Ice Hockey Action Recognition via Contextual Priors

Kseniia Buzko[1 [0009−0009−9930−8497]], Amir Nazemi[1 [0000−0002−8405−473X]], David A. Clausi[1 [0000−0002−6383−0875]], and Yuhao Chen[1 [0000−0001−6094−0545]]

Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
{kbuzko,amir.nazemi,dclausi,yuhao.chen1@}@uwaterloo.ca

Abstract. Skeleton-based action recognition models, which are developed for generic human-pose data, struggle with ice-hockey broadcasts player action recognition, where the players appear smaller, move abruptly, and wield sticks that are invisible to standard skeleton models. To address these issues, we propose CP-Hockey, a context-aware pipeline that incorporates two domain-specific priors. First, a temporal player's bounding-box normalization stabilizes player scale across the player tracklet, raising top-1 accuracy from 31 % to 57 % on a six-class NHL dataset. Second, we design hockey-specific skeletons that include stick end-points and optional detailed head landmarks. A 15-keypoint body-plus-stick model improves the accuracy to 64 %, while our full 20-keypoint configuration reaches 65 %. Experimental results with STGCN++ and 2s-AGCN show that both contextual priors are necessary: scale normalization reduces spatial jitter, and stick keypoints disambiguate visually similar movements such as stickwork versus striking a puck with a stick. CP-Hockey establishes a strong baseline for fine-grained ice-hockey analytics and provides a blueprint for adapting skeleton pipelines to other equipment-centric sports.

Keywords: Ice Hockey Action Recognition · Skeleton-based Action Recognition · Contextual Priors

1 Introduction

Player action recognition is a fundamental task in sports analytics, enabling the automatic identification and classification of particular player movements during the game, such as skating, striking a puck, or maintaining position. Understanding these actions allows deeper insights into player strategies and game analysis. Although sports such as basketball [2, 5] and soccer [3, 18] have seen extensive development in player action recognition methodologies due to their popularity, ice hockey remains comparatively underexplored. Traditional approaches like the Action Recognition Hourglass Network (ARHN) [1] rely mainly on static pose input, ignoring temporal and ice hockey contextual information. By 'contextual information' of broadcast ice hockey videos, we point to structured domain-specific cues such as camera-driven scale changes and stick pose that go beyond

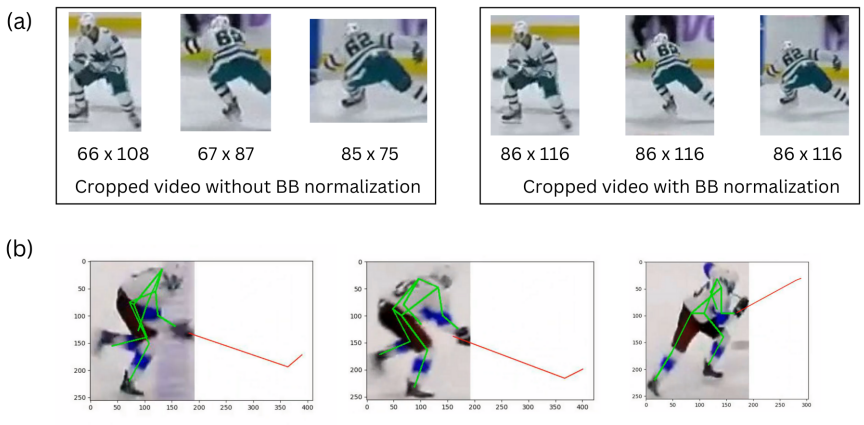

Fig. 1. Illustration of the two contextual priors exploited by CP-Hockey. (a) Temporal bounding-box (BB) normalization suppresses scale jitter by resizing every frame in the player tracklet to the maximum BB dimensions. (b) A 15-keypoint ice hockey skeleton (green body joints, red stick joints) with the stick pose which is vital for player action recognition and not included in the generic 17-keypoint COCO skeleton.

the joint coordinates; the two contextual priors that we leverage in this work are visualized in Figure 1.

Player action recognition in broadcast ice hockey footage is challenging for several reasons. The players bounding boxes are small and fast-moving (Figure 2); rapid camera pans create scale jitter; bulky protective gear and frequent occlusions confound generic pose estimators.

