Spatial Refinement for 3D Human Mesh Recovery in Ice Hockey Broadcast Videos

Zhibo Wang, Sirisha Rambhatla, Yuhao Chen, David Clausi
University of Waterloo
{zhibo.wang, sirisha.rambhatla, yuhao.chen1, dclausi}@uwaterloo.ca

Abstract

Incorporating 3D Human Mesh Recovery (HMR) is crucial for understanding player actions and game scenarios. However, challenges remain, including vertical drift along the Z-axis and the lack of a mechanism to align frame-based X–Y positions with a top-down rink template. In this work, we address these issues by introducing spatial refinement strategies within the HMR pipeline. First, we reduce vertical drift by estimating a temporally consistent ground plane and correcting each reconstructed mesh's Z-axis trajectory. Second, we link 3D predictions to tactical analysis by projecting players onto a standardized top-down rink template via homography estimation. Integrated into the GVHMR framework with a hockey-specific tracking model, our refined pipeline yields more stable and physically plausible meshes and enables accurate reconstruction of player trajectories. Experiments on NHL and AHL broadcast videos demonstrate substantial improvements in temporal consistency and spatial alignment, facilitating reliable downstream analysis of player behavior and game strategy.

1. Introduction

Ice hockey is one of the most popular winter sports worldwide [9, 27]. Beyond its cultural and economic significance, the sport's fast pace and strategic complexity pose substantial challenges for automated video analysis, making it difficult to accurately capture and interpret player actions on the rink [5]. A promising direction toward addressing these challenges is the use of **3D Human Mesh Recovery (HMR)**, which provides detailed player motion information and enables a more comprehensive understanding of game dynamics.

HMR reconstructs a full 3D mesh for each player, offering not only geometric cues such as position and orientation but also rich information useful for downstream tasks, including action recognition [4], interaction analysis [13], and team strategy modeling [23]. However, when applied to ice hockey footage, existing HMR models still face two key limitations here. First, frame-wise inference often introduces small but consistent errors along the vertical (Z) axis, which accumulate over time and manifest as unrealistic drift. Second, although the horizontal (X–Y) coordinates may be stable in 3D space, there is no established method for projecting these coordinates onto a standardized top-down rink template, which is essential for tactical and spatial analysis.

Figure 1. Visualization result on an AHL video. Reconstructed human meshes are shown in grey.

To address these issues, we introduce refinement strategies tailored for ice hockey video. For the Z-axis, we estimate a temporally consistent ground plane and adjust the reconstructed meshes accordingly to eliminate vertical drift. For the X–Y axis, we incorporate a homography estimation module to project player positions from the broadcast view onto the canonical rink template, enabling accurate reconstruction of full-game spatial layouts. A visualization result is shown by Figure 1.

In summary, the main contributions of this work are as follows:

- We integrate a tracking model specifically designed for ice hockey into the 3D HMR pipeline, enabling more accurate and temporally consistent player localization.

- We propose a robust Z-axis refinement approach that regresses a temporally smooth ground plane, significantly improving the physical plausibility of reconstructed vertical motion.
- We incorporate homography estimation to bridge the gap between 3D mesh predictions and 2D rink coordinates, enabling reconstruction of player trajectories in a standardized top-down view and supporting higher-level game analysis.

2. Related Works

2.1. 3D Human Mesh Recovery (HMR)

3D Human Mesh Recovery (HMR) aims to reconstruct the full 3D body mesh of a person from monocular RGB input, where the human body is typically represented using a parametric model such as SMPL [12] or SMPL-X [17]. Early approaches primarily relied on optimization-based methods, which minimize 2D reprojection errors to iteratively refine 3D pose and shape estimations. With the rapid development of deep neural networks, learning-based methods have become the dominant paradigm. The canonical work in this area, HMR [8], employs a pre-trained CNN backbone to extract image features and directly regress pose parameters, inspiring numerous subsequent improvements [10, 11, 28].

