Video-Based Player Re-Identification in Ice Hockey via Non-Contextual Implicit Features

Evan Iaboni, Amir Nazemi, Yuhao Chen, and David A. Clausi
Vision and Image Processing Lab, University of Waterloo

Abstract

Player re-identification (ReID) in ice hockey is difficult due to similar uniforms, motion blur, occlusions, and obscured jersey numbers. We introduce a video-based method that focuses on extracting implicit features by combining OSNet spatial features with a lightweight temporal transformer. Using an ice hockey dataset, our approach benefits from simple data augmentations and outperforms state-of-the-art video ReID models in a zero shot setting by +4.7% mAP and +4.3% rank-1 accuracy. Analysis of the learned embeddings shows that the implicit and non-contextual features learned by the model are efficient enough to capture explicit attributes such as team, handedness, and jersey number.

1. Introduction

Video-based person re-identification (ReID) is a computer vision task that aims to match and recognize individuals across different camera views in video sequences. Unlike image-based ReID, which relies on single still images, video-based ReID leverages temporal information, such as motion patterns, and appearance consistency over time, to improve recognition accuracy and robustness [8]. This approach is particularly useful in surveillance and security applications, where variations in pose, lighting, occlusion, and camera angles make identification challenging [14].

The sports domain introduces additional challenges, such as uniform similarity, compression artifacts, motion blur, background consistency, and optical character recognition (OCR) [6]. Prior work has attempted to uniquely identify the jersey number [2] [12]. However, this approach fails when the jersey number is partially occluded or entirely absent from the frame [3]. Recent work in person re-identification relies on large models such as vision transformers, which require substantial training data or pretraining on a closely matched data distribution [16] and typically incur longer inference times [7].

To address these gaps we propose a lightweight spacial encoder with a temporal transformer that pools frame level appearance features. This design improves retrieval performance while implicitly capturing attributes like jersey number, team color, and handedness. The result is a compact and effective framework tailored for the unique visual and temporal challenges of sports ReID. Our contributions are as follows:

- We reformulate an existing ice hockey tracking dataset [13] into a video-based ReID dataset.
- We compare our method against state-of-the-art person ReID models on this dataset.
- We evaluate our model's ability to detect useful player features such as jersey color, jersey number, and stick handedness.

2. Related Work

Image-Based Person ReID. Habel et al. [6] demonstrated that large-scale vision-language models can be repurposed for sports applications by leveraging a CLIP-based contrastive training objective. Their approach achieved top performance in both the Player Re-Identification Challenge 2022 [11] and the SoccerNet 2023 Player Re-Identification Challenge [4]. These results highlight the potential of pretrained visual encoders in recognizing athletes across multiple views without explicit supervision. However, such methods rely on still frames and thus fail to capture the temporal context in sports footage.

Video-Based Person ReID. A number of works have explored temporal modeling for person re-identification. Gu et al. [5] proposed AP3D, a 3D convolutional architecture that preserves appearance cues across frames. Liu et al. [7] presented a lightweight CNN framework for efficient video-based Re-ID. Alsehaim and Breckon [1] introduced ViD-Trans-ReID, a transformer-based model designed to enhance temporal attention in video sequences. Yu et al. [15] later proposed TF-CLIP, a text-free variant of CLIP that learns temporal embeddings for person Re-ID tasks. Despite their effectiveness on pedestrian datasets, these models tend to perform poorly when applied to ice hockey data.

Sun et al. [10] extended video-based Re-ID to the sports setting through a tracklet association framework that builds on top of multi-object tracking systems. However, their pairwise OSNet comparison across frames does not incor-

porate temporal information and is computationally inefficient at test time.

3. Methodology

3.1. Problem formulation

Video-based player ReID aims to match tracklets of individual players captured throughout the course of a game. Let a tracklet be defined by a sequence of frames $T = \{F_1, F_2, ..., F_m\}$ that depict the same player within a continuous period of time. Each tracklet is associated with an identity label $y \in \{1, 2, ..., C\}$ during training, while test identities are disjoint from the training set.