Most pose-based solutions for action recognition ignore the temporal contextual features of the data domain [22, 7] and rely on generic 17-keypoint human skeletons (such as keypoint COCO model [14]). However, static cropping and resizing of a player's bounding box on a frame-by-frame basis disregards the temporal context, often resulting in irregular player scales and spatial jitter across consecutive frames. This jitteriness of the players' poses decreases the performance of downstream models, as the player's pose rapidly changes in size or position due to camera effects rather than actual movement of the player's poses. Although generic skeletal configurations are widely used in human pose estimation [14], they are not tailored for the ice hockey domain. Many of the keypoints in these models correspond to body joints that can be easily occluded by hockey gear like helmets or pads. In addition, they do not consider key objects such as the hockey stick in their solution. From the perspective of contextual priors in ice hockey, we argue that the hockey stick serves as a vital cue to recognize actions in the sport. It can differentiate various actions, such as stickwork, where the stick primarily makes contact with the ice, and striking the puck, where the stick is in a winding-up motion.

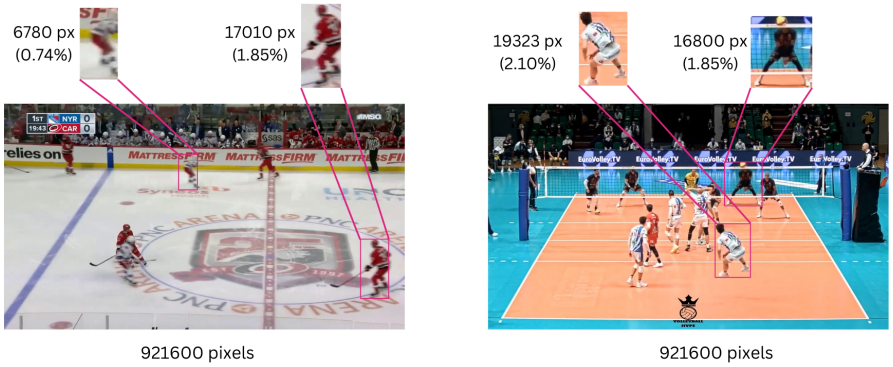

Fig. 2. Comparison of bounding-box pixel sizes between ice hockey (left) and volleyball (right) broadcast footage [4]. In addition to the rapid movement of players and the camera, the bounding boxes of players are relatively smaller in ice hockey broadcast videos, which makes player action recognition more complicated.

In this work, we propose CP-Hockey, a novel ice hockey action recognition pipeline that leverages contextual priors to address the mentioned challenges. Our CP-Hockey pipeline benefits from contextual priors at two points (Figure 1). First, a context-sensitive temporal normalization stabilizes the bounding box of each player's tracklet throughout the action clip, suppressing camera-driven scale variability. Second, our solution benefits from skeleton with stick end-points, giving the action recognition model explicit access to stick pose and motion. Through extensive experiments on an NHL hockey video dataset, we demonstrate that integrating such contextual priors markedly improves player action recognition performance in ice hockey broadcast videos.

This paper makes the following unique and novel contributions to player action recognition in ice hockey broadcast videos:

- We implement a temporal bounding box normalization method that reduces spatial jitter of a player's tracklet in broadcast footage and improves the skeleton-based player action recognition in ice hockey broadcast videos.
- We implement a novel skeletal configuration specifically tailored to the ice hockey domain, which integrates stick keypoints to capture important stick-related interactions, an aspect that to the best of our knowledge, has not been previously explored.
- We perform extensive evaluations comparing graph convolutional networks (GCNs) such as 2s-AGCN [9] and STGCN++ [6], generating empirical evidence advocating for the use of STGCN++ with contextual priors in recognizing ice hockey-specific actions.

2 Background Research

2.1 Pose Estimation

Human pose estimation models identify key anatomical points, such as joints and limbs, from images or videos, providing a structured representation of human posture. Pose estimation is particularly relevant in sports analytics, as it enhances the ability to capture precise body configurations necessary for effectively classifying specific movements. In ice hockey, accurate pose estimation can significantly improve action recognition by providing robust temporal skeletal representations of players' complex movements.

However, standard human pose estimation methods, typically trained on general datasets such as COCO [14] or MPII [15], face unique difficulties when applied to ice hockey broadcasts. Players wear bulky protective equipment that alters the shape of the body, leading to frequent misinterpretation of limb positions. In addition, uniform colors, such as white jerseys, visually blend with ice or boards, complicating limb identification. Pose estimation in ice hockey is further challenged by rapid and agile player movements, such as quick skating transitions or sudden turns, which often result in severe motion blur and unusual body poses [1, 10].

