Temporally Stable Rink Homography Estimation via 3D Reconstruction and Segmentation Fusion

Liam Salass, Bowen Dai, Yuhao Chen, John Zelek, David Clausi
{liam.salass, b27dai, yuhao.chen1, jzelek, dclausi}@uwaterloo.ca
University of Waterloo, Waterloo, ON, Canada, N2L 3G1

Abstract

Homography estimation in broadcast hockey video is challenging due to frequent occlusions, motion blur, and limited visibility of rink markings. Prior work estimates a homography independently for each frame, leading to substantial temporal jitter that undermines downstream tasks such as trajectory analysis and event recognition. We introduce a two-stage homography pipeline that leverages sequence-level 3D reconstruction and aggregated segmentation evidence. A stable rink plane is first extracted from a monocular 3D reconstruction, after which per-frame rink segmentations are warped into the plane and fused to recover the rink's global layout. A single plane-to-template homography is then estimated, producing a temporally stable and metrically consistent mapping from image space to rink coordinates. Experiments show that puck trajectories transformed using our homography are smooth and free of frame-to-frame jitter.

1. Introduction

Computer vision has become a prevalent tool in ice hockey analytics [1, 2, 10, 11, 14, 16, 18–20]. Its applications span fundamental tasks such as player detection and tracking [11, 20], pose estimation [2, 14], and jersey number recognition [1, 18], as well as more advanced analyses, including gameplay strategy assessment [10, 16, 19] and puck possession estimation [3, 17]. Estimating temporally stable homographies remains an ongoing challenge in both sports analytics and computer vision, as it provides a translation from the camera's perspective to the entire field of play and real-world coordinates.

Although several approaches have been developed for rink registration and homography estimation in hockey video[4, 8, 15], these methods compute a homography independently for each frame, making them highly sensitive to broadcast artifacts, occlusions, and segmentation noise. As a result, the recovered mappings often exhibit temporal jitter, which limits their usefulness for downstream analytics that require stable rink-space trajectories.

(a) (b)

Figure 1. Left: Input broadcast frame. Right: Corresponding 3D segmented reconstruction of the rink surface.

We address this problem with a two-stage homography estimation pipeline that combines monocular 3D reconstruction (Pi3[21]) with multi-frame segmentation aggregation. The 3D reconstruction provides a stable estimate of the rink plane across the entire sequence, while aggregated plane-space segmentations recover the rink's global orientation and scale. Together, these components yield a temporally stable and metrically consistent homography suitable for broadcast hockey analytics.

2. Related Works

2.1. Homography Estimation

A homography provides a projective transformation between two views of the same planar surface and is widely used in sports analytics, robotics, and augmented reality. In its classical formulation, a homography $\mathbf{H} \in \mathbb{R}^{3\times3}$ maps points between two planes using homogeneous coordinates:

$$\lambda \tilde{\mathbf{x}}' = \mathbf{H}\tilde{\mathbf{x}}, \tag{1}$$

where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ denote corresponding points in the source and target images. The most common method for estimating $\mathbf{H}$ is the Direct Linear Transform (DLT), which requires at least four point correspondences[7].

In practice, homography estimation in vision pipelines often relies on feature detection and matching, such as SIFT[9], ORB[13], or learned keypoint detectors. Matched

features provide noisy correspondences, and robust estimators such as RANSAC are typically used to reject outliers and obtain a stable solution[5]

2.2. Ice Hockey Homography Estimation

Early work on ice-rink registration in broadcast hockey video treats homography estimation as a frame-wise regression problem. Fani et al. [4] train a ResNet18-based network to directly regress the image coordinates of four control points on a canonical rink template for each input frame. The video is first segmented into shots, then every frame within a shot is processed independently, and a homography is recovered from the four predicted correspondences using DLT. Because the regressor operates on single frames, the resulting homographies exhibit frame-to-frame jitter; temporal consistency is enforced only *post hoc* by smoothing the control-point trajectories with a Hann window before recomputing the per-frame homographies.

