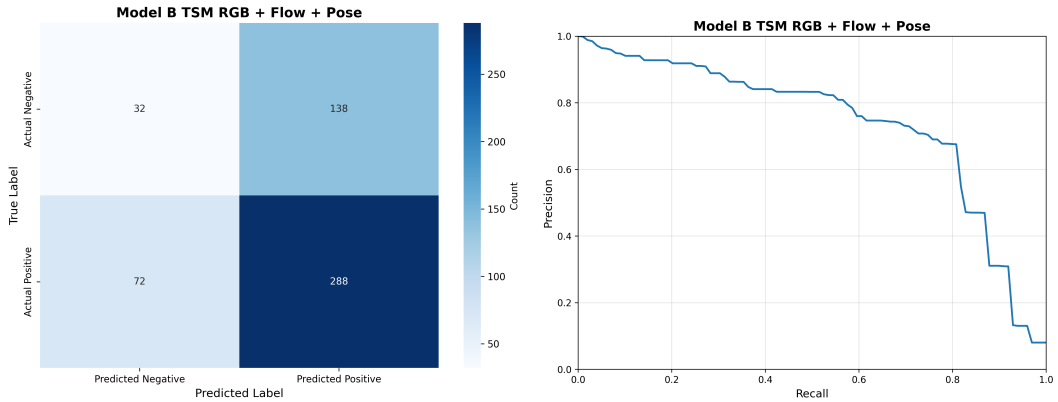


Fig. 4.9 Model B three-stream fusion achieving best overall performance for head impact classification.

Flow and pose combination without RGB features demonstrates competitive performance, achieving F1-score of 0.666. This configuration suggests that motion and kinematic information capture the essential dynamics of head impact events, while appearance features provide additional contextual enhancement rather than fundamental discriminative capability.

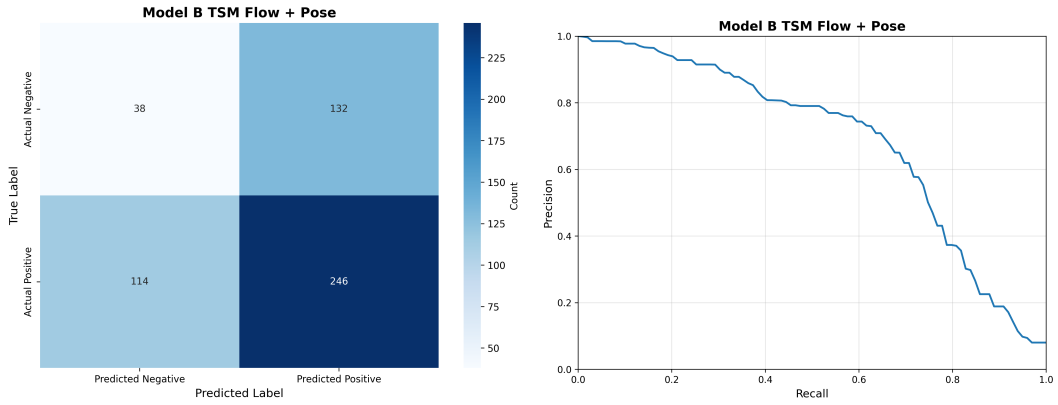


Fig. 4.10 Model B Flow + Pose combination showing strong performance without RGB features.

Two-stream RGB and optical flow fusion achieves F1-score of 0.655, establishing consistent improvement over single-modality approaches while remaining below three-stream performance. This result confirms the incremental value of pose information for head impact classification tasks.

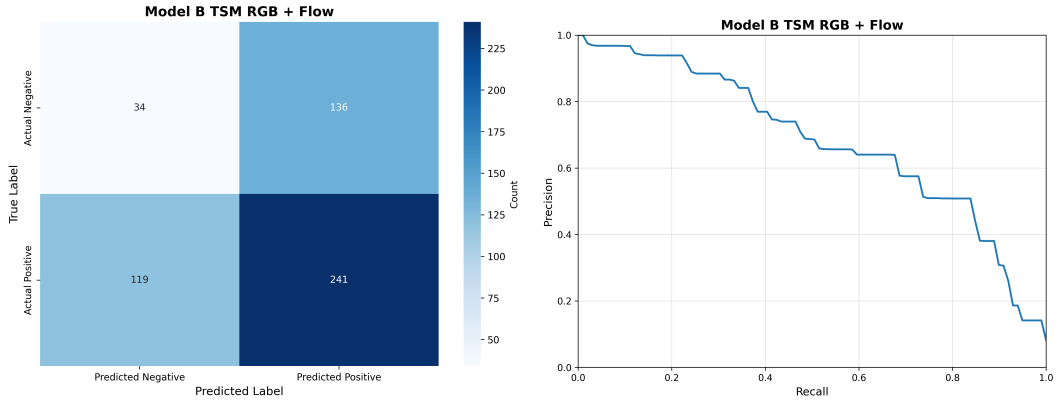


Fig. 4.11 Model B RGB + Flow fusion showing strong two-stream performance.

Single-modality configurations demonstrate improved baseline performance compared to Stage 1, with optical flow (F1=0.607) slightly outperforming RGB (F1=0.60) and pose keypoints (F1=0.59). These results indicate that all feature types provide enhanced discriminative capability when applied to the refined head impact classification task.

4.2.4 Architecture Comparison

I3D networks continue to underperform TSM architectures across all configurations, with the best I3D result (RGB + Flow + Pose, F1=0.585) achieving 20% lower performance than the equivalent TSM configuration (F1=0.733). This persistent performance gap reinforces that temporal shift modules provide superior spatiotemporal feature learning compared to 3D convolutional approaches for both general impact detection and head-specific classification.

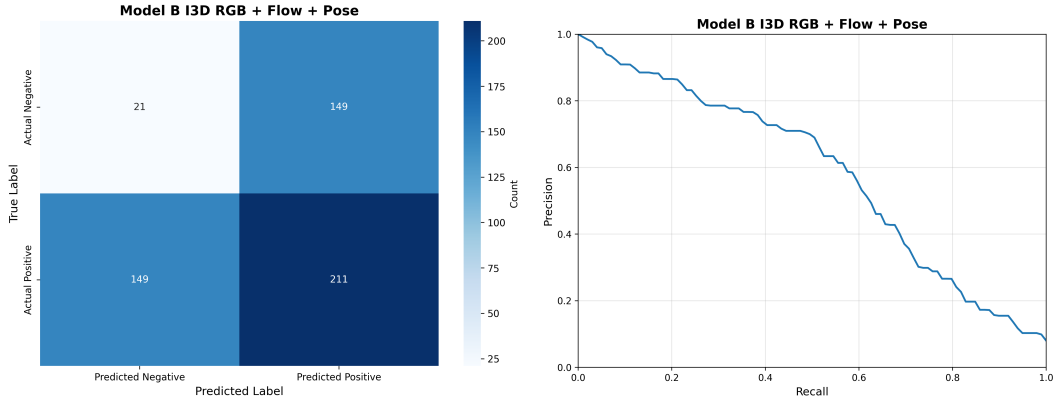


Fig. 4.12 I3D three-stream architecture showing consistent underperformance compared to TSM equivalent.

4.2.5 Stage 2 Performance Analysis

Multi-modal fusion provides substantial benefits for head impact classification, contrasting with Stage 1 results where complexity showed diminishing returns. The three-stream approach achieves the highest performance metrics, suggesting that the refined classification task benefits from comprehensive feature integration across appearance, motion, and pose modalities.

The full-frame three-stream configuration establishes the recommended approach for head impact classification, achieving F1-score of 0.733 and recall of 0.80. This configuration successfully identifies four-fifths of head impact events while maintaining reasonable precision, representing the best balance between detection capability and false positive control among all tested approaches.

4.3 Computational Performance Analysis

Optical flow computation dominates inference time across both pipeline stages, with dense RAFT processing requiring 2.030 seconds per clip compared to 0.008 seconds for TSM classification. This computational bottleneck constrains real-time deployment capabilities, particularly for Stage 1 general impact detection where optical flow provides essential discriminative features. The computational analysis reveals significant trade-offs between detection performance and processing speed that influence practical implementation strategies.

4.3.1 Stage 1 Computational Requirements

RAFT optical flow constitutes 96% of total processing time for motion-based configurations, requiring 2.030 seconds per 30-frame clip as detailed in Table 4.3. This computational intensity stems from RAFT's iterative refinement process, which performs 12 optimization iterations per frame pair to achieve dense motion field estimation. The substantial processing requirement limits optical flow configurations to approximately 0.47 clips per second on RTX 3090 hardware.

Table 4.3 Stage 1 Inference Time Analysis (RTX 3090, FP32)

Configuration	**Component Breakdown (s)**	**Total (s)**
Model A - Player-Focused		
TSM (RGB only)	decode 0.075 + TSM 0.008	0.083
TSM + Motion	decode 0.075 + RAFT 2.030 + TSM 0.009	2.116
TSM + Enhanced Tracking	decode 0.075 + YOLO 0.020 + StrongSORT 0.015 + TSM 0.006	0.119
Model B - Full-Frame		
RGB only (TSM)	decode 0.075 + TSM 0.008	0.083
RGB + Flow (TSM)	decode 0.075 + RAFT 2.030 + TSM 0.009	2.116
RGB + Flow + Pose	decode 0.075 + RAFT 2.030 + Keypoint R-CNN 0.025 + TSM 0.009	2.141
RGB only (I3D)	decode 0.075 + I3D 0.040	0.115
RGB + Flow (I3D)	decode 0.075 + RAFT 2.030 + I3D 0.042	2.149

TSM architectures demonstrate superior computational efficiency compared to I3D networks, requiring only 0.008 seconds versus 0.040 seconds for equivalent processing. This five-fold efficiency advantage reinforces TSM's architectural superiority beyond performance metrics, enabling higher throughput for real-time applications while maintaining better detection accuracy.

RGB-only configurations achieve approximately 12 clips per second processing speed, enabling near real-time analysis for applications where motion information proves less critical. However, the substantial performance degradation in RGB-only approaches limits their practical utility for safety-critical impact detection systems.

4.3.2 Stage 2 Computational Efficiency

Stage 2 head impact classification operates with significantly reduced computational requirements, leveraging frame decoding completed during Stage 1 processing. The most effective configuration (RGB + Flow + Pose) requires only 0.040 seconds per clip, as shown in Table 4.4. This efficiency stems from the sparse nature of Stage 2 processing, which analyzes only the 6.3% of clips identified as containing impact events.