Recent work has begun to address these challenges associated with ice hockey-specific poses. Balaji et al. [11] introduced a multimodal approach that utilizes language cues to manage occluded keypoints, focusing on the player's body and using textual prompts for expected stick positions. This method significantly improved pose accuracy in an ice-hockey dataset by leveraging domain context, such as equipment knowledge. We incorporate pose estimation [10, 11] into our CP-Hockey pipeline, as accurate skeletal representations are essential for reliable ice hockey action recognition, especially for subtle actions like puck striking and stick handling.

2.2 Skeleton-Based Action Recognition

Building on pose estimation, researchers have explored skeleton-based action recognition for sports, where player joint sequences are input to an action classifier. Graph convolutional networks (GCNs) have become a dominant paradigm for this problem [22], as they naturally model the human skeleton as a graph, where each vertex represents a body joint and edges are split into temporal and spatial connections. The edges that join vertices inside a frame are referred to as spatial edges, while the edges that join the same vertex across consecutive frames are referred to as temporal edges. This representation significantly outperformed convolutional neural networks (CNNs) and recurrent neural networks (RNNs) based approaches [16, 17] that employed the skeleton modality without taking joint dependence into account.

ST-GCN [8] demonstrated the effectiveness of spatio-temporal keypoint graphs for action recognition, outperforming purely appearance-driven approaches. Subsequent improvements, such as 2s-AGCN [9] and STGCN++ [6], introduced

adaptive graph structures and multi-stream inputs, such as joint and bone data, for improved performance. These models have proven particularly effective in sports analytics [23], where structured skeletal data remains reliable despite variations in clothing, lighting, and background clutter.

Despite advancements, action recognition in ice hockey remains underexplored compared to sports with a larger following, such as basketball or soccer. Prior research has focused mainly on coarse tasks such as player tracking [19] and puck tracking [20], often relying on traditional vision pipelines or CNN-based trackers. For particular action recognition, such as forward skating, stickwork, and rapid deceleration, remains underexplored.

2.3 Contextual and Object-Aware Action Recognition

Early skeleton pipelines inferred actions almost exclusively from joint coordinates, treating each frame or short clip in isolation. Recent work shows that contextual priors like temporal neighborhoods, interacting objects, and high-level semantics, can greatly sharpen recognition, especially when skeleton trajectories alone are ambiguous. For example, Cioppa et al. design CALF [12], a context-aware loss that spots soccer events by supervising not the single annotated frame but a short temporal window around it. While Wen et al. add dynamically detected object centers to the ST-VGCN graph [7], demonstrating that even coarse object localization can disambiguate actions sharing near-identical limb motions.

3 Dataset

To develop and evaluate our CP-Hockey pipeline, we collected video clips from National Hockey League (NHL) broadcasts captured at 30 frames per second (fps). These clips were sourced from multiple NHL games, covering 29 different teams, which provides a wide range of arenas, lighting conditions, and team uniforms for a diverse and challenging dataset. The raw footage was segmented into individual shots, where each clip lasts from a few seconds to over a minute.

Six ice-hockey-specific classes were selected for their frequency, tactical importance, and visual distinctness. Table 1 summarizes the class definitions and their distribution. In total, the dataset contains 1,547 annotated action instances, each spanning 2 seconds (60 frames), where the start and end frames are defined as ± 30 frames around the anchor action frame.

4 Methodology

4.1 Pipeline Overview

The CP-Hockey pipeline for ice-hockey action recognition comprises four stages that together inject two complementary contextual priors: a temporal prior that stabilizes player scale and position, and an object prior that models the hockey stick explicitly. Figure 3 gives a high-level view of the proposed pipeline.

Table 1. Action classes with descriptions and frequency.

Action Class	Description	Occurrences
Rapid Deceleration	Player abruptly reduces skating speed to change direction or avoid contact	52
Backwards Skating	Player skates in reverse, typically to maintain defensive positioning	116
Maintaining Position	Player remains largely stationary while reading the play or screening the goalie	164
Strike Puck with Stick	Player's stick makes direct contact with the puck for a pass, shot, or block	337
Forwards Skating	Player accelerates toward the offensive zone while controlling balance and speed	421
Stickwork	Player performs intricate stickhandling to retain or regain puck control	457
Total		**1547**

Each annotated action segment is first preprocessed by automatically extracting a player tracklet with the VIP-HTD tracker [19]. Bounding boxes are then temporally normalized to enforce scale consistency (Section 4.3, Contextual Prior #1). Next, an ice-hockey-specific pose estimator [10, 11] converts every normalized crop to a skeleton, yielding 2-D key-point sequences with six alternative configurations (Section 4.4, Contextual Prior #2). Finally, a GCN-based model classifies the action (Section 4.5).