More recent work exploits richer rink structure while still operating on individual frames. Shang et al. propose a rink-agnostic registration pipeline that first segments rink markings in a broadcast frame, then estimates and iteratively refines the homography between this segmentation map and an overhead rink template using ResNet-based regressors and a four-point parameterization [15]. Their method leverages domain adaptation and synthetic data so that a single model can generalize across NHL and non-NHL rink geometries, but each homography is still inferred independently from a single segmented frame. Complementary to these approaches, the HockeyRink dataset introduces dense keypoint annotations and a YOLOv8-based pose model that predicts 56 rink landmarks per frame, enabling precise per-frame homography estimation from keypoints without temporal coupling between frames [8].

Across all three lines of work, homography is ultimately derived from single-frame evidence, with temporal stability either ignored or handled by simple smoothing, in contrast to our 3D reconstruction–based approach that explicitly enforces consistency over time while adapting the rink-agnostic formulation of Shang et al. to a temporally coherent 3D rink representation.

3. Methodology

Our goal is to estimate a *temporally stable* homography for broadcast hockey video. We first use 3D reconstruction to obtain a stable rink plane across the sequence, avoiding the jitter inherent to single-frame 2D methods. Since the plane alone does not provide rink orientation or scale, we aggregate per-frame rink segmentations in plane space and compute a single plane-to-template homography, yielding a consistent and metrically accurate rink alignment.

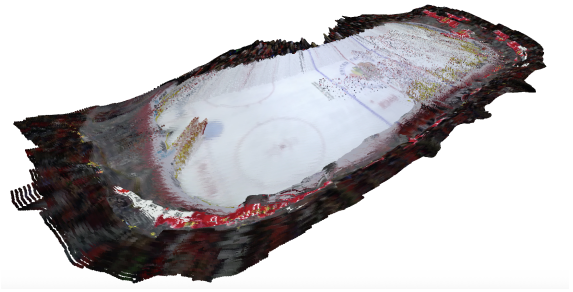

Figure 2. 3D ice hockey rink reconstruction with dynamic players filtered.

3.1. 3D Reconstruction and Dynamic Point Filtering

Each broadcast frame is processed using the Pi3 monocular depth estimation model [21], which produces: (1) a dense local point cloud $\mathbf{P}_{\text{local}} \in \mathbb{R}^{H\times W\times 3}$ in the camera coordinate frame, (2) its corresponding world-coordinate point cloud $\mathbf{P}_{\text{world}}$, and (3) the camera pose matrix $\mathbf{T} \in \mathbb{R}^{4\times 4}$.

Local points are mapped into the global frame by

$$\mathbf{P}_{\text{world}} = \mathbf{R}\mathbf{P}_{\text{local}}^{T} + \mathbf{t}, \tag{2}$$

where $\mathbf{R}$ and $\mathbf{t}$ are extracted from $\mathbf{T}$. Because Pi3 predicts consistent camera pose trajectories, the resulting 3D world points vary smoothly across frames, giving us a more stable geometric foundation than using 2D image evidence alone.

Raw reconstructions include both the ice surface and dynamic objects (players, officials), which must be removed to isolate the rink plane. We detect players using YOLOX [6] and segment them using SAM2 [12], producing a binary mask $\mathbf{M}$ of player pixels. The mask is resized to match the point cloud resolution:

$$\mathbf{M}_{\text{local}} = \text{Resize}(\mathbf{M}, (H_{\text{local}}, W_{\text{local}}), \text{Nearest}). \tag{3}$$

Player-associated 3D points are removed:

$$\mathbf{P}_{\text{cleaned}}[i,j] = \begin{cases} \text{NaN}, & \mathbf{M}_{\text{local}}[i,j] = 1, \\ \mathbf{P}_{\text{local}}[i,j], & \text{otherwise.} \end{cases} \tag{4}$$

This produces a filtered cloud containing only the static rink surface, shown in Figure 2