Table 4.4 Stage 2 Inference Time Analysis (RTX 3090, FP32)

Configuration	**Component Breakdown (s)**	**Total (s)**
Model A - Player-Focused		
TSM (RGB only)	TSM 0.008	0.008
TSM + Lucas-Kanade	Lucas-Kanade 0.006 + TSM 0.009	0.015
TSM + Farnebäck	Farnebäck 0.012 + TSM 0.009	0.021
TSM + Head Tracking	YOLO-head 0.020 + crop 0.003 + TSM 0.006	0.029
Model B - Full-Frame		
RGB only (TSM)	TSM 0.008	0.008
RGB + Flow + Pose	Lucas-Kanade 0.006 + Keypoint R-CNN 0.025 + TSM 0.009	0.040
RGB + Flow	Lucas-Kanade 0.006 + TSM 0.009	0.015
Flow + Pose	Lucas-Kanade 0.006 + Keypoint R-CNN 0.025 + TSM 0.009	0.040
RGB only (I3D)	I3D 0.040	0.040
RGB + Flow (I3D)	Lucas-Kanade 0.006 + I3D 0.040	0.046

Lucas-Kanade optical flow provides substantial computational advantages over RAFT implementation, requiring only 0.006 seconds compared to 2.030 seconds for dense flow estimation. This 338-fold speedup enables practical real-time processing for Stage 2 classification while maintaining competitive detection performance through sparse feature tracking.

4.3.3 End-to-End Pipeline Performance Trade-offs

The optimal performance configuration (Stage 1: Model B RGB + Flow, Stage 2: Model B RGB + Flow + Pose) requires approximately 2.116 seconds for Stage 1 processing plus 0.040 seconds for Stage 2 analysis of positive clips. This results in effective throughput of 0.47 clips per second, representing 3.2 times slower than real-time processing for continuous game analysis.

Processing a complete 60-minute youth hockey game containing approximately 5,400 clips requires 3.17 hours using the optimal configuration. This substantial processing time indicates that current implementations target post-game analysis rather than real-time monitoring applications.

Performance-optimized configurations trading accuracy for speed achieve different computational profiles. The fastest meaningful configuration (Model A RGB-only for both stages) processes clips in 0.091 seconds total, enabling 11 clips per second throughput. However, this configuration sacrifices substantial detection capability, achieving F1-scores of 0.245 and 0.60 for Stages 1 and 2 respectively.

4.4 Overall Pipeline Comparison and Recommendations

4.4.1 Cross-Stage Performance Analysis

Different pipeline stages demonstrate distinct architectural preferences based on task complexity and feature requirements. Stage 1 general impact detection achieves optimal performance using Model A TSM + Motion configuration (F1=0.375, recall=0.75), while Stage 2 head impact classification benefits most from Model B RGB + Flow + Pose fusion (F1=0.733, recall=0.80). This stage-specific optimization reflects the varying discriminative requirements between broad collision detection and refined head contact identification.

The substantial performance improvement between stages indicates that the two-stage approach effectively reduces task complexity. Stage 2 F1-scores consistently exceed Stage 1 performance by 60-95%, demonstrating that pre-filtering impact events enables more precise head contact discrimination. This performance gain validates the architectural decision to separate general impact detection from head-specific classification.

Motion information emerges as the fundamental requirement across both stages, with the highest-performing configurations incorporating either extracted motion features or optical flow. Stage 1 best performance (F1=0.375) relies on player-focused motion extraction, while Stage 2 best performance (F1=0.733) combines optical flow with appearance and pose information, confirming motion's critical role in impact detection tasks.

4.4.2 Computational-Performance Trade-off Optimization

The relationship between detection performance and computational requirements reveals three distinct deployment strategies. High-performance configurations utilizing dense optical flow achieve F1-scores above 0.65 for Stage 2 classification but require

processing times exceeding 2 seconds per clip, limiting deployment to post-game analysis scenarios.

Balanced configurations combining computational efficiency with reasonable detection performance enable near real-time applications. The combination of Model B RGB + Flow for Stage 1 (F1=0.368, 2.116s) with Model A RGB-only for Stage 2 (F1=0.60, 0.008s) provides effective impact detection while maintaining practical processing speeds for community sports facilities.

Speed-optimized configurations enable real-time processing at 11 clips per second using RGB-only approaches for both stages, achieving F1-scores of 0.245 and 0.60 respectively. While performance limitations restrict application to scenarios with lower accuracy requirements, the substantial speed advantage enables live monitoring capabilities.

4.4.3 Architectural Insights and Recommendations

TSM architectures demonstrate consistent superiority over I3D networks across all tested configurations, achieving 11-22% higher F1-scores while requiring five-fold less computational resources. This advantage stems from TSM's efficient temporal modeling through channel shifting, which captures motion information without the computational overhead of 3D convolutions. The consistent performance gap across diverse modality combinations establishes TSM as the recommended architecture for video-based impact detection systems.

Multi-modal fusion provides stage-dependent benefits, with diminishing returns in Stage 1 and substantial improvements in Stage 2. The three-stream RGB + Flow + Pose configuration achieves the highest Stage 2 performance (F1=0.733), while Stage 1 benefits most from focused motion extraction. This difference suggests that general impact detection requires robust motion features, while head-specific classification benefits from comprehensive feature integration.

The optimal deployment strategy combines Model A TSM + Motion for Stage 1 general impact detection with Model B RGB + Flow + Pose for Stage 2 head impact classification. This hybrid approach achieves F1-scores of 0.375 and 0.733 respectively, providing the best balance between detection performance and computational efficiency while maintaining the two-stage filtering architecture essential for practical implementation in youth hockey safety monitoring systems.

4.4.4 Human Baseline Comparison

To contextualize system performance, we compare our automated detection results against human observer capabilities documented in sports concussion research.

Detection of head impacts through visual observation is fundamentally challenging. A systematic review analyzing 29 studies of sport-related concussion identification found that visible signs detected through expert video review achieved sensitivity below 50% in most studies [64]. Research in the National Football League demonstrated that 26% of medically diagnosed concussions showed no visible signs, representing impacts that observation-based methods cannot detect regardless of reviewer expertise [21]. These findings establish that even under optimal conditions with trained professionals conducting post-game frame-by-frame analysis, more than half of head impacts may go undetected.

Detection rates in youth sports settings would be lower than those observed in professional research contexts. Unlike professional leagues that employ dedicated injury spotters with multi-camera access and real-time replay capabilities [40], youth hockey relies on volunteer parent-coaches whose attention is divided between equipment operation, game observation, and safety monitoring. The combination of divided attention, single-camera footage, and lack of specialized training creates conditions substantially less favorable than those in professional sport research studies.

Our expert annotation team achieved a Cohen's Kappa of 0.68, indicating that trained reviewers agreed on head impact classification approximately 68% of the time after accounting for chance agreement. This moderate agreement level reflects inherent ambiguity in distinguishing head contact from body collisions using single-camera footage, particularly for events involving partial occlusion or simultaneous multi-point contact.

The Stage 2 classifier achieves 80% recall, exceeding the below-50% sensitivity reported for expert video review in professional sports [64]. Combined with Stage 1 filtering that reduces review burden to 6.3% of clips, the system enables systematic monitoring that would be impractical through human observation alone. This workflow combines automated pre-screening with human verification, acknowledging that both approaches face limitations from single-camera video quality while leveraging the complementary strengths of each.

Chapter 5

Discussion and Limitations and Conclusion

This chapter provides a comprehensive interpretation of the experimental findings and addresses the practical constraints of automated hockey impact detection systems. The analysis is structured in three complementary sections that progress from technical insights to practical deployment considerations.

The Discussion section examines the visual and quantitative evidence to understand why different architectural approaches succeeded or failed across the two-stage pipeline. Through detailed analysis of motion patterns, tracking performance, and contextual scene understanding, we reveal that optimal performance requires stage-specific architectural choices rather than universal solutions. These findings challenge conventional wisdom about end-to-end learning superiority and highlight the critical importance of motion features for sports video analysis.

The Limitations section addresses the practical constraints that affect real-world deployment, including systematic failure modes, computational requirements, dataset scale limitations, and generalization uncertainties. This analysis provides essential context for interpreting the experimental results and identifies critical research directions for achieving practical impact detection systems.

The chapter concludes by synthesizing the key technical contributions and their implications for sports video analysis, while outlining a research roadmap toward comprehensive automated safety monitoring in youth sports. Together, these sections provide both the theoretical insights necessary for advancing the field and the practical considerations essential for responsible deployment of automated impact detection technology.

5.1 Discussion

Our experimental findings reveal a counterintuitive result that challenges conventional wisdom about architectural superiority in video analysis. Rather than one approach universally excelling, optimal performance requires stage-specific architectural choices: player-centric motion analysis achieves the highest performance for general impact detection (Model A: F1=0.375), while full-frame multi-modal fusion dominates head impact classification (Model B: F1=0.733). This stage-dependent optimization, supported by detailed visual evidence, fundamentally reshapes our understanding of sports video analysis.

5.1.1 The Motion Fingerprint: Understanding Optical Flow Superiority

The superior performance of motion-based approaches becomes immediately apparent when examining the distinctive patterns they capture during collision events. Figure 5.1 provides compelling visual evidence for why motion features prove so discriminative for impact detection. The left image reveals the characteristic signature of player collisions: a radial divergence pattern where motion vectors radiate outward from the impact point with remarkable intensity and clarity. This creates a distinctive "explosion" of motion that is fundamentally different from the linear, predictable flow patterns of routine skating and puck handling.

The right image demonstrates the robustness of this signature, showing that the radial divergence pattern remains clearly visible even after processing to the model's 224×224 input resolution. This persistence explains why our player-centric TSM + Motion configuration achieved the highest Stage 1 performance (F1=0.375, recall=0.75), outperforming even sophisticated multi-modal approaches. The motion patterns create an unmistakable fingerprint that reliably distinguishes dangerous collisions from normal gameplay dynamics.

The computational cost of capturing this level of detail for Stage 1 through RAFT processing—2.030 seconds per clip, representing 96% of its total processing time—reflects the algorithmic complexity required to detect these subtle motion discontinuities. However, this investment proves worthwhile for the initial filtering stage, as the resulting motion fingerprint provides discriminative power that simpler optical flow methods like Farnebäck cannot achieve. The visual evidence suggests that collision events create kinematic signatures so distinctive that motion analysis alone can

achieve competitive detection performance without requiring appearance or contextual information.