More recently, transformer-based architectures leveraging the Vision Transformer (ViT) [3] backbone have facilitated modeling temporal dependencies across video frames. Building on this, HMR 2.0 [6] incorporates temporal information to enhance motion stability and cross-frame consistency, establishing a robust foundation for video-based HMR. Several extensions have followed this direction. Multi-HMR [1] replaces the conventional two-stage pipeline with a one-stage design that directly regresses body parameters for multiple people within a single frame, thereby improving efficiency in multi-person scenarios. TRAM [26] unifies the human body coordinate system across all video frames to reconstruct more coherent motion trajectories, while CameraHMR [16] introduces a learned regressor to estimate camera intrinsics, resulting in more accurate 3D pose reconstruction. Beyond the camera coordinate space, some works [21, 22] also leverage a ground-based world coordinate system to produce physically more plausible human models.

Despite these advancements, existing HMR frameworks have primarily been evaluated on general-purpose datasets such as Human3.6M [7] and 3DPW [25], leaving their robustness in sports-specific contexts largely unexplored.

2.2. Sports and Ice Hockey Vision Applications

Computer vision for sports analytics has attracted growing research attention in recent years, driven by its potential to provide quantitative insights into player performance, team strategy, and overall game dynamics. In the context of ice hockey, a wide range of computer vision tasks have been explored to support automated understanding of the game. Broadly speaking, these tasks can be categorized into two levels: low-level understanding and high-level understanding.

Low-level tasks aim to extract information directly observable from the video, such as homography estimation [20], player and puck tracking [19, 24], and human pose estimation [14]. High-level tasks, in contrast, require abstract reasoning and temporal modeling to infer complex semantics, including player action recognition [2], team strategy analysis [23], and event detection [15].

A critical link between these two levels of understanding is the recovery of the **3D human pose**. On one hand, 3D pose estimation builds upon low-level visual cues such as player keypoints, spatial localization, and camera geometry. On the other hand, the recovered 3D poses provide rich motion cues—including body orientation, limb articulation, and facial direction—that can be leveraged to predict player intent and anticipate future actions. Consequently, accurate 3D human pose estimation serves as a fundamental bridge between perception-level tasks and higher-level reasoning in sports video analysis, offering a unified representation for reconstructing and interpreting complex game scenarios.

3. Method

3.1. Overall Framework

The baseline model employed in this work is GVHMR [21], which is built upon HMR 2.0 [6] and incorporates a gravity-based coordinate refinement module to enhance temporal stability, thereby achieving more consistent 3D reconstructions in video-based settings. However, when applied to ice hockey footage, the original GVHMR model exhibits suboptimal performance. Our analysis suggests that this limitation primarily arises from the tracking module—the first stage of GVHMR—which has not been trained on ice hockey data and thus fails to produce reliable player tracks. Consequently, the downstream components, including pose estimation, feature extraction, and human mesh reconstruction, also suffer in accuracy and temporal consistency.

To address this issue, we replace the original generic object tracker with a hockey-specific off-the-shelf tracking model that better adapts to the sport's unique motion patterns and appearance characteristics [18]. The overall framework of our proposed pipeline is illustrated in Fig. 2.

3.2. Ground-Consistent Z-Axis Refinement

The original GVHMR model suffers from accumulated per-frame errors that cause noticeable drift along the vertical (Z) axis. To mitigate this and ensure physically plausible foot–ground contact, we estimate a temporally consistent