3.2. Initial Image-to-Plane Homography

To recover the world–plane geometry, we fit a plane to the cleaned point cloud using RANSAC [5]. Each iteration samples three non-collinear points ($\mathbf{p}_n$), estimates a candidate normal $\mathbf{n}_k$ and plane offset d_k,

$$\mathbf{n}_k = \frac{(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1)}{\|(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1)\|}, \qquad d_k = -\mathbf{n}_k \cdot \mathbf{p}_1, \tag{5}$$

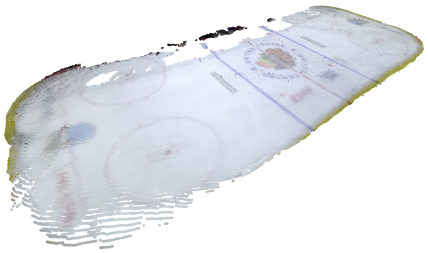

Figure 3. Filtered 3d rink reconstruction, only containing points withing $\epsilon = 0.001$ of the fit plane.

and evaluates its inliers. The best plane $(\mathbf{n}^*, d^*)$ is selected after N_{iter} iterations. This step isolates the rink surface as a clean, geometrically coherent plane.

Once the rink plane has been estimated, we recover a mapping from the broadcast image to the plane by constructing a canonical 2D representation of the fitted surface. Let $\mathcal{P}_{\text{inlier}}$ denote the set of 3D points whose distance to the plane is less than $\epsilon = 0.001$. These points form a dense sampling of the physical ice surface.

Flattening the Plane. We first construct an orthonormal basis $(\mathbf{u}, \mathbf{v})$ spanning the plane. All inlier points are projected onto this basis,

$$\mathbf{p}_{\text{plane}} = \begin{bmatrix} (\mathbf{p} - \mathbf{p}_0) \cdot \mathbf{u} \\ (\mathbf{p} - \mathbf{p}_0) \cdot \mathbf{v} \end{bmatrix}, \tag{6}$$

producing a 2D point cloud lying on the rink surface. The width of the plane image is chosen as the longest spatial extent of this flattened point set, while the height corresponds to the largest extent along the orthogonal direction. This yields a normalized 2D "plane image" whose coordinate system is aligned with the physical rink surface, shown in Figure 3.

Constructing Image–Plane Correspondences. For each inlier pixel (x_i, y_i) in the broadcast frame, Pi3 provides the associated 3D point $\mathbf{p}_i$. Since $\mathbf{p}_i$ is also mapped to a plane coordinate $\mathbf{p}_{\text{plane},i}$, we obtain direct correspondences

$$(x_i, y_i) \longleftrightarrow (u_i, v_i)$$

between image-space pixels and plane-image coordinates. Importantly, this step does not require camera intrinsics or knowledge of the camera matrix K; all correspondences are derived solely from the plane geometry and the Pi3-predicted 3D structure.

Estimating the Homography. Given these 2D–2D correspondences, the initial image-to-plane homography

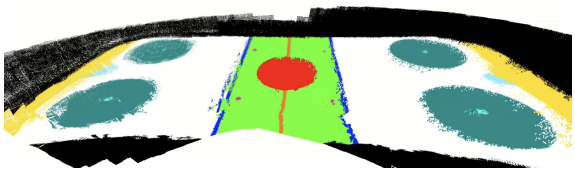

Figure 4. 3D reconstruction with segmentation repainting.

$\mathbf{H}_{\text{img}\to\text{plane}}$ is estimated using the Direct Linear Transform (DLT) with RANSAC to reject outliers:

$$\lambda \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \mathbf{H}_{\text{img}\to\text{plane}} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}. \tag{7}$$

This homography provides a stable mapping from the broadcast frame to the fitted 2D rink plane, enabling subsequent segmentation-guided normalization and template alignment.

3.3. Rink Segmentation Guided Plane Normalization

Although the fitted plane provides stable geometry, the plane alone cannot determine the *orientation*, *scale*, or exact *boundaries* of the rink. To resolve this ambiguity, we refine the plane representation using rink segmentation aggregated across multiple frames.