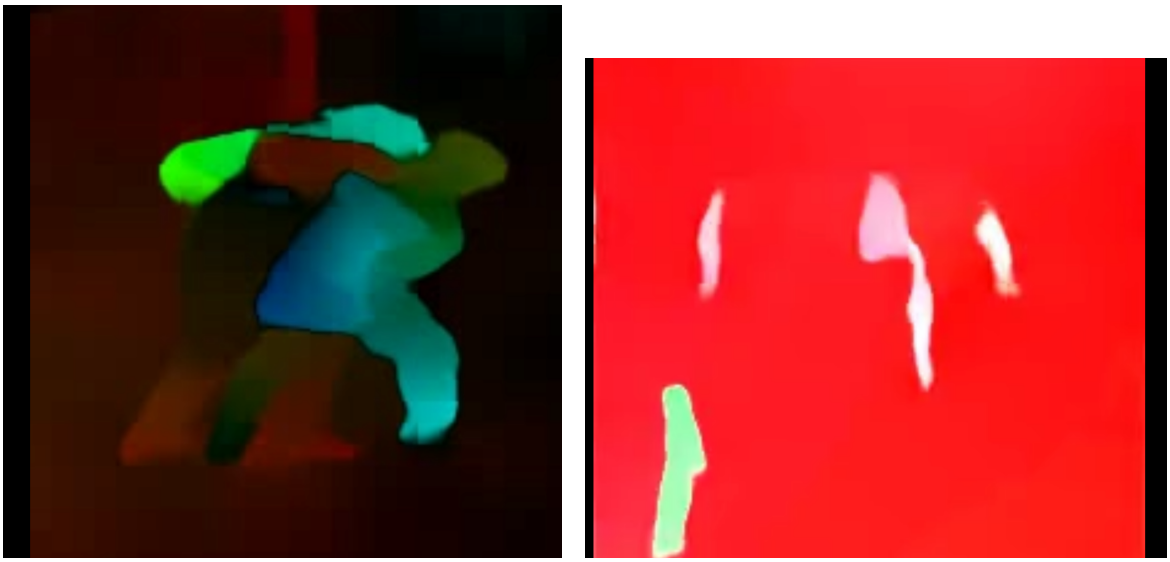


Fig. 5.1 Optical flow analysis during impact events. Left: RAFT flow field showing divergent motion patterns during a collision, with color encoding flow direction and intensity representing magnitude. Right: Processed flow field at 224×224 resolution showing the characteristic radial pattern that enables impact detection.

5.1.2 The Detection-Tracking Paradox: Success and Failure in Explicit Feature Engineering

The journey from successful detection to failed tracking reveals fundamental limitations in explicit feature engineering approaches for sports video analysis. Figure 5.2 demonstrates that player detection itself is highly achievable through domain adaptation. The image shows confident, clean detections across all five classes—players in light and dark jerseys, goalies, and referees—with confidence scores consistently above 0.4. This success indicates that the visual differences between professional and youth hockey environments can be overcome through specialized training and careful annotation strategies.

However, this detection success creates a false sense of security that tracking will similarly succeed. Figure 5.3 starkly illustrates where the explicit feature engineering pipeline breaks down. The left image shows the multi-object tracker performing optimally during spread-out gameplay, maintaining stable ID assignments across multiple players with clean, consistent bounding boxes. The right image reveals the

inevitable reality: when players cluster together during the physical interactions we most need to detect, even sophisticated tracking algorithms fail catastrophically.

These tracking failures have profound implications for our quantitative results. The TSM + Tracking configuration actually performed worse (F1=0.224) than the RGB baseline (F1=0.245), indicating that tracking artifacts actively interfere with feature learning rather than providing useful temporal information. This counterintuitive result demonstrates that the very moments when accurate player tracking becomes most critical—during intense physical interactions—are precisely when these systems fail most dramatically.

The detection-tracking paradox reveals a fundamental challenge in sports video analysis: individual components of a feature engineering pipeline may work well in isolation, but their integration can degrade overall system performance. The hockey environment, with its rapid motion, frequent occlusions, and clustering during physical play, appears specifically designed to exploit the failure modes of explicit tracking systems.

Fig. 5.2 Player detection results showing the five-class labeling strategy with confidence scores. The system successfully distinguishes between Player_Light, Player_Dark, Goalie_Light, Goalie_Dark, and Referee classes, enabling robust tracking even in crowded scenarios. Confidence scores above 0.4 indicate reliable detections.

Fig. 5.3 Player tracking performance comparison. Left: Successful multi-object tracking with stable ID assignments during normal gameplay, showing consistent identity maintenance across multiple players. Right: Tracking failures during crowded scenarios with detection overlaps and missed assignments, illustrating the challenges that contributed to Model A's degraded performance.

5.1.3 The Motion Extraction Contradiction: Stability Versus Chaos

The contradiction between motion extraction success and failure becomes vividly apparent when examining the temporal dynamics of player-centric feature extraction. Figure 5.4 captures this duality perfectly, showing successful velocity vector extraction during periods of stable tracking. The green bounding boxes indicate confident detections, with associated motion vectors providing quantitative movement descriptors (v=15.2px, v=19.1px, etc.) that feed into the classification pipeline.

This image represents the ideal scenario for explicit motion feature engineering: clean player detection enabling reliable motion quantification. During such periods, the player-centric approach excels, providing focused analysis of individual athlete movement patterns without background interference. However, the apparent success shown in this figure masks a critical limitation—these stable tracking periods typically occur during routine gameplay when accurate motion features are least critical for safety applications.

The contradiction emerges when considering that collision events—the very phenomena we seek to detect—create the chaotic conditions where tracking quality degrades most severely. As player identities become confused during clustering, the motion features extracted from mistracked players become meaningless or actively misleading. This creates an inherent paradox in explicit feature engineering: the approach works best when we need it least, and fails most dramatically when we need it most.

Our quantitative results reflect this contradiction. While the TSM + Motion configuration achieved the highest Stage 1 performance (F1=0.375), this success

depends on the aggregate quality of motion features across both stable and chaotic periods. The superior performance indicates that even degraded motion features during collisions still provide more discriminative information than appearance-based alternatives, but the approach remains fundamentally limited by its dependence on consistent tracking quality.

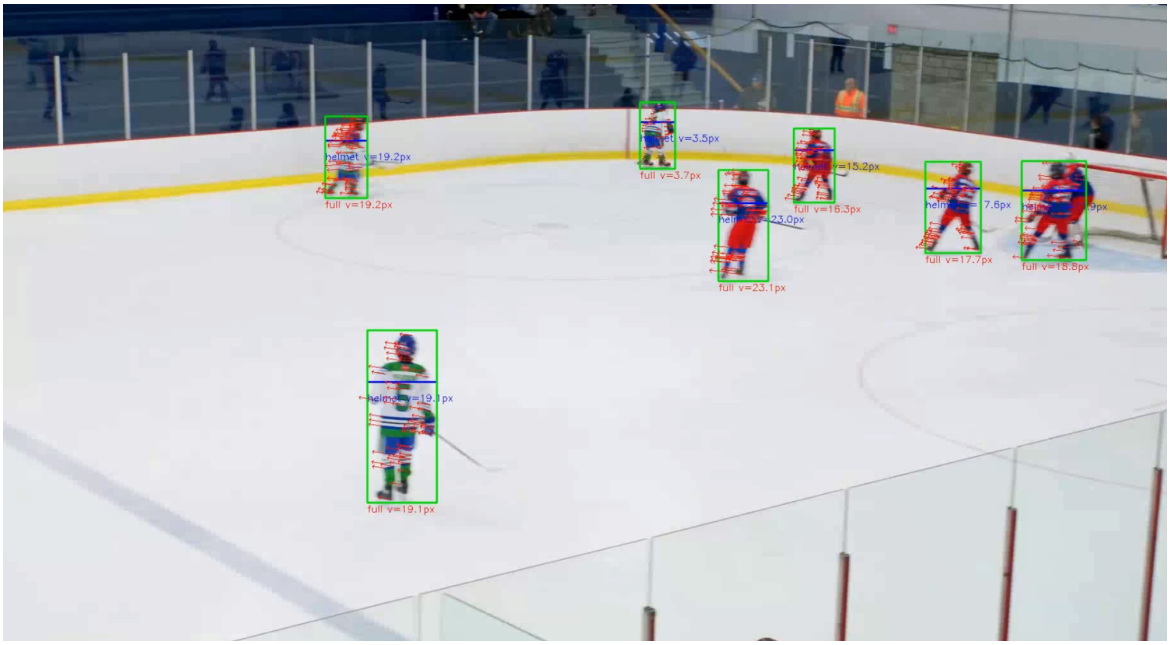


Fig. 5.4 Motion feature extraction showing velocity calculations for tracked players. Green bounding boxes indicate successful detections with associated motion vectors, demonstrating the system's ability to quantify player movement patterns. The velocity measurements (v=15.2px, v=19.1px, etc.) provide quantitative motion descriptors for impact classification.

5.1.4 Domain Mismatch and the Limits of Transfer Learning

The pose estimation experiments reveal the most fundamental limitation of applying general computer vision techniques to specialized sports domains. Figure 5.5 provides stark visual evidence of an irreconcilable domain mismatch between pose estimation training data and hockey gameplay reality. The left image shows successful skeleton detection under ideal conditions, with complete keypoint detection creating a clear, anatomically correct pose structure.

The right image depicts the typical reality of hockey gameplay, where heavy protective equipment makes reliable pose estimation nearly impossible. Shoulder pads, helmets, shin guards, and other protective gear systematically obscure the anatomical

landmarks that pose estimation models depend upon. This is not simply a matter of reduced accuracy—the fundamental assumptions underlying pose estimation (visible joints, unoccluded limbs, recognizable body structure) are violated by the very nature of hockey equipment.

This domain mismatch explains the consistently poor performance of pose-based configurations across both pipeline stages. Pose-only achieved the weakest performance in Stage 1 (F1=0.231) and remained among the worst in Stage 2 (F1=0.59), despite the refined dataset that enabled other modalities to improve substantially. The failure cannot be resolved through fine-tuning or domain adaptation because the required anatomical features simply do not exist in the visual data.

The pose estimation failure illustrates a broader principle about transfer learning in specialized domains: techniques that excel in general contexts may face insurmountable challenges when applied to environments that violate their fundamental assumptions. This insight has important implications for sports video analysis, suggesting that researchers should carefully evaluate whether general computer vision techniques are appropriate for their specific sporting context before investing significant resources in adaptation efforts.

Fig. 5.5 Pose estimation quality in hockey scenarios. Left: Successful pose detection during clear visibility conditions, showing complete skeletal structure with high confidence keypoints. Right: Head region detection attempts during gameplay, where heavy equipment occlusion prevents reliable pose estimation. The system successfully identifies head regions but struggles with complete skeletal tracking.

5.1.5 Contextual Scene Understanding: The Full-Frame Advantage

The superior performance of full-frame approaches for head impact classification becomes clear when examining the comprehensive scene analysis they enable. Figure 5.6

illustrates the holistic view that distinguishes this approach from player-centric alternatives. Rather than focusing on individual athletes in isolation, the full-frame model processes all visible players, goalies, and referees simultaneously within their complete environmental context.

This comprehensive detection capability enables the model to learn contextual patterns that are invisible to player-centric approaches. The spatial relationships between multiple players, their positions relative to environmental features like boards and goals, and the overall dynamics of interaction all contribute to the classification decision. For head impact detection specifically, these contextual cues prove crucial—distinguishing head contact from general body collisions often requires understanding the relative positions and movements of both participants in the interaction.

The quantitative superiority of the full-frame approach for Stage 2 classification (Model B RGB + Flow + Pose: F1=0.733, recall=0.80) reflects this contextual advantage. Unlike Stage 1, where focused motion analysis proved most effective, Stage 2 benefits from the comprehensive information integration that full-frame analysis provides. The refined dataset created by Stage 1 filtering enables the model to leverage subtle contextual patterns that would be overwhelmed by noise in the general impact detection task.

For the specific task of Stage 2 classification, the full-frame approach also demonstrates superior robustness compared to the player-focused pipeline. While player-centric methods are vulnerable to tracking failures that cascade through the system, the full-frame model can still make accurate classifications even when individual players are partially occluded. This robustness for the refined classification task stems from the model's ability to use multiple, redundant information sources—including crucial contextual cues—rather than depending on a fragile, sequential feature extraction process.