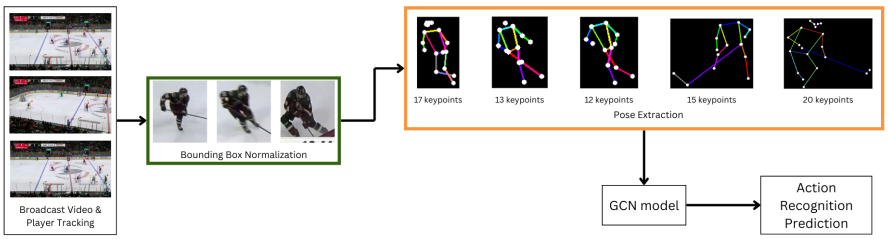

Fig. 3. Overview of the CP-Hockey pipeline. Contextual Prior #1 (green) stabilizes bounding-box scale; Prior #2 (orange) augments the skeleton with stick key-points. The enriched graph is fed to GCN-based model for action recognition.

4.2 Preprocessing

To automate bounding box extraction and ensure temporal consistency across frames, we leverage a tracking approach [19], specifically tailored for player detection in ice hockey broadcast footage. After extracting bounding boxes, frame-wise bounding box cropping and resizing to a normalized spatial representation

is performed. These cropped bounding boxes are the inputs for pose extraction, which uses ice hockey-optimized models presented in [10, 11].

4.3 Bounding Box Normalization

A significant challenge in using broadcast video for pose estimation is the variability in player size and position across frames. A naive approach might resize each frame independently, stretching the player's skeleton and distorting the body shape, which negatively impacts downstream performance. To mitigate this, we propose a robust bounding box normalization strategy. Specifically, we compute the maximum width and height of the bounding box for each annotated action across all frames within the action segment. We then re-extract the bounding box from the original video, expanding each frame's bounding box to consistently match these maximum dimensions, ensuring the player remains centered.

To illustrate the effectiveness of this approach, Figure 4 compares keypoint trajectories for a sample data sequence without and with bounding box normalization. On the left, the raw joint trajectories without normalization are scattered, showing significant variability due to changes. In contrast, the trajectories on the right, after applying bounding box normalization, are significantly more stable and coherent.

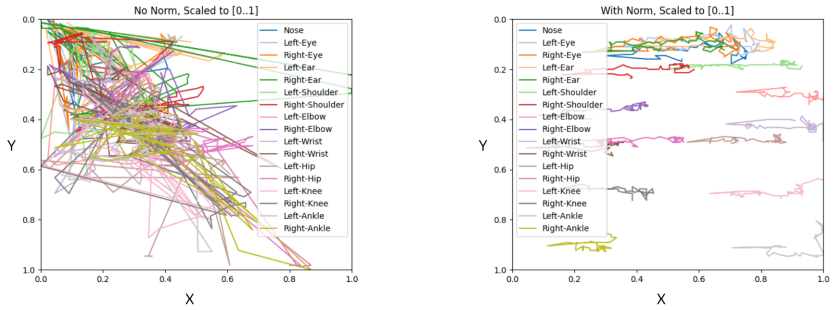

Fig. 4. Impact of bounding box normalization on joint trajectories for one data sample. Left: Without bounding box normalization, the joint trajectories are scattered and inconsistent due to variability in player scale. Right: After applying bounding box normalization, trajectories appear significantly more stable and coherent, clearly reducing positional variance and making movement patterns easier to interpret.

4.4 Pose Keypoint Extraction

We apply ice hockey-optimized 2D pose estimators [10, 11] to generate skeletal keypoints from players in normalized bounding boxes belonging to a sequence. We experimented with various keypoint configurations (Figure 5):

1. **17 Keypoints (COCO [14]):** Standard full-body representation.
2. **13 Keypoints:** Merged head keypoints.
3. **12 Keypoints:** Head keypoints omitted to focus solely on torso and limbs.
4. **15 Keypoints:** 12 body keypoints plus 3 ice hockey stick keypoints (butt end, heel, toe).
5. **20 Keypoints:** Comprehensive setup with 5 head, 12 body, and 3 stick keypoints.
6. **3 Stick Keypoints Only:** Focused solely on stick interaction.