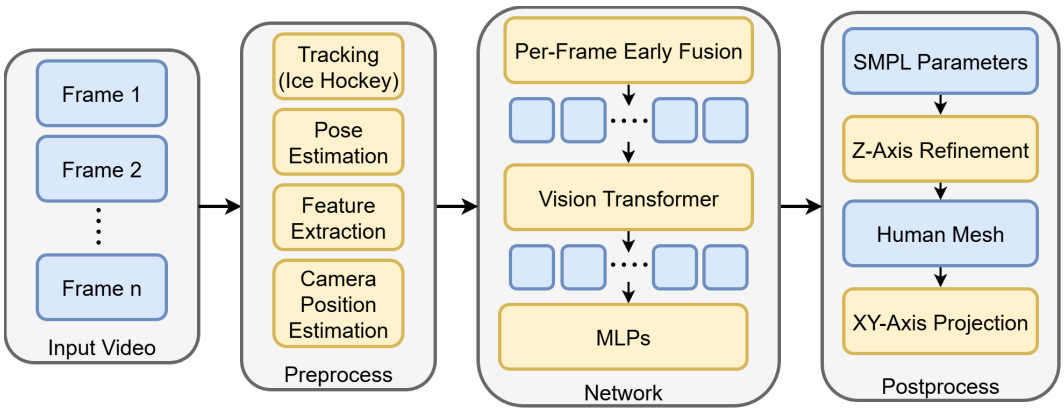


Figure 2. Overall framework of the proposed model. The input video is first preprocessed frame by frame using an ice-hockey-specific tracking module, after which the frames are fed into a ViT-based backbone to predict SMPL parameters, including pose θ, body shape β, and camera translation π. Finally, both Z-axis refinement and X–Y axis projection modules are applied to obtain the final temporally consistent 3D human mesh and corresponding 2D trajectories.

ground plane and refine the root-joint height accordingly.

For each frame I, we compute the minimal vertex height $z_{\min}$ as an approximate foot–ground contact indicator. For each contact frame t, we sample foot-surface points and form tuples (x, y, t, z), where (x, y) is the horizontal root position, z the sampled height, and t the frame index. Including t allows the model to capture slow ground-level variations caused by camera motion or drift.

We fit a plane

$$z = ax + by + ct + d, \tag{1}$$

using a RANSAC-regularized linear regressor over overlapping temporal windows of sizes $\{L, 2L\}$ (with $L = 15$). Each window produces a local estimate $\hat{z}_t$, and the final ground height is obtained via a kernel-weighted average:

$$z_t = \frac{\sum w_t \hat{z}_t}{\sum w_t}, \qquad w_t = \exp\left(-\frac{|t - t_c|}{\tau}\right), \tag{2}$$

where t_c is the window midpoint and $\tau = L/2$.

The refined root height is then computed as

$$\Delta z_t = z_t - z_{\min}, \tag{3}$$

yielding a smooth, ground-aligned vertical trajectory. This correction removes long-term drift and suppresses frame-to-frame jitter, leading to more stable and physically consistent meshes.

3.3. X–Y Axis Projection and Homography-Based Alignment

Unlike the Z-axis, which often exhibits noticeable drift, the X and Y coordinates predicted by GVHMR [21] remain relatively stable due to its gravity-aligned coordinate system. However, effective analysis of player behavior and team strategy requires mapping these coordinates onto a consistent 2D rink template. This necessitates establishing a reliable connection between the 3D reconstruction and the canonical rink plane.

To accomplish this, we adopt a homography-based transformation that links the broadcast view to a top-down template. By applying the estimated homography to the X–Y coordinates derived from the 3D mesh, we obtain accurate player positions on the standardized rink plane, enabling tasks such as action understanding, and tactical interpretation.

Specifically, starting from the original GVHMR output, we first project the reconstructed human mesh back onto the broadcast frame. We then extract the mesh vertices corresponding to each player's left and right feet. By connecting these two foot points, we obtain a line segment whose midpoint provides a stable approximation of the player's current position on the image plane. Finally, using the homography matrix predicted by the off-the-shelf model [20], we map the player's image-plane position onto the standardized rink template according to:

$$[x,\, y,\, 1]^\top = H\,[x',\, y',\, 1]^\top,$$

where (x', y') denotes the player's position in the broadcast frame, (x, y) denotes the corresponding position on the rink template, and H is the estimated 3×3 homography matrix.

This process allows us to recover each player's complete trajectory on the top-down rink template, providing a consistent spatial representation that benefits downstream tasks.

4. Experiments

4.1. Implementation Detail and Dataset

All experiments are conducted on a server equipped with two NVIDIA RTX 6000 Ada GPUs. We use the pre-

trained checkpoints provided by the official GVHMR repository [21] for all modules except the tracking component, for which we adopt the model and pretrained weights from [18].