Given two tracklets T_i and T_j, we wish to learn a mapping $f(\cdot)$ that produces a discriminative tracklet-level embedding such that the cosine similarity $s(f(T_i), f(T_j))$ is high when $y_i = y_j$ and low otherwise. During inference, a query tracklet is compared against a gallery of tracklets, and ranking scores derived from $s(\cdot)$ guide retrieval.

3.2. Architecture

We adopt OSNet [17] as the spacial feature extractor for our network. Its lightweight architecture allows efficient processing over long frame sequences. To aggregate spacial features across each traklet, we employ a transformer with CLS token pooling. The architecture is depicted in figure 1.

3.3. Confidence Awareness

Reliable frame level information is essential for producing stable tracklet embeddings, since many frames in broadcast hockey footage suffer from motion blur, low resolution, and occlusions. To help the temporal transformer reason about which frames contain trustworthy visual cues, we propose incorporating an explicit measure of frame confidence into the model's input. This explicit feature is not from ice hockey context and is not in contrast with the goal of this research which is extracting non-contextual implicit features.

As a first step toward confidence awareness, we supply the transformer with a score that reflects the amount of motion blur present in each frame. We compute blur score using the variance of Laplacian method introduced by Pech-Pacheco et al. [9]. For each frame, we calculate its variance of Laplacian, normalized to be between 0 and 1, and append it to the 512-dimensional OSNet output feature vector. the resulting 513-dimensional vector is then padded with zeros so its size remains divisible by the number of heads in the transformer.

4. Experiments

4.1. Dataset

We introduce VIP-VHReID (VIP - Video Hockey ReID), a restructured version of the NHL player tracklets dataset introduced in Kanav et al. [13]. The original dataset consists of 84 NHL broadcast clips captured from a single camera setup with a resolution of 1280×720 pixels and a frame rate of 30 frames per second, with an average tracklet length of 191 frames.

We modify the dataset by excluding referees, replacing jersey-number labels with identity labels, reorganizing the train–test split by team to allow evaluation of zero-shot performance, and converting the test portion into the standard ReID query–gallery structure by assigning the first detection of each identity to the query set and placing all remaining detections in the gallery.

The training set consists of 1537 player tracklets with 261 unique identities and the test set consists of 1193 tracklets with 206 unique identities along with 22 distractors, which are tracklets that do not correspond to any labeled identity and are included to make the evaluation more challenging

4.2. Training Details

Our training pipeline has two stages. In the first stage, OSNet is trained for 10 epochs on identity classification using a cross entropy objective. The model uses a learning rate of 0.0003, weight decay of 0.0005, and a batch size of 32 frames on a single GPU

In the second phase, we freeze OSNet and train the temporal transformer for 30 epochs, also with an identity classification objective based on cross-entropy loss. This model is trained with a learning rate of 0.0003, weight decay of 0.0005, and a batch size of 20 tracklets on a single GPU. To reduce overfitting and improve sample diversity, the transformer is trained on 20 randomly sampled frames from each tracklet which are resampled every epoch. For the transformer architecture, we use 1 layer, 8 heads, and a feed forward dimension of 2048. We find that larger transformer architectures tend to overfit to our dataset.

4.3. Results

We evaluate performance on VIP-VHReID using standard retrieval accuracy metrics: mean average precision (mAP) and rank-1 accuracy. Table 1 shows the performance of several person ReID models on our ice hockey dataset. CFANN [7] and AP3D [5] are based on convolutional neural network (CNN) architectures, while Clip-ReIDent [6], ViD-Trans-ReID [1], and TF-CLIP [15] are based on more complex transformer architectures. The more complex models tend to perform poorly on our dataset.

4.4. Feature Classification

In this subsection, we evaluate our model's ability to implicitly detect stick handedness, team, and jersey number. Figure 2 shows a player tracklet frame and the corresponding activation map from OSNet's final convolutional layer.