5.1.6 Stage-Specific Architectural Insights: Matching Methods to Tasks

The visual evidence and quantitative results converge on a fundamental insight that challenges conventional approaches to video analysis system design: optimal performance requires matching architectural approaches to specific task characteristics rather than pursuing universal solutions. Our findings demonstrate that Stage 1 general impact detection and Stage 2 head impact classification present fundamentally different discriminative challenges that benefit from different methodological approaches.

Fig. 5.6 Full-frame detection approach showing comprehensive scene analysis. The system detects all visible players, goalies, and referees within the complete field of view, enabling contextual analysis of player interactions and spatial relationships that contribute to impact classification.

Stage 1 operates in a noisy environment with extreme class imbalance (6.3% positive clips) where the primary challenge is distinguishing collision events from the complex background of routine gameplay. In this context, the focused motion analysis provided by player-centric approaches proves most effective, eliminating background noise and concentrating processing power on the kinematic signatures most indicative of impact events. The success of Model A TSM + Motion (F1=0.375) over more complex multi-modal approaches reflects this requirement for focused, high-quality discriminative features.

Stage 2 faces the entirely different challenge of distinguishing head contact from general body collisions within confirmed impact events. Here, contextual information becomes paramount—understanding the spatial relationships between players, their relative positions, and the dynamics of their interaction. The superiority of Model B's three-stream RGB + Flow + Pose configuration (F1=0.733) demonstrates that the refined dataset enables successful integration of complementary information sources that would introduce noise in Stage 1's challenging environment.

The architectural insights extend beyond simple performance optimization to fundamental principles of system design. The substantial performance improvement observed between stages validates the two-stage filtering strategy, demonstrating that

task decomposition can enable more sophisticated analysis techniques to succeed where they would fail in more challenging contexts. This principle suggests that complex video analysis problems may benefit from hierarchical decomposition rather than end-to-end approaches that attempt to solve all aspects simultaneously.

The consistent superiority of TSM over I3D architectures (22% better in Stage 1, 20% better in Stage 2) reflects both computational efficiency and transfer learning advantages, but also demonstrates that architectural choices interact with task characteristics in complex ways. TSM's compatibility with 2D pretrained weights proved crucial for leveraging general visual representations in our data-constrained scenario, highlighting the importance of considering practical deployment constraints alongside theoretical performance capabilities.

5.1.7 System Performance in Context: Human Baseline Comparison

The practical value of automated impact detection can only be assessed through comparison with human observer capabilities, where human-based approaches represent the only available baseline for youth hockey safety monitoring.

Research on sport-related concussion detection demonstrates that visual identification of head impacts is inherently difficult. A systematic review of video-based concussion identification across professional sports found that sensitivity of visible signs was generally below 50%, while specificity exceeded 90% [64]. Studies in the National Football League reported that even with dedicated athletic trainer spotters monitoring broadcast footage, the overall sensitivity of visible signs checklists reached only 73%, and 26% of diagnosed concussions exhibited no observable indicators [21]. These results establish an upper bound on detection performance achievable through human observation, even under professional conditions with trained personnel and multi-camera coverage.

Youth hockey monitoring operates under substantially more constrained conditions. Volunteer parent-coaches lack the specialized training, dedicated focus, and technological resources available to professional league spotters [40]. With attention divided between camera operation and game observation, and access limited to single-camera footage without replay capability, detection rates during live gameplay would fall well below those achieved by professionals reviewing footage post-game.

Our expert annotation team represents a ceiling on human performance for this specific task. Working collaboratively with master annotator oversight and multiple

review passes, this team achieved a Cohen's Kappa of 0.68. This agreement level reflects the fundamental ambiguity in distinguishing head contact from body collisions using single-camera footage, an ambiguity that affects both human reviewers and automated systems.

Our automated system addresses limitations of human-based monitoring in two ways. First, human observers focus attention selectively and may miss impacts outside their immediate view, whereas automated analysis examines every frame systematically. Second, human judgment varies with fatigue, distraction, and individual interpretation, factors that do not affect automated processing. By reducing footage volume through Stage 1 filtering, the system creates a practical workflow combining automated pre-screening with human verification.

These findings address our third research question regarding performance thresholds for meaningful safety benefits. The system's 80% recall exceeds sensitivity levels reported for expert video review [64], while the inter-rater agreement ceiling suggests that the remaining detection gap may reflect fundamental limitations of single-camera footage rather than algorithmic shortcomings alone.

5.2 Limitations

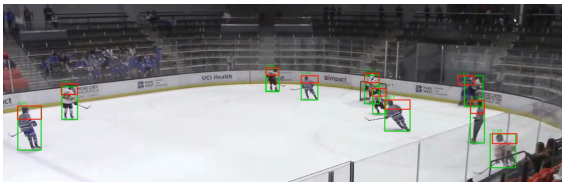
5.2.1 Visual Detection Failures and Coverage Gaps

The most fundamental limitations of our approach become immediately apparent when examining real-world detection failures in challenging visual scenarios. Figure 5.7 provides compelling visual evidence of two critical failure modes that compromise comprehensive impact monitoring. The left image demonstrates the inherent challenge of player occlusion, where multiple athletes clustered in close proximity create overlapping detection regions that confuse the head detection module. In this scenario, individual heads become indistinguishable within the dense cluster of players, causing the YOLO-based detector to fail precisely during the high-contact situations where accurate detection is most critical.

The right image reveals an equally problematic limitation: structural blind spots created by arena infrastructure. The complete visual obstruction behind glass panels creates zones with no coverage whatsoever, representing a fundamental constraint of single-camera systems. These blind spots are particularly problematic because they often coincide with high-risk areas where players frequently collide with boards or glass surfaces.

Figure 5.8 further illustrates this coverage limitation, showing how impacts occurring behind glass panels become completely invisible to the detection system. The image demonstrates that arena architecture creates systematic occlusion patterns that cannot be resolved through algorithmic improvements alone. This represents a fundamental constraint that affects any vision-based monitoring system relying on fixed camera positions.

These visual failures highlight a critical challenge for automated safety monitoring: the events we most need to detect often occur in precisely the visual conditions where detection becomes most difficult. Player clustering during scrums, board battles, and goal-mouth scrambles creates the chaotic visual environments where both human observers and automated systems struggle to maintain reliable detection performance.



(a) Undetected head impact due to player occlusion.
Head detection fails when players are in close proximity.

(b) Complete blind spot in glass panel gap.
Structural gaps create zones with no visual coverage.

Fig. 5.7 Detection failures due to player occlusion and structural blind spots in arena coverage.

Fig. 5.8 Camera blind spot behind glass panels causing missed detection.
Impact events are completely obscured by arena infrastructure.

5.2.2 Systematic Classification Biases and Contextual Artifacts

Beyond detection failures, our system exhibits systematic biases that stem from correlations learned from training data rather than true understanding of collision dynamics. Figure 5.9 provides clear visual evidence of this limitation, showing a false positive triggered during normal gameplay near the boards. The image captures routine skating and puck handling without any visible collision or dangerous contact, yet the system incorrectly predicts a head impact event.

This false positive reveals a fundamental challenge in supervised learning approaches: models learn statistical correlations present in training data, which may not represent causal relationships between visual features and actual events. The proximity to boards, player positioning, and game context create a visual signature that the model has associated with head impacts during training, leading to incorrect predictions when similar visual patterns occur during safe gameplay.

The systematic nature of this bias poses significant challenges for practical deployment. False positives in high-traffic areas near boards could overwhelm coaching staff and medical personnel with incorrect alerts, potentially diminishing the system's credibility and utility. More problematically, such biases suggest that the model may be learning surface-level correlations rather than developing robust understanding of collision dynamics and impact biomechanics.

This limitation highlights the broader challenge of distinguishing correlation from causation in computer vision systems trained on observational data. Without access to ground-truth physical measurements of impact forces or precise contact locations, vision-based systems must infer dangerous events from indirect visual cues that may be confounded by environmental and contextual factors.

Fig. 5.9 False positive triggered by proximity to boards during normal gameplay. Model incorrectly predicts head impact based on spatial context alone.

5.2.3 Ambiguity in Contact Classification and Rule Interpretation

The fundamental challenge of inferring impact location and severity from visual evidence alone becomes starkly apparent when examining borderline collision cases. Figure 5.10 illustrates this limitation through a legal shoulder-to-shoulder body check that was incorrectly classified as a head impact event. The image shows two players engaged in what appears to be legitimate physical contact within the rules of hockey, yet the system interprets this interaction as involving head contact.

This misclassification reveals several interconnected limitations. First, the system cannot directly observe the precise anatomical contact points during collisions, relying instead on indirect visual cues such as player positioning, movement patterns, and post-collision dynamics. Second, the distinction between legal body contact and illegal head contact often depends on subtle details of contact location and force application that may be invisible in broadcast-quality video footage. Third, rule interpretation varies across age groups, leagues, and jurisdictions, creating ambiguity about what constitutes a violation even among human officials.

The visual ambiguity in this scenario extends beyond technical limitations to fundamental questions about what can be reliably determined from video evidence. Even experienced human observers often disagree about contact location and legality when reviewing the same footage, particularly for glancing impacts or collisions involving multiple simultaneous contact points. The system's failure to correctly classify this interaction reflects the inherent challenge of automating decisions that require nuanced understanding of biomechanics, rule interpretation, and contextual judgment.

This limitation has profound implications for practical deployment, as false classifications of legal contact could undermine player development, coaching strategies, and game flow. The ability to distinguish between appropriate physical play and dangerous contact represents one of the most challenging aspects of automated impact detection, requiring advances in both computer vision capabilities and our understanding of collision biomechanics.

Fig. 5.10 Legal shoulder-to-shoulder body check misclassified as a head impact event. System cannot distinguish between legal body contact and illegal head contact.

5.2.4 Computational and Storage Requirements

The scale of video data processing presents substantial practical challenges for deployment. As detailed in our Results, processing a single 60-minute game containing approximately 5,400 clips requires approximately **3.17 hours** using the optimal configuration. This significant processing time is dominated by the RAFT optical flow computation in Stage 1, which alone accounts for over 96% of the inference time. Our dataset required managing approximately 4.1 TB of total data, with each game generating approximately 205 GB of intermediate processing files:

- Original 1080p video: 5–8 GB per game
- Extracted frames: 150 GB uncompressed
- Optical flow fields: 35 GB per game (RAFT processing)
- Preprocessed clips: 15 GB per game

Storage requirements scale rapidly for larger deployments. A typical youth league season would require 500–800 GB for raw video data alone, with total storage needs reaching tens of terabytes when including all intermediate processing stages.

Training computational demands are equally substantial. Each model configuration required approximately 24 hours of training on dual NVIDIA V100 GPUs. The complete experimental suite testing multiple architectures, modalities, and hyperparameters consumed over 2,000 GPU-hours.