For evaluation, we curate a collection of video clips from both the NHL (National Hockey League) and AHL (American Hockey League), covering a diverse range of game scenarios. The video lengths vary from a few seconds to approximately one minute, while all clips share a uniform resolution of 1280×720 at 30 fps.

4.2. Qualitative Results

4.2.1. 3D Human Mesh

Using an AHL broadcast video, we generated the 3D human meshes and projected them back onto the original frames. Representative visualization results are shown in Fig. 1.

As illustrated in the figure, the reconstructed meshes (shown in gray) align closely with the players' silhouettes in the broadcast footage, demonstrating the effectiveness and visual fidelity of our approach.

4.2.2. Z-Axis Stability Evaluation

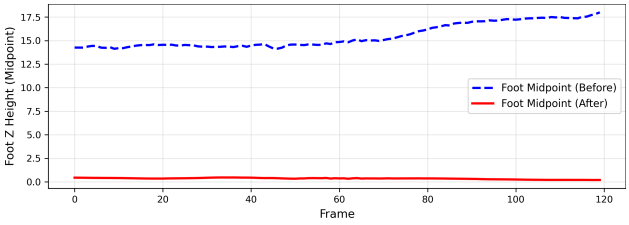
For a video sequence of 120 frames, we visualize the Z-axis trajectory of several players before and after applying our refinement module. The results are shown in Fig. 3.

As illustrated in Fig. 3, the original GVHMR predictions exhibit noticeable vertical drift. Player 1 shows a gradual upward trend, as if stepping onto an elevated surface, whereas Player 3 exhibits significant frame-to-frame jitter. Additionally, the initial Z-axis values are not centered around zero, which we attribute to inconsistencies in scale across the training data. In contrast, after applying our Z-axis refinement, the trajectories remain stable and consistently close to zero throughout the entire sequence, demonstrating the effectiveness of our method in eliminating vertical drift and stabilizing the reconstruction.

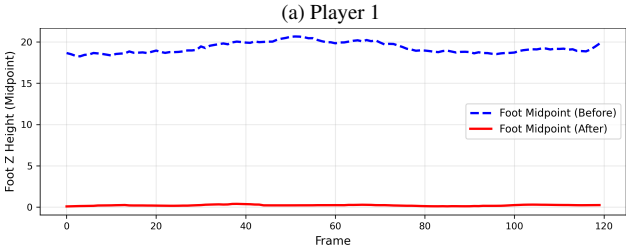
4.2.3. X–Y Axis Projection onto the Rink Template

Using the homography matrix estimated from the broadcast frame, we map each player's image-plane position onto the standardized rink template. This allows us to reconstruct full player trajectories throughout the sequence, as illustrated in Fig. 4.

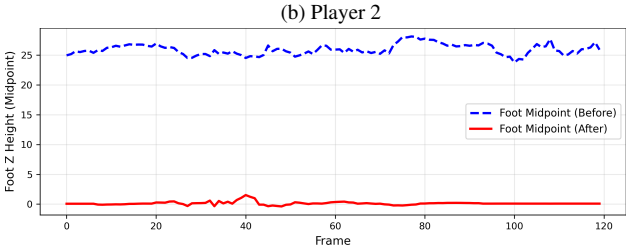
We use the same test video as in Fig. 1. By comparing each player's starting point (shown as a red dot) with their corresponding location in the first broadcast frame, we observe that the projected positions accurately align with the template. This confirms that the homography transformation effectively preserves spatial structure and enables reliable trajectory reconstruction.



(a) Player 1



(b) Player 2



(c) Player 3

Figure 3. Z-axis values of the foot midpoint (the midpoint between the left and right foot vertices) for three players. The blue dashed curve shows the original GVHMR output, while the red solid curve corresponds to our refined result.