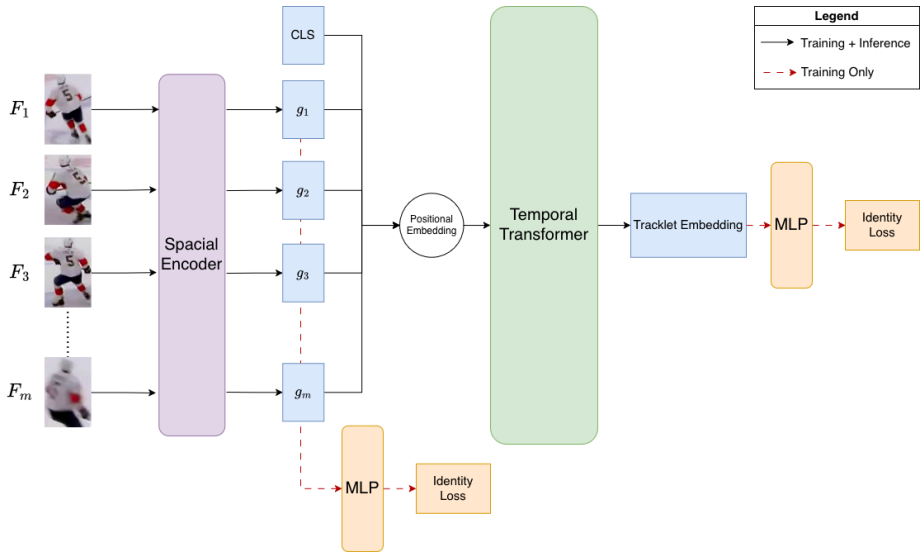

Figure 1. **Overview of the proposed architecture.** The input to our network is a sequence of frames $F_1, F_2, ..., F_m \in \mathbb{R}^{256\times128}$. Each frame F_i initially goes through OSNet to extract spacial features. Next, the outputted 512-dimensional OSNet embeddings get concatenated and passed through a temporal transformer to extract spatiotemporal features. Black lines get run during both training and inference, while red dashed lines are only run during training.

Method	mAP	Rank-1
OSNet [17]	42.1	48.2
AP3D [5]	48.8	61.7
CF-ANN [7]	40.3	43.5
Clip-ReIDent [6]	32.4	32.5
VID-Trans-ReID [1]	21.0	20.9
TF-CLIP [15]	40.2	45.6
Ours	53.5	66.0
Ours + Blur Score	**54.1**	**69.9**

Table 1. **Performance comparison on VIP-VHReID dataset.** Mean Average Precision (mAP) and Rank-1 accuracy are reported for several baseline and state-of-the-art methods.

Figure 2. **OSNet activation map.** A player frame is shown alongside the corresponding OSNet final layer activation map. The network focuses on the jersey number region, which indicates that OSNet learns the importance of jersey number even without explicit supervision.

It appears that OSNet primarily pays attention to the jersey number region, indicating that OSNet has the ability to recognize that jersey number is an important attribute.

4.4.1. Handedness

To evaluate the model's ability to infer player handedness, we train a linear classifier on top of the frozen model. All other model parameters remain fixed while the classifier is trained using ground-truth handedness labels. The resulting classifier achieves 75.8% accuracy in predicting player handedness, outperforming the random baseline of 63.1% that reflects the dataset's skew toward left handed players. These results demonstrate that our model is able to implicitly identify handedness to some extent, however further performance improvements could be achieved by incorporating explicit handedness labels or using additional information to aid in handedness identification (e.g., pose keypoints).

4.4.2. Team

Figure 3 shows a t-SNE plot of our model's ability to cluster players by team and by whether they are home or away. The model does an excellent job at team-based clustering, which is especially impressive since we use a zero-shot dataset, meaning no teams in the training set appear in the test set. Note that the micro-clusters in the top-middle of the figure are goaltenders.

4.4.3. Jersey Number

The dataset contains only a small number of examples per jersey number, making supervised classification difficult.