Fig. 5. Five skeletal configurations used in this study, from left to right: 17-keypoint COCO, 13-keypoint merged head, 12-keypoint head-omitted, 15-keypoint with ice hockey stick endpoints, 20-keypoint comprehensive configuration.

4.5 GCN for Action Recognition

We employ the 2s-AGCN [9] and STGCN++ [6] architectures to classify skeleton-based action sequences. Training and evaluation leverage the mmAction2 toolbox [21]. Our approach exclusively utilizes the joint modality (keypoint coordinates) to capture spatial-temporal dynamics. This choice simplifies model inputs while effectively capturing both static postures and dynamic movements necessary for accurate ice hockey action recognition.

5 Experiments

5.1 Experimental Setup

To ensure a robust model evaluation, we partitioned the dataset described in Section 3 into training and validation sets using a 70%-30% stratified split, maintaining an approximately proportional representation of each action class. We evaluate action recognition performance using standard metrics.

We adopt the mmAction2 toolbox [21] for training and evaluation. The training process leverages the GCN-based architectures (described in Section 4.5) with cross-entropy loss. Experiments were conducted on an NVIDIA RTX 3090 Ti GPU. The training setup for our GCN experiments is as follows: a learning rate of 0.1, weight decay of 5×10^{-4}, batch size of 16, and training for a total of 20 epochs using stochastic gradient descent (SGD) as the optimizer.

5.2 Experimental Results

We evaluated our ice hockey action recognition framework using exclusively the joint modality, focusing clearly on different skeletal configurations and their impact on performance. Results across these configurations are presented in Table 2.

Table 2. Comparison of action recognition accuracy using joint modality across different skeletal configurations and models.

Model	Keypoint Configuration	Accuracy (%)
STGCN++	17 Keypoints (No Bounding Box Norm)	31
	12 Keypoints	40
	13 Keypoints	44
	17 Keypoints	57
	15 Keypoints (12 body + 3 stick)	64
	20 Keypoints (5 head + 12 body + 3 stick)	**65**
	3 Stick Keypoints Only	52
2s-AGCN	17 Keypoints	51
	15 Keypoints (12 body + 3 stick)	56

The 20-keypoint skeleton with STGCN++ attains the highest 65 % accuracy. Using the same model, bounding-box normalization plus the standard 17-keypoint human skeleton yields 57 %. Adding only three stick joints to that baseline raises accuracy by 7 pp to 64 %, and further appending five fine-grained head landmarks adds another 1 pp, reaching 65 %. Conversely, dropping head landmarks altogether (13 or 12 keypoints) lowers accuracy to 44 % and 40 %, respectively. These step-wise gains and losses show that both detailed head cues and explicit stick endpoints are essential for recognizing hockey-specific actions. Additionally, results from the 3 stick keypoints alone configuration (52%) motivate the significant role of stick interactions. STGCN++ consistently outperformed 2s-AGCN, validating our selection of STGCN++ as the core model.

5.3 Qualitative Results

Figure 6 illustrates qualitative predictions from our models on six representative action clips. Each image captures the midpoint frame of a two-second action sequence along with ground-truth labels and predictions from five experimental setups: (1) without bounding box normalization, (2) 2s-AGCN with 17 keypoints, (3) STGCN++ with 15 keypoints, (4) STGCN++ with 17 keypoints, and (5) STGCN++ with our optimized 20-keypoint configuration.

The model without bounding box normalization often misclassifies subtle actions, mistaking stick interactions for general stickwork due to spatial inconsistencies. While the 2s-AGCN model with 17 keypoints struggles to differentiate stick actions from skating movements, our 15-keypoint and 17-keypoint STGCN++ configurations significantly improve accuracy. However, the 20-keypoint