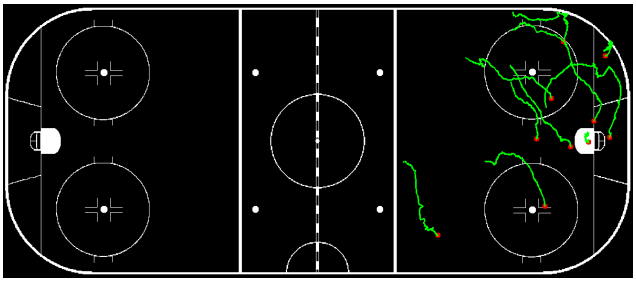


Figure 4. Reconstructed player trajectories on the rink template from a broadcast hockey clip. Red dots indicate each player's starting position.

5. Conclusion

In this paper, we introduced spatial refinement modules for 3D Human Mesh Recovery in ice hockey, addressing errors in both the Z-axis and X–Y plane. Integrated into the GVHMR baseline, these refinements improve temporal stability, reduce drift, and enable accurate projection onto the rink template. Experiments on broadcast videos demonstrate clear gains in reconstruction quality and spatial consistency.

References

[1] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 2

[2] Kseniia Buzko, Amir Nazemi, David A Clausi, and Yuhao Chen. Ice hockey action recognition via contextual priors. In *Linköping Hockey Analytics Conference*, pages 2–14, 2025. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[4] Mehrnaz Fani, Helmut Neher, David A Clausi, Alexander Wong, and John Zelek. Hockey action recognition via integrated stacked hourglass network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 29–37, 2017. 1

[5] Keisuke Fujii. Computer vision for sports analytics. In *Machine Learning in Sports: Open Approach for Next Play Analytics*, pages 21–57. Springer, 2025. 1

[6] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14783–14794, 2023. 2

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[8] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[9] Bruce Kidd. Canada's 'national'sport. *Sport in society*, 16 (4):351–361, 2013. 1

[10] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 2

[11] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 2

[13] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9687–9697, 2024. 1

[14] Helmut Neher, Mehrnaz Fani, David A Clausi, Alex Wong, and John Zelek. Pose estimation of players in hockey videos using convolutional neural networks. In *2017 Ottawa Hockey Analytics Conference (OTTHAC), Ottawa, Canada*, 2017. 2

[15] Ken Nsiempba, Amir Nazemi, David Clausi, and John Zelek. Leveraging player tracking for event detection in ice hockey. *Journal of Computational Vision and Imaging Systems*, 10 (1):69–74, 2025. 2

[16] Priyanka Patel and Michael J. Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571, 2025. 2

[17] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[18] Harish Prakash, Jia Cheng Shang, Ken M Nsiempba, Yuhao Chen, David A Clausi, and John S Zelek. Multi player tracking in ice hockey with homographic projections. *arXiv preprint arXiv:2405.13397*, 2024. 2, 4

[19] Liam Salass, Jerrin Bright, Amir Nazemi, Yuhao Chen, John Zelek, and David Clausi. Ice hockey puck localization using contextual cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6059–6068, 2025. 2

[20] Jia Cheng Shang, Yuhao Chen, Mohammad Javad Shafiee, and David A Clausi. Rink-agnostic hockey rink registration. In *Proceedings of the ACM International Workshop on Multimedia Content Analysis in Sports*, pages 73–81, 2023. 2, 3

[21] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 4

[22] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, 2024. 2

[23] Craig A. Staunton and Glenn Björklund. A framework for the standardization of game analysis in ice hockey. *International Journal of Sports Physiology and Performance*, 18(5): 458 – 464, 2023. 1, 2

[24] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A. Clausi, and John S. Zelek. Player tracking and identification in ice hockey. *Expert Systems with Applications*, 213: 119250, 2023. 2

[25] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[26] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *Computer Vision – ECCV 2024*, pages 467–487, 2025. 2

[27] John Wong and Scott R Jedlicka. When culture meets capital: commercialism, national identity, and vancouver's initial attempt to join the nhl. *Sport History Review*, 50(2):225–243, 2019. 1

[28] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023. 2