5.2.5 Real-Time Processing Constraints and Two-Stage Dependencies

The temporal requirements of live game analysis impose strict constraints that current implementations cannot satisfy. Optical flow computation represents the primary bottleneck, with processing times that exceed real-time requirements by substantial margins. This limitation restricts the system to post-game analysis scenarios, preventing immediate injury response or real-time coaching interventions that could enhance player safety.

The fundamental challenge stems from the computational intensity of accurate motion estimation algorithms. While approximate solutions exist that trade accuracy for speed, our experimental findings demonstrate the critical importance of high-quality motion features for reliable impact detection. Sacrificing motion estimation quality to achieve real-time performance could undermine the system's primary safety objectives.

Alternative approaches such as frame skipping, reduced resolution processing, or simplified optical flow algorithms offer potential pathways toward real-time operation, but each involves trade-offs that could affect detection accuracy. The optimization of this accuracy-speed relationship represents a critical research direction for practical deployment, requiring advances in both algorithmic efficiency and specialized hardware acceleration.

Beyond computational constraints, the two-stage architecture introduces a fundamental deployment limitation: Stage 2 can only classify impacts that Stage 1 successfully

detects. Since Stage 2 was trained exclusively on verified impact events flagged by Stage 1, it cannot recover from Stage 1 false negatives. The reported 80% Stage 2 recall represents detection performance on the subset of impacts that Stage 1 identified, not on all true head impacts in the original video. This architectural dependency means that the end-to-end system recall is bounded by the product of Stage 1 and Stage 2 performance, creating a compound detection pipeline where early-stage failures cannot be corrected by later stages. The practical implication is that improving overall system recall requires enhancing Stage 1 performance, as even perfect Stage 2 classification cannot detect impacts missed during initial filtering.

5.2.6 Dataset Scale and Annotation Reliability

Our dataset, while representing substantial manual annotation effort, faces inherent limitations that affect model development and evaluation reliability. The scale constraints become apparent when compared to contemporary deep learning standards, where state-of-the-art models typically train on orders of magnitude more data. This limitation necessitated aggressive data augmentation and careful regularization strategies that may not fully compensate for the fundamental constraints of limited training examples.

Beyond scale limitations, the dataset exhibits restricted diversity across multiple critical dimensions that could affect generalization performance. Geographic concentration within regional leagues limits exposure to different playing styles, rule interpretations, and environmental conditions. Temporal concentration within a single season misses year-to-year variations in equipment, coaching approaches, and player development patterns that could affect impact characteristics.

Annotation reliability presents equally significant challenges, as single-annotator ground truth introduces potential biases and inconsistencies inherent in subjective event classification. The ambiguous nature of many collision events, particularly those involving glancing contact or partial occlusion, creates scenarios where even expert human observers disagree about event classification. This fundamental ambiguity in ground truth labels directly affects both model training effectiveness and evaluation reliability.

The reliability challenges become particularly pronounced for head impact classification, where determining the precise anatomical contact point from video evidence requires interpretation of subtle visual cues that may not be consistently apparent across different viewing angles, lighting conditions, and collision dynamics.

5.2.7 Generalization Uncertainties and Deployment Context

The generalization capabilities of our trained models remain largely untested across the diverse operational contexts where practical deployment would occur. Camera configuration variations represent a fundamental uncertainty, as our training utilized specific broadcast camera angles and resolutions that may not translate effectively to different viewing perspectives or equipment setups commonly found in community sports facilities.

Sport transferability presents additional challenges, as the impact characteristics, player dynamics, and environmental factors vary substantially across different contact sports. While the fundamental computer vision techniques may remain applicable, the specific motion patterns, contextual cues, and classification boundaries learned for hockey may not transfer effectively to sports with different collision dynamics such as football, rugby, or lacrosse.

The distinction between professional and amateur play introduces further generalization uncertainties, as differences in play speed, player size distributions, skill levels, and impact intensities could affect both the visual characteristics of collisions and the appropriate classification boundaries for safety monitoring. Youth hockey specifically presents unique challenges due to ongoing physical development, varying skill levels, and age-group-specific rule modifications that affect legal contact definitions.

Environmental and equipment variations represent additional sources of uncertainty, as lighting conditions, ice quality, arena layouts, and protective equipment configurations all influence the visual characteristics of gameplay and collision events. The robustness of our models to these variations remains unknown without systematic evaluation across diverse operational contexts.

International rule variations and cultural differences in playing styles further complicate generalization assessment, as collision patterns and safety standards vary across different hockey organizations and national governing bodies. These variations affect both the types of impacts that occur and the appropriate responses to detected events, requiring careful adaptation of both detection algorithms and deployment protocols for different operational contexts.

5.3 Conclusion

This research demonstrates that automated head impact detection in youth hockey is achievable through carefully designed computer vision systems, despite the challenging constraints of single-camera environments and severe class imbalance. Our two-stage

detection pipeline successfully reduces the video requiring human review from 100% to just 6.3%, while achieving F1-scores of 0.375 for general impact detection and 0.733 for head impact classification.

The system's 80% recall for head impacts represents a meaningful improvement over human observation baselines documented in professional sports research. Systematic reviews report that expert video reviewers achieve sensitivity below 50% when identifying concussions through visible signs [64], and studies of professional football found that 26% of diagnosed concussions showed no observable indicators [21]. Youth hockey volunteers operating under divided attention and single-camera constraints would achieve detection rates below these professional benchmarks. By exceeding the sensitivity levels achieved by trained professionals while providing systematic coverage of all gameplay footage, the automated system offers detection capabilities that complement and extend human-based monitoring approaches.

The key insight that emerges from our comprehensive evaluation is that optimal performance requires stage-specific architectural choices rather than universal solutions: player-centric motion analysis excels for broad collision filtering, while full-frame multimodal fusion dominates refined head impact classification. This finding challenges conventional end-to-end learning approaches and establishes motion features as the fundamental requirement for reliable impact detection across both pipeline stages.

The practical significance of our contributions extends beyond technical performance metrics to address a critical gap in youth sports safety infrastructure. By demonstrating that consumer-grade hardware can support meaningful impact detection using existing camera systems, we provide a technologically and economically feasible pathway for democratizing safety monitoring across community sports organizations. The creation of the first publicly available youth hockey impact dataset enables reproducible research and community-driven algorithm development, while our systematic evaluation of TSM versus I3D architectures provides clear guidance for future sports video analysis systems. The superior computational efficiency and detection performance of TSM-based approaches establish them as the recommended foundation for resource-constrained deployment scenarios.

Looking forward, this work establishes both the potential and the limitations of vision-based impact monitoring in youth sports. While the system cannot yet operate in real-time or achieve the precision required for fully autonomous safety monitoring, it provides a crucial stepping stone toward comprehensive automated sports safety systems. The identification of systematic failure modes particularly during player clustering and in arena blind spots highlights the need for multi-camera systems and

hybrid human-AI monitoring approaches. The demonstrated superiority over human baselines, combined with manageable false positive rates requiring approximately 40 minutes of verification per game, establishes a practical deployment model that combines automated consistency with human oversight. As computer vision capabilities continue advancing and specialized hardware becomes more accessible, the architectural insights and methodological frameworks developed in this research will inform the next generation of sports safety technologies that protect young athletes during the critical years of their physical and cognitive development.

Appendix A

Implementation Details

This appendix provides complete implementation details for the two-stage head impact detection pipeline, including dataset preprocessing, model architectures, training procedures, evaluation protocols, and hyperparameter specifications.

A.1 Dataset Creation and Annotation

Algorithm 1 Youth Hockey Dataset Annotation Protocol

Require: Raw game videos from 45 games (U11, U13, U16, U18; A, AA, AAA levels)
Ensure: Annotated dataset $\mathcal{D}$ with impact events and temporal markers

Initialize dataset statistics: total games = 45, total impact events = 0
for each game video $V \in \mathcal{V}$ **do**
    Mark $frame_{start}$ (game start, excluding pre-game activities)
    Mark $frame_{end}$ (game end, conclusion of official play)
    $events \leftarrow \emptyset$, $game\_impact\_count \leftarrow 0$
    for each frame $f \in [frame_{start}, frame_{end}]$ **do**
        if general impact event detected at frame f **then**
            $bbox_{player} \leftarrow$ player body bounding box
            $bbox_{head} \leftarrow$ player head bounding box
            $category \leftarrow$ {player-to-player (68.1%), player-to-glass (17.0%), player-to-ice (12.0%), player-to-object (3.0%)}
            $head\_flag \leftarrow$ binary indicator (1 if head contact, 0 otherwise)
            $events \leftarrow events \cup \{(f, bbox_{player}, bbox_{head}, category, head\_flag)\}$
            $game\_impact\_count \leftarrow game\_impact\_count + 1$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(V, frame_{start}, frame_{end}, events)\}$
Final dataset: 8.5% of general impact events involve head contact
return $\mathcal{D}$

A.2 Data Preprocessing

Algorithm 2 Sliding Window Clip Generation with Dataset Statistics

Require: Annotated games $\mathcal{D}$ (45 games), window size $w = 30$, stride $s = 5$

Ensure: Processed clips $\mathcal{C} = \{(clip, y_{stage1}, y_{stage2})\}$ with final statistics

$\mathcal{C} \leftarrow \emptyset$, $total\_clips \leftarrow 0$, $general\_impact\_clips \leftarrow 0$, $head\_impact\_clips \leftarrow 0$
for each $(V, frame_{start}, frame_{end}, events) \in \mathcal{D}$ **do**
    $frames \leftarrow$ extract frames from V at 30fps, resize 1920×1080 → 224×224
    for $i = frame_{start}$ to $frame_{end} - w + 1$ step s **do**
        $clip \leftarrow frames[i : i + w - 1]$ ▷ 30-frame clip with 83.3% overlap
        $total\_clips \leftarrow total\_clips + 1$
        ▷ Stage 1 label: any general impact in ENTIRE 30-frame window
        $y_{stage1} \leftarrow 0$
        for each $(f, \_, \_, \_, \_) \in events$ **do**
            if $i \leq f \leq i + w - 1$ **then**
                $y_{stage1} \leftarrow 1$, $general\_impact\_clips \leftarrow general\_impact\_clips + 1$; **break**
        ▷ Stage 2 label: head impact ONLY in central region (frames 10-20 of 30)
        $y_{stage2} \leftarrow 0$
        $central\_start \leftarrow i + 9$, $central\_end \leftarrow i + 19$ ▷ Frames 10-20 for focused attention
        for each $(f, \_, \_, \_, head\_flag) \in events$ **do**
            if $central\_start \leq f \leq central\_end$ **and** $head\_flag = 1$ **then**
                $y_{stage2} \leftarrow 1$, $head\_impact\_clips \leftarrow head\_impact\_clips + 1$; **break**
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{(clip, y_{stage1}, y_{stage2})\}$
▷ Final statistics: 93.7% non-event, 6.3% general impact, 0.5% head impact clips
Verify: $\frac{general\_impact\_clips}{total\_clips} \approx 0.063$, $\frac{head\_impact\_clips}{total\_clips} \approx 0.005$
return $\mathcal{C}$, dataset statistics