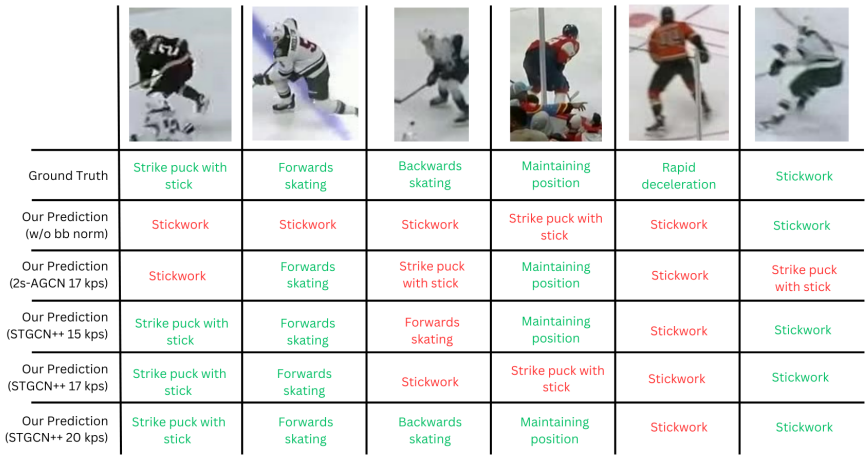

Fig. 6. Qualitative comparison of ice hockey action recognition predictions across different model setups on six action clips from our dataset. Ground-truth labels are indicated in green (correct predictions), while incorrect predictions are shown in red. The optimized STGCN++ model with 20 keypoints demonstrates improved accuracy, effectively differentiating subtle actions compared to models without bounding box normalization or fewer keypoints.

STGCN++ model excels, effectively distinguishing visually similar actions such as forward skating versus backward skating, and reliably differentiating "strike puck with stick" from general stickwork. These findings highlight the importance of incorporating detailed head and stick keypoints, along with effective bounding box normalization, to enhance ice hockey action recognition accuracy.

6 Conclusion

In this paper, we introduced CP-Hockey, a pipeline that injects two complementary contextual priors: a temporal prior that stabilizes player scale and an object prior that models the hockey stick into the skeleton-based action-recognition stack for broadcast ice-hockey footage. By counteracting camera-induced scale jitter and making stick pose explicit, CP-Hockey tackles the key challenges of tiny, fast-moving players, heavy occlusion from protective gear, and subtle stick-centric motions.

Exploiting the temporal prior alone, our bounding-box normalization lifts top-1 accuracy from 31 % to 57 % with the general 17-keypoint skeleton. Adding the object prior (three stick end-points) further raises accuracy to 64 %. Our most expressive 20-keypoint configuration, which also restores detailed head landmarks, tops out at 65 %. These gains confirm that action recognition benefits not merely from more keypoints but from the right context-aware keypoints, aligned with the physics and semantics of ice hockey.

Although CP-Hockey demonstrates strong performance on pre-segmented action clips, extending the method to the action spotting [24] task would be challenging. Future work should focus on developing mechanisms for automatic action spotting in continuous video streams. Integrating multi-view or depth information may also enable more accurate 3D pose estimation. Additionally, incorporating contextual signals such as puck tracking and player interactions could further enhance recognition of complex, team-based plays.

Overall, our results demonstrate that combining robust spatial normalization, sport-specific skeletal representations, and spatio-temporal graph models substantially improves action recognition in visually challenging sports scenarios. Validated on real NHL broadcast footage, CP-Hockey lays a strong foundation for advanced ice hockey analytics and provides a blueprint for extending action recognition methods to other demanding sports domains.

Acknowledgments. This study is funded by Stathletes Inc and NSERC (Natural Sciences and Engineering Research Council) Alliance program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fani, M., Neher, H., Clausi, D.A., Wong, A., Zelek, J.: Hockey Action Recognition via Integrated Stacked Hourglass Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 85–93 (2017). https://doi.org/10.1109/CVPRW.2017.17
2. Cheng, X., Li, Z.: Research on Basketball Shooting Action Recognition and Optimization System Based on Deep Learning. In: ICMIII, Melbourne, Australia, pp. 546–551 (2024). https://doi.org/10.1109/ICMIII62623.2024.00108
3. Zhu, H., Liang, J., Lin, C., Zhang, J., Hu, J.: A Transformer-based System for Action Spotting in Soccer Videos. In: ACMMM, pp. 103–109 (2022). https://doi.org/10.1145/3552437.3555693
4. Klyukin, M.: volleyball_dataset: An Open Source Volleyball Action Recognition Dataset. Roboflow Universe (2022). Available at: https://universe.roboflow.com/mikhail-klyukin/volleyball_dataset
5. Khobdeh, S.B., Yamaghani, M.R., Sareshkeh, S.K.: Basketball Action Recognition Based on the Combination of YOLO and a Deep Fuzzy LSTM Network. J. Supercomput. **80**, 3528–3553 (2024). https://doi.org/10.1007/s11227-023-05611-7. Available at: https://doi.org/10.1007/s11227-023-05611-7
6. Duan, H., Wang, J., Chen, K., Lin, D.: Pyskl: Towards Good Practices for Skeleton Action Recognition. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 7351–7354 (2022).
7. Wen, H., Lu, Z., Shen, F., et al.: Improving Skeleton-Based Action Recognition with Interactive Object Information. Int. J. Multimed. Info. Retr.
8. Yu, B., Yin, H., Zhu, Z.: Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 3634–3640 (2018). https://doi.org/10.24963/ijcai.2018/505

9. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In: Proceedings of CVPR, pp. 12018–12027 (2019). https://doi.org/10.1109/CVPR.2019.01230
10. Balaji, B., Clausi, D.A.: Towards Agile Human Pose Estimation: A Benchmark Study in Ice Hockey Analytics. JCVI 9(1), 54–57 (2024). https://doi.org/10.15353/jcvis.v9i1.10014
11. Balaji, B., Bright, J., Chen, Y., Rambhatla, S., Zelek, J., Clausi, D.A.: Seeing Beyond the Crop: Using Language Priors for Out-of-Bounding Box Keypoint Prediction. In: NIPS, pp. 102897–102918 (2024).
12. Cioppa, A., Deliège, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M., Gade, R., Moeslund, T.B.: A Context-Aware Loss Function for Action Spotting in Soccer Videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2020).
13. Shahroudy, A., Liu, J., Ng, T.-T., Wang, G.: NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. arXiv:1604.02808 (2016)
14. Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context. arXiv:1405.0312 (2015). Available at: https://arxiv.org/abs/1405.0312
15. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693 (2014). https://doi.org/10.1109/CVPR.2014.471
16. Hou, R., Chen, C., Shah, M.: Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5823–5832 (2017). https://doi.org/10.1109/ICCV.2017.620
17. Krishnan, K., Prabhu, N., Babu, R.V.: ARRNET: Action Recognition through Recurrent Neural Networks. In: 2016 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5 (2016). https://doi.org/10.1109/SPCOM.2016.7746614
18. Sen, A., Minhaz Hossain, S.M., Ashraf Russo, M., Deb, K., Jo, K.H.: Fine-Grained Soccer Actions Classification Using Deep Neural Network. In: 2022 15th International Conference on Human System Interaction (HSI), pp. 1–6 (2022). https://doi.org/10.1109/HSI55341.2022.9869480
19. Prakash, H., Chen, Y., Rambhatla, S., Clausi, D.A., Zelek, J.: VIP-HTD: A Public Benchmark for Multi-Player Tracking in Ice Hockey. Journal of Computational Vision and Imaging Systems **9**(1), 22–25 (2024). https://doi.org/10.15353/jcvis.v9i1.10006. Available at: https://openjournals.uwaterloo.ca/index.php/vsl/article/view/5858
20. Li, M., Hu, H., Yan, H.: Ice Hockey Puck Tracking through Broadcast Video. Neurocomputing **551**, 126484 (2023). https://doi.org/10.1016/j.neucom.2023.126484. Available at: https://www.sciencedirect.com/science/article/pii/S0925231223006070
21. MMAction2 Contributors: OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. Github repository (2020). Available at: https://github.com/open-mmlab/mmaction2
22. Yin, H., Sinnott, R.O., Jayaputera, G.T.: A Survey of Video-Based Human Action Recognition in Team Sports. Artif. Intell. Rev. **57**, 293 (2024). https://doi.org/10.1007/s10462-024-10934-9
23. Wei, J., Yu, B., Zhang, H., Liu, J.: Skeleton Based Graph Convolutional Network Method for Action Recognition in Sports: A Review. In: 2023 8th IEEE International

Conference on Network Intelligence and Digital Content (IC-NIDC), Beijing, China, pp. 66–70 (2023). https://doi.org/10.1109/IC-NIDC59918.2023.10390711

24. Denize, J., Liashuha, M., Rabarisoa, J., Orcesi, A., Héralt, R.: COMEDIAN: Self-Supervised Learning and Knowledge Distillation for Action Spotting Using Transformers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, pp. 530–540 (2024).