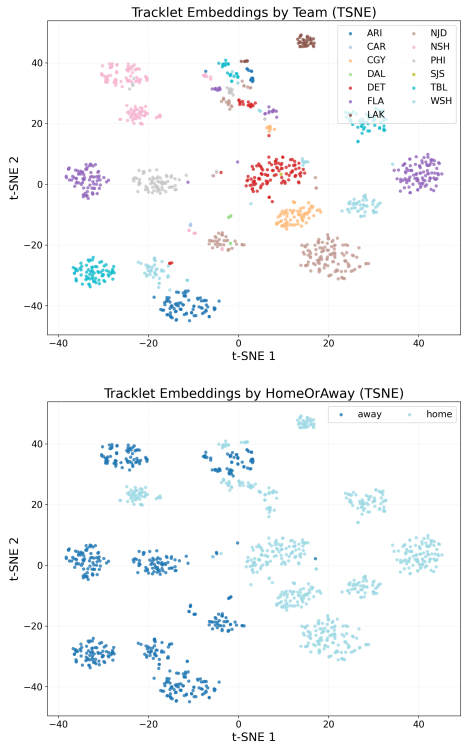

Figure 3. **t-SNE plot of tracklet embeddings colored by team (top) and home or away (bottom).** These plots show clear team-based clustering despite the zero shot evaluation protocol where no test teams appear during training.

Instead of training a classifier, we evaluate whether players who share a jersey number have more similar embeddings than those who do not. We make the following definitions:

- **Intra-number similarity:** the cosine similarity between two player embeddings with the same jersey number.
- **Inter-number similarity:** the cosine similarity between two player embeddings with different jersey numbers.

We draw 1000 bootstrap samples from the set of all possible intra-number similarity and inter-number similarity pairs. For each sample we compute the sample mean, and we report the average of these bootstrap means and their variance. Note that we exclude pairs of players from the same team because teammates tend to have similar embeddings due to uniform similarity. Since teammates always have different jersey numbers, including these pairs would artificially raise the average inter-number similarity.

The resulting average inter-number similarity is 0.0850 ± 0.0002, and the average intra-number similarity is 0.1120 ± 0.0015. This indicates that player's with the same jersey number are, on average, more similar in latent space. Hence, our model has some understanding of the differences between jersey numbers.

4.5. Ablation Study

We conduct ablations to assess how the number of frames per tracklet and data augmentations influence performance. Table 2 examines the effect of varying the number of frames per tracklet. Using 20 frames yields the strongest results, while both shorter and longer sequences lead to small drops. This suggests that 20 frames strike the best balance between temporal diversity and redundancy.

Frames per Tracklet	mAP	Rank-1
10	50.1	61.2
20	**53.5**	**66.0**
30	53.4	65.5
40	52.9	64.1

Table 2. **Ablation study to determine the best number of frames per tracklet.** We choose to use 20 frames as it achieves the best performance.

Table 3 evaluates several augmentation strategies. Each individual augmentation provides an improvement over the baseline, with random erase giving the largest single boost. The combination of random crop and random erase achieves the strongest overall performance in mAP, whereas applying all augmentations together results in a small drop, suggesting that overly aggressive augmentation is less effective.

Augmentations	mAP	Rank-1
No Augmentations	48.6	58.7
Random Crop	50.6	65.0
Random Erase	53.0	**68.5**
Random Rotation	51.2	62.1
Color Jitter	51.3	60.7
Random Crop + Random Erase	**53.5**	66.0
All augmentations	51.3	63.6

Table 3. **Ablation study to determine augmentations.** We achieve the best performance with random crop + random erase.

5. Conclusion

In this paper we introduce a re-identification model that performs well on our ice hockey dataset in the zero-shot setting. Additionally, we show that our model has the ability to identify implicit player attributes. Future work will look at incorporating explicit feature labeling and domain specific constraints like temporal consistency and shift distribution information.

6. Acknowledgment

This work was supported by a grant with the Natural Sciences and Engineering Research Council (NSERC) partnered with Stathletes, Inc.