Segmenting Each Broadcast Frame. Following Shang *et al.* [15], we predict a semantic rink mask

$$S_t \in \{0, 1, \ldots, C\}^{H \times W}$$

for each broadcast frame t, capturing markings such as blue lines, faceoff circles, and goal creases.

Warping Segmentations Into the Plane. For each frame, we use the corresponding image-to-plane homography $\mathbf{H}^{(t)}_{\text{img}\to\text{plane}}$ to warp its segmentation mask into the plane coordinate system:

$$S_{\text{plane},t} = \mathbf{H}^{(t)}_{\text{img}\to\text{plane}}\, S_t. \tag{8}$$

Each warped mask contributes evidence about rink structure from a different viewpoint and time step, producing a temporally aggregated representation of the rink on the plane, shown in Figure 4.

Multi-Frame Plane-Space Aggregation. We fuse the warped segmentations

$$S_{\text{plane},1},\ S_{\text{plane},2},\ \ldots,\ S_{\text{plane},T}$$

by repainting the 2D plane image with the markings observed across all frames. This yields a consolidated plane-space segmentation S_{agg} that is:

- more complete than any single-frame mask (fills occlusions),
- less noisy (errors average out across multiple frames),
- geometrically aligned via the 3D reconstruction.

The aggregated segmentation provides a far more reliable estimate of the true rink boundaries and markings.

Template Alignment. Following the approach of Shang *et al.* [15] we estimate the plane-to-rink homography directly from the aggregated segmentation image. In their pipeline, a segmentation map is paired with a template, and an initial homography is estimated and then refined using a second regression stage. We apply the same two-step strategy, but entirely in the plane coordinate system: the aggregated plane-space segmentation S_{agg} plays the role of the input segmentation, and the canonical rink template T_{rink} serves as the reference. The initial homography aligns S_{agg} to the template, and the refinement step further corrects residual misalignment. This produces a single, globally consistent homography

$$\mathbf{H}_{\text{plane}\to\text{rink}}$$

that maps the 2D plane representation into the canonical rink coordinate system.

Final Two-Stage Homography. Combining the per-frame image-to-plane mappings with the global plane-to-rink alignment yields a final per-frame image-to-rink homography:

$$\mathbf{H}^{(t)}_{\text{img}\to\text{rink}} = \mathbf{H}_{\text{plane}\to\text{rink}}\,\mathbf{H}^{(t)}_{\text{img}\to\text{plane}}. \tag{9}$$

This mapping converts any image-space coordinate from frame t directly into rink coordinates. By first stabilizing geometry through 3D reconstruction and then correcting orientation and scale through multi-frame segmentation aggregation, this two-stage formulation produces a temporally stable and metrically consistent rink-space representation across the entire sequence.

3.4. Rink-Space Coordinate Transformation

Let (μ_x, μ_y) denote the puck's image coordinates in frame t. In homogeneous form:

$$\tilde{\mathbf{p}} = \begin{bmatrix} \mu_x \\ \mu_y \\ 1 \end{bmatrix}. \tag{10}$$

Warping into rink coordinates uses the per-frame homography:

$$\tilde{\mathbf{p}}' = \mathbf{H}^{(t)}_{\text{img}\to\text{rink}}\,\tilde{\mathbf{p}}. \tag{11}$$

After normalization,

$$(p_{\text{warp},x}, p_{\text{warp},y}) = \left(\frac{\tilde{p}'_x}{\tilde{p}'_z}, \frac{\tilde{p}'_y}{\tilde{p}'_z}\right). \tag{12}$$

Given the template resolution (1280×720) and NHL rink dimensions ($L = 61$ m, $W = 25.9$ m):

$$x_{\text{rink}} = \frac{p_{\text{warp},x}}{1280}\,L, \qquad y_{\text{rink}} = \frac{p_{\text{warp},y}}{720}\,W. \tag{13}$$

Temporal Stability. 3D reconstruction stabilizes frame-to-frame geometry; segmentation stabilizes scale and orientation. Their combination eliminates the drift and jitter typical of single-frame homography estimation, producing consistent rink-space trajectories across entire sequences.