A.3 Model A: Player-Focused TSM Implementation

A.3.1 Player Detection and Tracking

Algorithm 3 YOLOv8 Player Detection with Domain Adaptation

Require: Frame I, fine-tuned YOLOv8 model for youth hockey
Ensure: Player detections $\mathcal{B} = \{(x, y, w, h, class, conf)\}$

$detections \leftarrow$ YOLOv8(I) ▷ 5 classes: Player_Light, Player_Dark, Goalie_Light, Goalie_Dark, Referee
Apply class-aware NMS with IoU threshold = 0.85 ▷ Prevent removal of close players from different teams
Filter detections with confidence > 0.5
return $\mathcal{B}$

Algorithm 4 IoU-Assisted StrongSORT Tracking (TSM + Tracking Configuration)

Require: Detection sequence $\{\mathcal{B}_t\}$, tracking parameters $\alpha = 0.6$
Ensure: Player trajectories $\mathcal{T} = \{track\_id : [(t, bbox)]\}$

Initialize StrongSORT tracker with appearance and motion models
for each frame t **do**
    $\mathcal{B}_t \leftarrow$ YOLOv8 detections (Alg. 3)
    for each existing track $track_i$ and detection det_j **do**
        $C_{appearance} \leftarrow$ ReID feature similarity$(track_i, det_j)$
        $C_{IoU} \leftarrow$ IoU$(track_i.bbox, det_j.bbox)$
        $C_{total} \leftarrow \alpha \cdot C_{appearance} + (1 - \alpha) \cdot C_{IoU}$ ▷ Balanced cost matrix
    Associate detections to tracks using Hungarian algorithm on C_{total}
    Update track states with Kalman filter
    Initialize new tracks for unmatched detections
Apply trajectory smoothing: $p_t^{smooth} = 0.3 \cdot p_t + 0.7 \cdot p_{t-1}^{smooth}$
return $\mathcal{T}$

Algorithm 5 Enhanced Tracking with Head Detection (TSM + Enhanced Tracking Configuration)

Require: Detection sequence $\{\mathcal{B}_t\}$, dedicated head detection model $\mathcal{H}$

Ensure: Enhanced trajectories with head-specific features $\mathcal{T}_{enhanced}$

Initialize StrongSORT tracker as in Alg. 4
for each frame t **do**
    $\mathcal{B}_t \leftarrow$ YOLOv8 detections
    for each player detection $bbox_{player}$ in $\mathcal{B}_t$ **do**
        $bbox_{head} \leftarrow \mathcal{H}(crop(I_t, bbox_{player}))$ ▷ Detect helmet within player region
        Augment detection: $(bbox_{player}, bbox_{head}, features_{player}, features_{head})$
    Perform tracking association using both player and head features
    Store head-specific trajectories for Stage 2 classification
return $\mathcal{T}_{enhanced}$ with head region tracking

A.3.2 Motion Feature Extraction

Algorithm 6 RAFT Optical Flow (Model A Stage 1 - TSM + Motion Configuration)

Require: Consecutive frames I_t, I_{t+1} from player crops

Ensure: Dense flow field F_t for motion augmentation

Pad frames to multiples of 64 pixels
$F_t \leftarrow \text{RAFT}(I_t, I_{t+1})$ with 12 refinement iterations
Post-process: clip magnitudes at 95th percentile, remove outliers
Compute motion descriptors: $\mathbf{f}_{motion} = [\text{median}(\Delta x), \text{median}(\Delta y), \text{median}(|\mathbf{v}|)]$
return F_t, $\mathbf{f}_{motion}$

Algorithm 7 Lucas-Kanade Optical Flow (Model A Stage 2 - TSM + Motion Lucas-Kanade)

Require: Player crop sequence $\{C_t\}$, helmet region (upper 30% of bbox)

Ensure: Motion descriptors focused on helmet region

Detect corner features using Shi-Tomasi (max 60 features in helmet region)
Track features across consecutive frames using Lucas-Kanade with 3-level pyramid
Compute displacement vectors $\{\Delta x, \Delta y, |\mathbf{v}|\}$ for helmet and full player regions
Aggregate: $\mathbf{f}_{helmet} = [\text{median}(\Delta x_{helmet}), \text{median}(\Delta y_{helmet}), \text{median}(|\mathbf{v}_{helmet}|)]$
Aggregate: $\mathbf{f}_{player} = [\text{median}(\Delta x_{player}), \text{median}(\Delta y_{player}), \text{median}(|\mathbf{v}_{player}|)]$
return concatenated motion descriptor $[\mathbf{f}_{helmet}, \mathbf{f}_{player}]$ (6-dim)

Algorithm 8 Farnebäck Optical Flow (Model A Stage 2 - TSM + Motion Farnebäck)

Require: Player crop sequence $\{C_t\}$

Ensure: Dense motion field via polynomial expansion

1: Apply Farnebäck algorithm: $f(\mathbf{x}) \approx \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$
2: Estimate displacement field from coefficient changes: $\mathbf{A}_{t+1} - \mathbf{A}_t, \mathbf{b}_{t+1} - \mathbf{b}_t$
3: Extract dense flow within helmet region (upper 30%) and full player region
4: Aggregate motion statistics from both regions as in Alg. 7
5: **return** aggregated motion descriptors (6-dim)

A.3.3 TSM Network Architecture and Feature Fusion

Algorithm 9 TSM ResNet-50 with Motion Feature Integration

Require: Player crop sequence $\{C_t\}_{t=1}^{30}$, optional motion features $\{M_t\}$ (3 or 6-dim)

Ensure: Classification probability p

1: Sample 8 temporal segments from 30 frames (segments = 30/8 4 frames per segment)
2: Apply temporal shift with division=8 to channels before each ResNet block:
3: **for** channel c in feature tensor where C = total channels **do**
4: **if** $c < C/4$ **then**
5: shift backward: $X_{t,c} \leftarrow X_{t-1,c}$
6: **else if** $C/4 \leq c < C/2$ **then**
7: shift forward: $X_{t,c} \leftarrow X_{t+1,c}$
8: **else**
9: keep current: $X_{t,c} \leftarrow X_{t,c}$
10: **if** motion features available **then**
11: ▷ Early fusion via spatial replication and channel concatenation
12: Spatially replicate motion descriptors: $M_{spatial} \leftarrow \text{repeat}(M_t, [H, W])$ to match RGB dimensions
13: Create augmented input: $X_{augmented} = [X_{RGB}, M_{spatial}]$ ▷ 3+3 or 3+6 channels
14: Modify first conv layer: $\text{Conv}_{first}(X_{augmented}) \rightarrow$ ResNet input dimensions
15: Apply 1×1 convolution for dimension reduction: $\text{Conv}_{1\times1}(X_{augmented}) \rightarrow$ 3 channels
16: Process through ResNet-50 backbone with TSM modules at each residual block
17: Global average pool: $\text{features}_{global} = \text{GAP}(\text{final\_conv\_features})$
18: Apply classification head: $p \leftarrow \sigma(W_{cls} \cdot \text{features}_{global} + b)$
19: **return** p

A.4 Model B: Full-Frame Multi-Modal Fusion

A.4.1 Input Modality Processing

Algorithm 10 RGB Stream Processing (Model B)

Require: 30-frame clip $\{I_t\}$, target resolution 224×224

Ensure: Normalized RGB tensor $(30, 3, 224, 224)$

1: Resize frames maintaining aspect ratio with zero-padding to 224×224
2: Normalize pixel values: $I_{norm} = I_{raw}/255.0$
3: Apply TSM temporal shifting (8 segments, shift division = 8)
4: Process through ResNet-50 + TSM backbone
5: Apply bidirectional GRU for temporal aggregation
6: **return** RGB embedding (1024-dim)

Algorithm 11 Dense Optical Flow Stream (Model B)

Require: 30-frame clip $\{I_t\}$

Ensure: Dense flow tensor $(29, 2, 224, 224)$ and flow embedding

1: Pad frames to multiples of 64 pixels for RAFT processing
2: **for** $t = 1$ to 29 **do**
3: $F_t \leftarrow \text{RAFT}(I_t, I_{t+1})$ with 12 refinement iterations
4: Post-process: clip magnitudes at 95th percentile, remove outliers
5: Modify ResNet-50 first conv layer: $\text{Conv2d}(2, 64, \text{kernel} = 7)$ for 2-channel flow input
6: Apply TSM temporal shifting (8 segments, shift division = 8)
7: Process through modified ResNet-50 + TSM backbone
8: Apply bidirectional GRU for temporal aggregation
9: **return** Flow embedding (1024-dim)

Algorithm 12 Pose Keypoint Stream (Model B)

Require: 30-frame clip $\{I_t\}$

Ensure: Pose tensor $(30, 17, 2)$ and pose embedding

for each frame I_t **do**
  $skeletons_t \leftarrow$ Keypoint R-CNN(I_t) ▷ 17 COCO keypoints per detected person
  Rank skeletons by composite score with $\alpha_{pose} = 0.7$, $\sigma = 112$:
    $S_i = \alpha_{pose} \cdot \frac{1}{17} \sum_{j=1}^{17} c_{i,j} + (1 - \alpha_{pose}) \cdot \exp(-d_i^2/\sigma^2)$
  Select top 10 ranked skeletons based on confidence and proximity
  $pose_t \leftarrow$ weighted average of selected skeletons: $\frac{\sum_{i=1}^{10} S_i \cdot skeleton_i}{\sum_{i=1}^{10} S_i}$
Apply Transformer encoder (8 attention heads, 2 layers) for spatial keypoint relationships:
  Multi-head attention: $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}$
Apply bidirectional GRU for temporal aggregation across 30 frames
return pose embedding (512-dim)

A.4.2 Multi-Modal Fusion Architecture

Algorithm 13 Multi-Modal Fusion with Squeeze-and-Excitation Attention

Require: RGB embedding (1024-dim), Flow embedding (1024-dim), Pose embedding (512-dim)

Ensure: Fused representation for classification

▷ Dimension alignment via projection
Project to common dimension: $e_{RGB}^{proj} = W_{RGB} \cdot e_{RGB} + b_{RGB}$ (1024→512)
Project to common dimension: $e_{flow}^{proj} = W_{flow} \cdot e_{flow} + b_{flow}$ (1024→512)
Keep pose embedding: $e_{pose}^{proj} = e_{pose}$ (already 512-dim)
▷ Late fusion via concatenation
Concatenate modalities: $e_{concat} = [e_{RGB}^{proj}, e_{flow}^{proj}, e_{pose}^{proj}]$ (1536-dim total)
▷ Squeeze-and-Excitation attention mechanism
Global average pooling: $e_{squeezed} = \text{GAP}(e_{concat})$
First FC layer: $e_{excited1} = \text{ReLU}(W_1 \cdot e_{squeezed} + b_1)$ (1536→384)
Second FC layer: $e_{excited2} = \sigma(W_2 \cdot e_{excited1} + b_2)$ (384→1536)
Apply attention weights: $e_{final} = e_{excited2} \odot e_{concat}$ (element-wise multiplication)
Apply final classification head: $p = \sigma(W_{cls} \cdot e_{final} + b_{cls})$
return p