References

[1] A. Alsehaim and T.P. Breckon. Vid-trans-reid: Enhanced video transformers for person re-identification. 2022. 1, 2, 3

[2] Bavesh Balaji, Jerrin Bright, Harish Prakash, Yuhao Chen, David A Clausi, and John Zelek. Jersey number recognition using keyframe identification from low-resolution broadcast videos, 2023. 1

[3] Alvin Chan, Martin D. Levine, and Mehrsan Javan. Player identification in hockey broadcast videos. *Expert Systems with Applications*, 165:113891, 2021. 1

[4] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdullah Kamal, Adrien Maglo, Albert Clapés, Amr Abdelaziz, Artur Xarles, Astrid Orcesi, Atom Scott, Bin Liu, Byoungkwon Lim, Chen Chen, Fabian Deuser, Feng Yan, Fufu Yu, Gal Shitrit, Guanshuo Wang, Gyusik Choi, Hankyul Kim, Hao Guo, Hasby Fahrudin, Hidenari Koguchi, Håkan Ardö, Ibrahim Salah, Ido Yerushalmy, Iftikar Muhammad, Ikuma Uchida, Ishay Be'ery, Jaonary Rabarisoa, Jeongae Lee, Jiajun Fu, Jianqin Yin, Jinghang Xu, Jongho Nang, Julien Denize, Junjie Li, Junpei Zhang, Juntae Kim, Kamil Synowiec, Kenji Kobayashi, Kexin Zhang, Konrad Habel, Kota Nakajima, Licheng Jiao, Lin Ma, Lizhi Wang, Luping Wang, Menglong Li, Mengying Zhou, Mohamed Nasr, Mohamed Abdelwahed, Mykola Liashuha, Nikolay Falaleev, Norbert Oswald, Qiong Jia, Quoc-Cuong Pham, Ran Song, Romain Hérault, Rui Peng, Ruilong Chen, Ruixuan Liu, Ruslan Baikulov, Ryuto Fukushima, Sergio Escalera, Seungcheon Lee, Shimin Chen, Shouhong Ding, Taiga Someya, Thomas B. Moeslund, Tianjiao Li, Wei Shen, Wei Zhang, Wei Li, Wei Dai, Weixin Luo, Wending Zhao, Wenjie Zhang, Xinquan Yang, Yanbiao Ma, Yeeun Joo, Yingsen Zeng, Yiyang Gan, Yongqiang Zhu, Yujie Zhong, Zheng Ruan, Zhiheng Li, Zhijian Huang, and Ziyu Meng. Soccernet 2023 challenges results. *Sports Engineering*, 27(2), 2024. 1

[5] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. *Appearance-Preserving 3D Convolution for Video-Based Person Re-identification*, pages 228–243. 2020. 1, 2, 3

[6] Konrad Habel, Fabian Deuser, and Norbert Oswald. Clipreident: Contrastive training for player re-identification. 2022. 1, 2, 3

[7] Chih-Ting Liu, Jun-Cheng Chen, Chu-Song Chen, and Shao-Yi Chien. Video-based person re-identification without bells and whistles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1491–1500. IEEE, 2021. 1, 2, 3

[8] Haifei Ma, Canlong Zhang, Yifeng Zhang, Zhixin Li, Zhiwen Wang, and Chunrong Wei. A review on video person re-identification based on deep learning. *Neurocomputing*, 609:128479, 2024. 1

[9] J.L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, pages 314–317 vol.3, 2000. 2

[10] Jiacheng Sun, Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, and Jenq-Neng Hwang. Gta: Global tracklet association for multi-object tracking in sports, 2024. 1

[11] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. Deepsportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, page 1–8. ACM, 2022. 1

[12] Kanav Vats, William J. McNally, Pascale Walters, David A. Clausi, and John S. Zelek. Ice hockey player identification via transformers. *CoRR*, abs/2111.11535, 2021. 1

[13] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A. Clausi, and John Zelek. Player tracking and identification in ice hockey, 2021. 1, 2

[14] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook, 2021. 1

[15] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification, 2023. 1, 2, 3

[16] Xiaoyu Zhang, Rui Cai, Ning Jiang, Minwen Xing, Ke Xu, Huicheng Yang, Wenbo Zhu, and Yaocong Hu. Te-transreid: Towards efficient person re-identification via local feature embedding and lightweight transformer. *Sensors (Basel, Switzerland)*, 25, 2025. 1

[17] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification, 2019. 2, 3