4. Results and Analysis

Because no publicly available hockey dataset includes ground-truth temporally consistent homographies, we evaluate our approach qualitatively. Figure 5 shows the puck trajectory reprojected into rink coordinates using our final two-stage homography. The trajectory is smooth and free of the frame-to-frame oscillations typically produced by single-frame homography estimation. This demonstrates that integrating 3D plane estimation with aggregated segmentation provides a stable geometric mapping that preserves temporal coherence throughout the sequence.

Figure 5. Puck trajectory projected onto the rink plane reconstruction.

5. Conclusion

We presented a two-stage homography estimation pipeline that combines 3D reconstruction with multi-frame segmentation aggregation to produce temporally stable rink-space mappings from broadcast hockey video. By isolating a consistent rink plane and aligning it to a canonical template using fused segmentation evidence, our method eliminates the jitter characteristic of single-frame approaches. The resulting homographies enable smooth and reliable trajectory analysis and offer a foundation for more robust broadcast-view hockey analytics.

Acknowledgments

This work was supported by a grant with the Natural Sciences and Engineering Research Council (NSERC) partnered with Stathletes, Inc. Stathletes also provided the puck annotation data used in this research.

References

[1] Bavesh Balaji, Jerrin Bright, Sirisha Rambhatla, Yuhao Chen, Alexander Wong, John Zelek, and David A Clausi. Domain-guided masked autoencoders for unique player identification. 1

[2] Bavesh Balaji, Jerrin Bright, Yuhao Chen, Sirisha Rambhatla, John Zelek, and David Clausi. Seeing beyond the crop: Using language priors for out-of-bounding box keypoint prediction. *Advances in Neural Information Processing Systems*, 37:102897–102918, 2024. 1

[3] Xin Duan. *Automatic determination of puck possession and location in broadcast hockey video*. PhD thesis, University of British Columbia, 2011. 1

[4] Mehrnaz Fani, Pascale Berunelle Walters, David A. Clausi, John Zelek, and Alexander Wong. Localization of ice-rink for broadcast hockey videos, 2021. 1, 2

[5] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 2

[6] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021, 2021. 2

[7] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 1

[8] Mehdi Houshmand Sarkhoosh, Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Tomas Kupka, and Pål Halvorsen. Hockeyrink: A dataset for precise ice hockey rink keypoint mapping and analytics. In *Proceedings of the 16th ACM Multimedia Systems Conference*, page 249–255, New York, NY, USA, 2025. Association for Computing Machinery. 1, 2

[9] Tony Lindeberg. *Scale Invariant Feature Transform*. 2012. 1

[10] 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, 2024. 1

[11] 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. 1

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[13] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. 1

[14] Marjan Shahi, David Clausi, and Alexander Wong. Goalienet: A multi-stage network for joint goalie, equipment, and net pose estimation in ice hockey. *arXiv preprint arXiv:2306.15853*, 2023. 1

[15] Jia Cheng Shang, Yuhao Chen, Mohammad Javad Shafiee, and David A. Clausi. Rink-agnostic hockey rink registration. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, page 73–81, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2, 3, 4

[16] Sijia Tian. *Group event recognition in ice hockey*. PhD thesis, University of British Columbia, 2018. 1

[17] Moumita Roy Tora, Jianhui Chen, and James J Little. Classification of puck possession events in ice hockey. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 147–154. IEEE, 2017. 1

[18] Kanav Vats, Mehrnaz Fani, David A Clausi, and John Zelek. Multi-task learning for jersey number recognition in ice hockey. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, pages 11–15, 2021. 1

[19] Kanav Vats, Mehrnaz Fani, David A Clausi, and John Zelek. Puck localization and multi-task event recognition in broadcast hockey videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4567–4575, 2021. 1

[20] 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. 1

[21] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning, 2025. 1, 2