A.4.3 I3D Architecture Comparison

Algorithm 14 I3D ResNet-50 Implementation (Model B Comparison)

Require: 30-frame clip $\{I_t\}$, modality streams (RGB, Flow, Pose)
Ensure: I3D-based classification probability

▷ Inflate 2D ResNet-50 to 3D using temporal kernel size T=3
for each 2D convolutional layer with kernel (k_h, k_w) **do**
    Inflate to 3D: $(k_t, k_h, k_w) = (3, k_h, k_w)$
    Initialize 3D weights: $\mathbf{W}_{3D}(i, j, k) = \frac{1}{T}\mathbf{W}_{2D}(i, j)$ where $T = 3$
▷ Process each modality stream
$e_{RGB} \leftarrow$ I3D-ResNet-50(RGB_frames) ▷ Direct 3D convolution processing
if Flow stream enabled **then**
    Modify first layer for 2-channel input: Conv3d$(2, 64, \text{kernel} = (3, 7, 7))$
    $e_{flow} \leftarrow$ I3D-ResNet-50($Flow\_frames$)
if Pose stream enabled **then**
    Convert pose keypoints to heatmap representation $(30, 17, 224, 224)$
    $e_{pose} \leftarrow$ I3D-ResNet-50($Pose\_heatmaps$)
Apply fusion mechanism (same as TSM version - Alg. 13)
return classification probability

A.5 Training Procedures

Algorithm 15 Stage 1 Training: General Impact Detection

Require: Training clips $\mathcal{D}_{train}$ with labels $y \in \{0, 1\}$ (6.3% positive class)

Ensure: Trained Stage 1 model Θ_1 and optimal threshold τ_1^*

Initialize model: TSM ResNet-50 (Model A) or Multi-modal fusion (Model B)

Load Kinetics-400 pre-trained weights for backbone initialization

▷ Configure loss function for class imbalance

Calculate class weights: $w_{pos} = \frac{1}{0.063} \approx 15.87$, $w_{neg} = \frac{1}{0.937} \approx 1.07$

Setup weighted BCE loss: $\mathcal{L}_{BCE}(y, \hat{y}) = -[w_{pos} \cdot y \log(\hat{y}) + w_{neg} \cdot (1 - y) \log(1 - \hat{y})]$

Setup AdamW optimizer: $lr_{max} = 5 \times 10^{-5}$, $lr_{min} = 1 \times 10^{-7}$, weight decay $= 1 \times 10^{-5}$

Apply cosine annealing: $\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{t\pi}{T_{max}}))$

for epoch $e = 1$ to max epochs **do**

    for each mini-batch $\{(x_i, y_i)\}$ of size 16 **do**

        Apply weighted random sampling: $P(x_i) = \frac{w_{y_i}}{\sum_j w_{y_j}}$ for class balance

        Forward pass: $\hat{y}_i = f_{\Theta_1}(x_i)$

        Compute loss: $\mathcal{L}_{batch} = \frac{1}{N} \sum_i \mathcal{L}_{BCE}(y_i, \hat{y}_i)$

        Backward pass with gradient clipping: $\|\nabla\| \leq 1.0$

        Update parameters: $\Theta_1 \leftarrow \Theta_1 - \eta_t \cdot \nabla_{\Theta_1} \mathcal{L}_{batch}$

    Validate on held-out games, compute F1-score

    Early stopping if no improvement for 10 epochs

▷ Optimize classification threshold on validation set

$\tau_1^* = \arg\max_\tau F1(\tau) = \arg\max_\tau \frac{2 \cdot Precision(\tau) \cdot Recall(\tau)}{Precision(\tau) + Recall(\tau)}$

return Θ_1, τ_1^*

Algorithm 16 Stage 2 Training: Head Impact Classification

Require: Stage 1 positive clips with head impact labels (0.5% of all clips are head impacts)

Ensure: Trained Stage 2 model Θ_2 and optimal threshold τ_2^*

Initialize player-focused TSM model (Model A architecture only)
Load Kinetics-400 pre-trained ResNet-50 backbone
▷ Configure Focal Loss for extreme class imbalance
Setup Focal Loss: $\mathcal{L}_{Focal}(y, \hat{y}) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$
Set parameters: $\alpha_t = 0.25$, $\gamma = 2.0$ where $p_t = \hat{y}$ if $y = 1$, else $1 - \hat{y}$
▷ Preprocess all training clips through tracking pipeline
for each training clip with general impact **do**
    Apply YOLOv8 + StrongSORT tracking:
    if Enhanced tracking configuration **then**
        Use Alg. 5 with head detection
    else
        Use Alg. 4 for standard tracking
    Extract player crops, resize to 224×224 with aspect ratio preservation
    Apply trajectory smoothing: $p_t^{smooth} = 0.3 \cdot p_t + 0.7 \cdot p_{t-1}^{smooth}$
    Interpolate short gaps 15 frames using cubic splines
    if Motion features enabled **then**
        Extract Lucas-Kanade or Farnebäck features (Alg. 7 or 8)
for epoch $e = 1$ to max epochs **do**
    for each mini-batch of player crops **do**
        Forward pass through TSM network (Alg. 9)
        Compute Focal Loss: $\mathcal{L}_{batch} = \frac{1}{N} \sum_i \mathcal{L}_{Focal}(y_i, \hat{y}_i)$
        Update parameters with AdamW optimizer
    Validate on held-out games, optimize for head-class F1-score
    Early stopping if no head-class F1 improvement for 10 epochs
Optimize threshold: $\tau_2^* = \arg\max_\tau F1_{head}(\tau)$ on validation set
return Θ_2, τ_2^*

A.6 Evaluation Protocol

Algorithm 17 Game-Based Dataset Partitioning (45 Games Total)

Require: 45 annotated games across age groups (U11, U13, U16, U18) and levels (A, AA, AAA)

Ensure: Train/validation/test splits with complete game-level isolation

1: Randomly assign complete games to splits to prevent data leakage
2: Typical split: 70% train (31 games), 15% validation (7 games), 15% test (7 games)
3: Extract clips only from games assigned to each respective split
4: Verify temporal isolation: no clips from test games appear in train/validation sets
5: Maintain age/level distribution across splits for representative sampling
6: **return** game-based partitions with complete isolation

Algorithm 18 End-to-End Two-Stage Pipeline Evaluation

Require: Test videos, trained models Θ_1, Θ_2, optimized thresholds τ_1^*, τ_2^*

Ensure: Performance metrics with temporal NMS post-processing

$predictions \leftarrow \emptyset$

for each test clip x_i with ground truth label y_i **do**

▷ Stage 1: General Impact Detection

  $p_1 = f_{\Theta_1}(x_i)$ ▷ Stage 1 general impact probability

  if $p_1 \geq \tau_1^*$ **then**

    ▷ Stage 2: Head Impact Classification

    Extract player crops from x_i using tracking pipeline

    $p_2 = g_{\Theta_2}(crops_i)$ ▷ Stage 2 head impact probability

    Final prediction: $\hat{y}_i = 1$ if $p_2 \geq \tau_2^*$, else $\hat{y}_i = 0$

    Store: $(clip_i, timestamp_i, p_1, p_2, \hat{y}_i)$

  else

    $\hat{y}_i = 0$, $p_2 = 0$ ▷ No general impact detected

    Store: $(clip_i, timestamp_i, p_1, 0, 0)$

  $predictions \leftarrow predictions \cup (\hat{y}_i, y_i, p_1, p_2)$

▷ Apply temporal Non-Maximum Suppression

Group overlapping positive detections within temporal windows

Keep highest-confidence detection per temporal cluster

Compute final confusion matrix: TP, FP, FN, TN

Calculate metrics: $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$

return Performance metrics, confusion matrix, per-game statistics

A.7 Final Hyperparameters

Component	Parameter	Stage 1	Stage 2
Dataset Configuration			
Total games		45 (U11, U13, U16, U18)	
Window size	frames	30	30
Stride	frames	5	N/A
Overlap	percentage	83.3%	
Input resolution	pixels	224×224	224×224
Frame rate	fps	30	30
Class Distribution			
General impact clips	percentage	6.3%	N/A
Head impact clips	percentage	0.5%	0.5% (of all clips)
Head impact ratio		N/A	8.5% (of general impacts)
Network Architecture			
Backbone		ResNet-50 + TSM	ResNet-50 + TSM
Temporal segments		8	8
Shift division		8	8
Pre-training dataset		Kinetics-400	Kinetics-400
Output dimensions		1024 (RGB/Flow), 512 (Pose)	1024 (single stream)
Training Configuration			
Learning rate (initial)		5×10^{-5}	5×10^{-5}
Learning rate (minimum)		1×10^{-7}	1×10^{-7}
Schedule		Cosine annealing	Cosine annealing
Batch size	clips	16	16
Weight decay		1×10^{-5}	1×10^{-5}
Gradient clipping	max norm	1.0	1.0
Early stopping patience	epochs	10	10
Loss Function			
Loss type		Weighted BCE	Focal Loss
Positive class weight		15.87	N/A
Negative class weight		1.07	N/A
Focal α		N/A	0.25
Focal γ		N/A	2.0
Player Tracking (Model A)			
YOLOv8 classes		5 (Player_Light/Dark, Goalie_Light/Dark, Referee)	
IoU-appearance balance	α	0.6	0.6
Class-aware NMS threshold		0.85	0.85
Detection confidence		0.5	0.5
Trajectory smoothing	α_{smooth}	0.3	0.3
Gap interpolation limit	frames	15	15
Optical Flow Processing			
RAFT refinement iterations		12	N/A
Frame padding	pixels	Multiple of 64	N/A
Lucas-Kanade pyramid levels		N/A	3
Max corner features (helmet)		N/A	60
Helmet region		N/A	Upper 30% of bbox
Pose Processing (Model B)			
Keypoint standard		17 (COCO format)	N/A
Top skeletons per frame		10	N/A
Transformer attention heads		8	N/A
Transformer encoder layers		2	N/A
Confidence weight α_{pose}		0.7	N/A
Distance weight σ		112	N/A
Multi-Modal Fusion (Model B)			
RGB/Flow embedding dim		$1024 \rightarrow 512$	N/A
Pose embedding dim		512	N/A
Fused representation		$1536 \rightarrow$ final	N/A
SE reduction ratio		4:1 ($1536 \rightarrow 384 \rightarrow 1536$)	N/A

References

[1] Associated Press (2013). Mls commissioner don garber says league won't adopt goal-line technology by 2014. AP via MLSSoccer: GoalControl $260,000 per stadium + $3,900 per match.

[2] Bacigalupe, C. (2019). The video assistant referee (var) protocol. In *The use of video technologies in refereeing football and other sports*, pages 183–207. Routledge.

[3] Bai, S., Zhang, Z., Wang, H., and Mei, T. (2020). Temporal interlacing network. In *Computer Vision – ECCV 2020*, pages 512–528.

[4] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

[5] Bertasius, G., Wang, H., and Torresani, L. (2021). Is space–time attention all you need for video understanding? *Proceedings of the 38th International Conference on Machine Learning*, 139:813–824.

