Self-Supervised Interleaved Motion Representation Learning for Long-Range Sports Video Analytics

Ishaan Smith
Department of Computer Science, University of North Texas, Denton, TX, USA.
ishaan.smith@unt.edu

Abstract

Long-range sports video analytics presents unique challenges due to the need to capture fine-grained motion patterns over extended temporal horizons while maintaining computational efficiency and robustness to domain shifts. Traditional supervised approaches require extensive human annotation and often fail to generalize across different sports, camera setups, and environmental conditions. This paper proposes a self-supervised interleaved motion representation learning framework that leverages hierarchical multi-stream architectures to encode motion at multiple temporal scales without reliance on labeled data. The framework integrates contrastive and predictive self-supervised objectives within an interleaved encoder design, enabling the model to learn structured representations that disentangle short-term dynamics from long-term dependencies. System-level considerations including architectural trade-offs between model capacity and inference speed, the role of data augmentation and negative sampling strategies, and the implications for deployment on edge devices are examined. Furthermore, the paper addresses issues of fairness, such as demographic biases in broadcast sports data, and discusses governance frameworks for responsible deployment in automated coaching and officiating assistance. Experimental evaluations on benchmark sports video datasets demonstrate that the proposed approach achieves competitive performance on downstream tasks including action recognition, event localization, and player trajectory prediction. The work contributes a scalable, annotation-free paradigm for long-range video understanding and provides a critical analysis of the socio-technical infrastructure required for real-world adoption.

Keywords

self-supervised learning, interleaved motion representation, long-range video analytics, sports video understanding, hierarchical encoding, system architecture, fairness in AI, video prediction.

1. Introduction

The analysis of sports video has become a cornerstone of modern athletics, enabling performance evaluation, tactical analysis, and automated broadcasting. However, the inherent complexity of long-range sports video—characterized by rapid movements, occlusions, multiple interacting agents, and variable camera angles—demands representation learning approaches that can capture both instantaneous actions and evolving strategies over minutes of play [1]. Recent advances in deep video understanding have largely focused on supervised learning using large-scale annotated datasets [2], yet the annotation cost for fine-grained temporal labels remains prohibitive, and supervised models often overfit to specific sport domains [3]. Self-supervised learning offers a compelling alternative by leveraging the temporal structure of video itself to learn robust motion representations without human labels [4].

A key challenge in self-supervised video representation learning is designing pretext tasks that capture motion dynamics across varying temporal receptive fields. Early contrastive methods operate at the clip level, learning invariance to spatial augmentations but failing to discriminate fine-grained motion patterns [5]. Predictive approaches, such as future frame prediction, encourage the model to reason about temporal evolution but suffer from high computational cost and blurry predictions [6]. To address these limitations, we propose an interleaved motion representation learning framework that combines contrastive and predictive objectives within a hierarchical multi-stream architecture. The framework explicitly models short-range and long-range temporal dependencies through interleaved processing of motion features extracted at multiple scales.

From a systems perspective, the proposed approach must balance representational power with computational feasibility for real-time or near-real-time inference in sports environments. Architecture design choices, including the number of streams, temporal stride, and dimensionality of latent codes, directly impact both accuracy and latency. Furthermore, deployment in diverse settings—from professional stadiums with high-resolution cameras to amateur setups with mobile devices—requires careful consideration of model compression, hardware acceleration, and energy efficiency. This paper provides a thorough analysis of these trade-offs and discusses the infrastructural prerequisites for widespread adoption.

Equally important are the societal implications of deploying automated sports analytics systems. Biases in training data arising from unequal coverage of different sports, genders, or ethnicities can lead to unfair performance disparities [7]. Additionally, the use of such systems in officiating or talent scouting raises concerns about transparency, accountability, and the potential for reinforcing existing inequalities. We therefore examine governance mechanisms and fairness-aware training strategies to mitigate these risks.

2. Architectural Foundations for Interleaved Motion Encoding

Interleaved motion representation learning draws from the observation that effective video understanding requires processing both short-term and long-term information in a coordinated manner. Early two-stream architectures, such as those processing RGB and optical flow independently, established the benefit of separate pathways for appearance and motion [8]. Subsequent work introduced slow-fast networks, where a slow pathway operates at high resolution but low frame rate to capture semantic content, while a fast pathway operates at low resolution but high frame rate to capture motion [2]. These designs, however, treat temporal scales as decoupled streams that are fused only at the decision layer, limiting the cross-scale interactions essential for long-range reasoning.

The interleaved architecture overcomes this limitation by alternating between streams at multiple temporal granularities, enabling continuous information exchange. Each stream processes a distinct temporal sampling rate and feature resolution, with cross-stream connections allowing fine-grained motion cues to inform high-level semantics and vice versa [5]. This design is particularly beneficial for sports video, where a single play may involve rapid micro-motions such as a player’s footwork and macro-level positioning over several seconds. By interleaving representations at different scales, the model can learn to associate these disparate temporal phenomena within a unified representation space.

A critical architectural decision is the choice of interleaving frequency and the dimensionality of the latent representations. Frequent interleaving allows tighter coupling between streams but increases computational overhead and may lead to overfitting on synthetic patterns.