[6] Bevir, G. (2023). Inside the production of the 2023 iihf ice hockey world championship. https://www.svgeurope.org/blog/headlines/inside-the-production-of-the-2023-iihf-ice-hockey-world-championship/. SVG Europe, 26 May 2023. Includes the "IIHF 2023 Camera Plan" graphic by Infront Productions. Accessed 30 July 2025.

[7] Camarillo, D. B., Shull, P. B., Mattson, J., Shultz, R., and Garza, D. (2013). An instrumented mouthguard for measuring linear and angular head-impact kinematics in american football. *Annals of Biomedical Engineering*, 41(9):1939–1949.

[8] Carey, L. J., Terry, D. P., McIntosh, A. S., Stanwell, P., Iverson, G. L., and Gardner, A. J. (2021). Video analysis and verification of direct head impacts recorded by wearable sensors in junior rugby league players. *Sports Medicine – Open*, 7:66.

[9] Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

[10] Centers for Disease Control and Prevention (2024). HEADS UP: Data on Sports- and Recreation-Related Concussions. https://www.cdc.gov/heads-up/data/index.html. Accessed 26 May 2025.

[11] Chambers, R. M., Gabbett, T. J., Cole, M. H., and Beard, A. (2019a). Automatic detection of one-on-one tackles and ruck events using microtechnology in rugby union. *Journal of Science and Medicine in Sport*, 22(9):1043–1048.

[12] Chambers, R. M., Gabbett, T. J., Cole, M. H., Gupta, R., Josman, C., Bown, R., Stridgeon, P., and Beard, A. (2019b). Automatic detection of one-on-one tackles and ruck events using microtechnology in rugby union. *Journal of Science and Medicine in Sport*, 22(7):827–832.

[13] Chaurasia, A., Jocher, G., Qiu, J., and Stoken, J. (2023). Yolov8: Ultralytics real-time object detector. https://github.com/ultralytics/ultralytics. Accessed 26 May 2025.

[14] Chen, W., Post, A., Karton, C., Gilchrist, M. D., Robidoux, M., and Hoshizaki, T. B. (2023a). A comparison of frequency and magnitude of head impacts between pee wee and bantam youth ice hockey. *Sports Biomechanics*, 22(6):728–751.

[15] Chen, W., Post, A., Karton, C., Gilchrist, M. D., Robidoux, M., and Hoshizaki, T. B. (2023b). A comparison of frequency and magnitude of head impacts between pee wee and bantam youth ice hockey players. *Sports Biomechanics*, 22(6):728–751.

[16] Ciaparrone, G., Sánchez, F. L., Tabik, S., Troiano, L., Tagliaferri, R., and Herrera, F. (2020). Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88.

[17] Dawson, J., Myer, G. D., Davis, M. P., Group, S. S. S., et al. (2019). Engaging athletic trainers in concussion detection: overview of the national football league atc spotter program, 2011–2017. *Journal of Athletic Training*, 54(9):939–949.

[18] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634.

[19] Du, P., Turan, M., and Yang, Y. (2022). Strongsort: Make deepsort great again. https://arxiv.org/abs/2202.13514.

[20] Duan, H., Mei, W., and Li, W. t. (2022). Revisiting skeleton-based action recognition. In *CVPR*, pages 270–280.

[21] Elbin, R. J., Sufrinko, A., Anderson, M. N., Mohler, S., Schatz, P., Covassin, T., Stolz, S., and Kontos, A. P. (2020). Sensitivity and specificity of on-field visible signs of concussion in the national football league. *Neurosurgery*, 87(3):530–537.

[22] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Image Analysis (SCIA)*, pages 363–370.

[23] Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941.

[24] Gibson, O. (2013). Fifa snubs hawk-eye in favour of german goalline technology goalcontrol. *The Guardian*.

[25] Handa, A., Newcombe, R. A., Angeli, A., and Davison, A. J. (2012). Real-time camera tracking: when is high frame-rate best? In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*, pages 222–235. Springer.

[26] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[27] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[28] Houshmand Sarkhoosh, M., Gautam, S., Midoglu, C., Shafiee Sabet, S., Kupka, T., and Halvorsen, P. (2025). Hockeyrink: A dataset for precise ice hockey rink keypoint mapping and analytics. In *Proceedings of the 16th ACM Multimedia Systems Conference*, pages 249–255.

[29] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *CVPR*.

[30] Kassab, E. J., Solberg, H. M., Gautam, S., Sabet, S. S., Torjusen, T., Riegler, M., Halvorsen, P., and Midoglu, C. (2024). TACDEC: Dataset of tackle events in soccer game videos. In *Proceedings of the 15th ACM Multimedia Systems Conference (MMSys '24)*, pages 250–256.

[31] Kay, W., Carreira, J., Simonyan, K., et al. (2017). The kinetics human action video dataset. *arXiv:1705.06950*.

[32] Kuo, C., Wu, L., Loza, J., Senif, D., Anderson, S. C., and Camarillo, D. B. (2018). Comparison of video-based and sensor-based head impact exposure. *PloS one*, 13(6):177–188.

[33] Laptev, I. (2005). On space–time interest points. *International Journal of Computer Vision*, 64(2-3):107–123.

[34] Lin, J., Gan, C., and Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093.

[35] Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2999–3007.

[36] Lin, T., Zhao, X., Su, H., Wang, C., and Yang, M. (2018). Bsn: Boundary sensitive network for temporal action proposal generation. In *Computer Vision – ECCV 2018: 15th European Conference on Computer Vision, Munich, Germany, September 8–14, 2018, Proceedings, Part IV*, volume 11219 of *Lecture Notes in Computer Science*, pages 3–19. Springer International Publishing.

[37] Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

[38] Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *arXiv:1711.05101*. Updated version commonly cited as AdamW.

[39] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI*.

[40] Mack, C., Myers, E., Barnes, R., Solomon, G., and Sills, A. (2019). Engaging athletic trainers in concussion detection: Overview of the national football league atc spotter program, 2011–2017. *Journal of Athletic Training*, 54(8):852–857.

[41] McCrea, M., Hammeke, T., Olsen, G., Leo, P., and Guskiewicz, K. (2004). Unreported concussion in high school football players: implications for prevention. *Clinical Journal of Sport Medicine*, 14(1):13–17.

[42] McCrory, P., Meeuwisse, W., Dvorak, J., et al. (2017). Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in berlin, october 2016. *British Journal of Sports Medicine*, 51(11):838–847.

[43] McKay, C. D., Cumming, S. P., and Blake, T. (2019). Youth sport: friend or foe? *Best Practice & Research Clinical Rheumatology*, 33(1):141–157.

[44] National Football League (2021a). Nfl 1st & future – impact detection challenge. https://www.kaggle.com/competitions/nfl-impact-detection.

[45] National Football League (2021b). NFL 1st & Future Impact Detection Challenge Dataset. https://www.kaggle.com/c/nfl-impact-detection.

[46] National Football League & Kaggle (2021). NFL 1st & Future Impact Detection Challenge Dataset. https://www.kaggle.com/competitions/nfl-impact-detection. Accessed 26 May 2025.

[47] Nievas, E. B., Tabik, S., Suárez, S., and Rivas, J. F. (2011). Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns*, volume 6854 of *Lecture Notes in Computer Science*, pages 332–339.

[48] NVIDIA Corporation (2022). Nvidia announces availability of jetson agx orin developer kit to advance robotics and edge ai. MSRP $1,999.

[49] NVIDIA Corporation (2025). Deepstream documentation — overview. DeepStream ingests RTSP/USB/CSI and supports multi-stream analytics.

[50] O'Connor, K. L., Rowson, S., Duma, S. M., and Broglio, S. P. (2017). Head-impact-measurement devices: A systematic review. *Journal of Athletic Training*, 52(3):206–227.

[51] Patel, B. H., Okoroha, K. R., Jildeh, T. R., Nwachukwu, B. U., and Forsythe, B. (2019). Concussions in the national basketball association: Analysis of incidence, return to play, and performance from 1999 to 2018. *Orthopaedic Journal of Sports Medicine*, 7(6):2325967119854199.

[52] Perrett, T., Sinha, S., Burghardt, T., Mirmehdi, M., and Damen, D. (2023). Use your head: Improving long-tail video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2415–2425.

[53] Pulli, K., Baksheev, A., Kornyakov, K., and Eruhimov, V. (2012). Realtime computer vision with opencv. *Communications of the ACM*, 55(6):61–69.

[54] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

[55] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99.

[56] Rezaei, M. and Wu, L. (2022). Automated soccer head impact exposure tracking using video and deep learning. *Scientific Reports*, 12:10421.

[57] Rowson, S., Duma, S. M., Beckwith, J. G., Chu, J. J., Greenwald, R. M., Crisco, J. J., Brolinson, P. G., Duhaime, A.-C., McAllister, T. W., and Maerlender, A. C. (2018). Head-impact–measurement devices: a systematic review. *Journal of Athletic Training*, 53(4):351–360.

[58] Sadlier, D. A. and O'Connor, N. E. (2005). Event detection in field sports video using audio–visual features and a support vector machine. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1225–1233.

[59] Scottish Professional Football League (2022). Var approved by spfl clubs. States £1.2m per season to operate VAR; club cost breakdown provided.

[60] Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576.

[61] Sozykin, K., Protasov, S., Khan, A., Hussain, R., and Lee, J. (2018). Multi-label class-imbalanced action recognition in hockey videos via 3d convolutional neural networks. In *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 146–151. IEEE.

[62] Swenson, A. G., Pritchard, N. S., Miller, L. E., Urban, J. E., and Stitzel, J. D. (2023). Characterization of head impact exposure in boys' youth ice hockey. *Research in Sports Medicine*, 31(4):440–450.

[63] Teed, Z. and Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer.

[64] Townsend, D. C., Saker, A., Scandrett, K., Green, M., Brownlow, M., Riley, P., Gillett, M., and Belli, A. (2025). Role of video review for sport-related concussion identification: a systematic review. *British Journal of Sports Medicine*, pages bjsports–2024–109603. Online ahead of print.

[65] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497.

[66] Wallace, J., Covassin, T., Nogle, S., Gould, D., and Kovan, J. (2017a). Knowledge of concussion and reporting behaviors in high school athletes. *Journal of Athletic Training*, 52(3):228–235.

[67] Wallace, J., Covassin, T., Nogle, S., Gould, D., and Kovan, J. (2017b). Knowledge of concussion and reporting behaviors in high school athletes. *Journal of Athletic Training*, 52(3):228–235.

[68] Wilcox, B. J., Machan, J. T., Beckwith, J. G., Greenwald, R. M., and Crisco, J. J. (2014). Head-impact mechanisms in men's and women's collegiate ice hockey. *Journal of Athletic Training*, 49(4):514–520.

[69] Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649.

[70] World Rugby (2023). World rugby integrates smart mouthguard technology into the head injury assessment. https://www.world.rugby/news/875212. Press release, 9 Oct 2023.

[71] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2.

[72] Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7444–7452.

[73] Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y., and Davis, L. (2021). VideoLT: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10021.