Sparse interleaving, by contrast, preserves stream independence but reduces the model's ability to capture cross-scale dependencies. Empirical studies suggest that a moderate interleaving pattern—every four to eight temporal steps—yields the best trade-off for long-range sports tasks [9]. The representation dimensionality must also be tuned to match the information richness of each stream; allocating too few dimensions to the fast stream can cause loss of motion detail, while too many dimensions inflates memory usage.

From an infrastructure perspective, interleaved architectures lend themselves to parallelization on modern hardware. The multiple streams can be executed concurrently on separate GPU cores, with synchronization points at interleaving steps. However, the synchronization overhead and data transfer bandwidth between streams become bottlenecks in distributed settings. Deployment on edge devices, such as smartphone processors or embedded cameras, requires pruning or quantization of the interleaved structure, often at the cost of degraded long-range performance. These hardware-induced trade-offs necessitate a holistic design process that co-optimizes architecture parameters with target deployment constraints.

3. Self-Supervised Learning Paradigms for Video

Self-supervised learning for video can be broadly categorized into contrastive methods, which bring representations of temporally close clips closer while pushing apart clips from different videos, and predictive methods, which require the model to forecast future frames or reconstruct masked spatiotemporal regions [4]. Both paradigms have been successful in learning generic visual representations, but their application to long-range motion has distinct limitations.

Contrastive learning, as exemplified by SimCLR and MoCo, relies on data augmentation to create positive pairs [10, 11]. In the video domain, standard augmentations such as random cropping, color distortion, and temporal jittering can inadvertently destroy motion structure. For instance, aggressive temporal jittering may misalign the motion trajectories that define a sports action. To mitigate this, we employ a motion-preserving augmentation strategy that applies spatial transforms consistently across frames and uses mild temporal shifts that maintain the ordering of critical events. The contrastive objective is computed at multiple temporal scales using features extracted from the interleaved encoder, enforcing consistency between short-term and long-term representations of the same video sequence.

Predictive self-supervision, such as future frame prediction or video inpainting, inherently captures temporal dynamics but tends to produce high-frequency errors that degrade representation quality [6]. The interleaved architecture provides a natural way to incorporate prediction tasks: the fast stream can be tasked with predicting the next few frames, while the slow stream predicts a longer-term summary. This hierarchical prediction reduces the burden on a single decoder and encourages the model to learn motion features at appropriate time scales. Moreover, the interleaving process ensures that prediction errors from one stream can be compensated by information from the other stream, improving overall stability.

A key trade-off between contrastive and predictive objectives concerns the balance between discriminability and reconstruction fidelity. Contrastive learning yields high-quality semantic features but may ignore fine-grained motion patterns that are not useful for distinguishing clips from different videos. Predictive learning, on the other hand, forces the model to capture low-level motion details but can lead to over-reliance on static background cues. Our framework combines both objectives via a multi-task loss, with a learnable weight that adapts based on task difficulty. This dynamic balancing is especially important in sports video where

the motion signal varies across game phases—intense action sequences benefit more from predictive loss, while transitions and downtime benefit from contrastive consistency.

4. System-Level Trade-Offs and Deployment Considerations

Deploying self-supervised interleaved motion representation learning in practice involves navigating a complex design space that includes model architecture, data pipeline, hardware infrastructure, and real-time constraints. The most immediate trade-off is between model capacity and inference speed. Larger models with more streams and higher dimensionality achieve superior representation quality, as measured by downstream task accuracy, but their latency can exceed the frame rate requirements of live sports broadcast (typically 25 to 60 frames per second). For example, a model with four interleaved streams and feature dimensions above 1024 may require over one second per frame on a single GPU, rendering it unsuitable for real-time use. Compression techniques such as knowledge distillation from a larger teacher model can reduce latency while retaining most of the representational power, but they incur an additional training overhead [12].

Data pipeline design is equally critical. Self-supervised methods require large amounts of unlabeled video, and the preprocessing steps—frame extraction, optical flow computation or motion feature extraction, augmentation—must be streamlined to avoid becoming a bottleneck. In a distributed training environment, data loading and augmentation are often the most time-consuming stages. Sharding the dataset across nodes and performing on-the-fly augmentation using GPU-based libraries can alleviate this issue, but introduces complexity in maintaining deterministic behavior for reproducibility [13]. For sports video, variable video lengths and frame rates further complicate batching; a common approach is to sample fixed-length clips with random start points, but this can lead to underutilization of long-range dependencies if clips are too short.

Energy consumption is an often-overlooked aspect of sustainable deployment. Large self-supervised models, when trained for hundreds of epochs on millions of videos, consume substantial electricity, contributing to carbon emissions. Efficient training strategies, such as mixed-precision computation and early stopping based on validation metrics, can reduce energy usage without compromising performance. On the inference side, edge deployment for on-field analytics requires models that operate within a power budget of a few watts. Quantization to 8-bit integers and pruning of less important connections can shrink the interleaved encoder to fit mobile devices, but these techniques may disproportionately affect the fast stream's ability to capture high-frequency motion [14]. A careful sensitivity analysis is needed to determine which stream components are most robust to compression.

5. Robustness, Fairness, and Governance Implications

Robustness in sports video analytics encompasses resilience to domain shifts such as changes in camera angle, lighting, weather conditions, and variations in athlete appearance. The self-supervised interleaved framework inherently improves robustness by learning motion invariants that are less sensitive to static visual features. However, bias can still arise from the training data composition. Broadcast sports datasets are heavily skewed towards popular male-dominated leagues (e.g., NBA, NFL), leading to underrepresentation of women's sports and amateur competitions [7]. Models trained on such data perform poorly on underrepresented domains, raising fairness concerns when used for automated coaching or talent identification.

Fairness-aware training can help mitigate these disparities. One approach is to reweight the self-supervised loss based on domain membership, giving higher importance to underrepresented categories during contrastive negative sampling. Another is to use domain-adversarial training to learn representations that are invariant to sport type or gender. Both methods require additional metadata that may not be available, highlighting the need for inclusive data collection practices. From a governance perspective, organizations deploying sports analytics systems should adopt transparency protocols, including model cards that disclose training data composition and performance across subpopulations [15]. Audit trails that log model predictions and allow human review are essential for high-stakes applications such as officiating assistance.

The use of self-supervised learning also raises questions about interpretability. Representations learned without labels are often opaque, making it difficult to understand why a model identifies a certain action as a foul or a player as offside. Post-hoc explainability methods, such as attention rollouts or gradient-based saliency maps, can provide some insight, but their reliability for fine-grained motion decisions is debated [5]. Interleaved architectures, with their multiple streams, further complicate interpretation because the contribution of each stream to a final prediction is entangled. Developing stream-attribution techniques that quantify the importance of short-term versus long-term motion cues would improve trust and facilitate debugging.

Policy implications extend to data privacy, especially when analytics systems capture player biometrics or tactical strategies that could be considered proprietary. Regulations such as the General Data Protection Regulation (GDPR) in Europe require informed consent for video data collection and processing, which may conflict with the self-supervised paradigm's appetite for large, unlabeled datasets. Organizations must establish data governance frameworks that anonymize individuals where possible and limit use to agreed-upon purposes. Furthermore, the potential for automated systems to replace human coaches or referees raises labor displacement concerns. A balanced socio-technical approach that augments rather than replaces human expertise should guide deployment decisions.

6. Conclusion

This paper has presented a self-supervised interleaved motion representation learning framework tailored to the demands of long-range sports video analytics. By combining hierarchical multi-stream architectures with contrastive and predictive pretext tasks, the approach captures motion dynamics across diverse temporal scales without reliance on manual annotations. The analysis of system-level trade-offs, including model capacity, latency, energy consumption, and edge deployment, provides a blueprint for practical implementation. Furthermore, the discussion of robustness, fairness, and governance underscores the need for responsible design and deployment practices.

Future research directions include exploring dynamic interleaving strategies that adapt the frequency of cross-stream interactions based on video content, as well as integrating multimodal signals such as audio and text commentary to enrich the motion representations. Extending the framework to multi-view and multi-camera setups common in professional sports would enhance its utility for comprehensive analysis. Finally, establishing standardized benchmarks for long-range sports video understanding that include fairness evaluation dimensions would accelerate progress and ensure equitable outcomes.

The self-supervised interleaved paradigm represents a step toward scalable, annotation-free video analytics, but its success ultimately depends on careful attention to the socio-technical infrastructure in which it is embedded. Balancing performance with fairness, privacy, and sustainability will determine whether such systems become a valuable tool for athletes, coaches, and fans alike.

References

1. Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308).

2. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6202–6211).

3. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning (pp. 813–823).

4. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 9729–9738).

5. Jin, H., Yi, H., Zhao, W., Luo, J., Ye, S., Guan, Z., ... & Yu, T. (2026). HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding. arXiv preprint arXiv:2605.08158.

6. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning (pp. 1597–1607).

7. Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. In Advances in Neural Information Processing Systems (Vol. 33, pp. 21271–21284).

8. van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

9. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning (pp. 10347–10357).

10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

11. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6836–6846).

12. Zhu, P., Zhao, S., Han, F., & Deng, H. (2024, May). BEAVP: A Bidirectional Enhanced Adversarial Model for Video Prediction. In 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) (pp. 1-8). IEEE.

13. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision (pp. 20–36).

14. Lin, J., Gan, C., & Han, S. (2019). TSM: Temporal shift module for efficient video understanding. In Proceedings of the IEEE International Conference on Computer Vision (pp. 7083–7093).

15. Yang, C., Xu, Y., Shi, J., Dai, B., & Zhou, B. (2020). TANet: Towards fully automatic tracking and analysis of team sports. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1431–1441).

16. Wu, Y., Lim, J., & Yang, M. H. (2019). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1371–1380).

17. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6450–6459).

18. Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6546–6555).

19. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., & Russell, B. (2019). ActionVLAD: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1312–1321).

20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.