Digital Twin-Oriented Spatio-Temporal Modeling of Crowd Dynamics Using Hierarchical Multi-Stream Video Representations

Terry J. Bowman
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
terryjbowman348@unr.edu

Zhuting Liu
Department of Computer Science, University of Houston, Houston, TX, USA.
liu1995@uh.edu

Bennett C. Steele
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
bennetts@ucf.edu

Abstract

The convergence of digital twin technology and advanced video representation learning offers a transformative paradigm for modeling crowd dynamics in complex urban and infrastructural environments. This paper presents a systematic framework for constructing digital twin-oriented spatio-temporal models that leverage hierarchical multi-stream video representations to capture the multi-scale, multi-modal nature of human movement. The proposed architecture integrates high-level semantic reasoning with low-level motion encodings, enabling robust and scalable simulation of crowd behaviors for applications in smart city management, event security, and transportation planning. We examine the structural trade-offs between model granularity, computational efficiency, and predictive accuracy, and discuss implications for system governance, data fairness, and infrastructure sustainability. A critical analysis of deployment strategies reveals the need for adaptive streaming pipelines and federated learning mechanisms to ensure real-time responsiveness and privacy compliance. The paper further considers the role of policy frameworks in governing the use of crowd models, particularly with respect to bias mitigation, accountability, and ethical boundaries. By situating the technical contributions within a broader socio-technical context, we argue that hierarchical multi-stream video representations are not merely a computational improvement but a foundational component for trustworthy and responsible digital twin ecosystems. The study draws on recent advances in motion encoding, trajectory prediction, and large-scale video understanding, and proposes a roadmap for future research that balances innovation with societal resilience.

Keywords

digital twin, crowd dynamics, spatio-temporal modeling, hierarchical multi-stream video, motion representation, smart city infrastructure, governance, fairness.

1. Introduction

Digital twins have emerged as a powerful abstraction for representing, monitoring, and simulating physical systems in real time. In the domain of crowd dynamics, digital twins serve as virtual replicas of public spaces, enabling authorities to anticipate congestion, detect anomalous behaviors, and optimize resource allocation. However, the fidelity of such models depends critically on the quality and structure of the underlying observational data. Video feeds from surveillance cameras, public transport hubs, and pedestrian walkways offer a rich but challenging source of information: high-dimensional, noisy, and spatio-temporally complex. Conventional processing pipelines often reduce this information to coarse aggregated statistics, losing the fine-grained motion patterns that are essential for accurate prediction and simulation. Recent advances in video representation learning, particularly hierarchical multi-stream architectures, provide a pathway to preserve and exploit these details while maintaining computational tractability [1], [2]. This paper argues that a digital twin-oriented approach to crowd dynamics must adopt a hierarchical, multi-stream paradigm that separates motion cues at multiple temporal scales and spatial resolutions. Such an approach not only improves predictive performance but also facilitates interpretability and modular system design.

2. Related Work and Conceptual Foundations

The literature on crowd dynamics modeling spans several decades, with early work relying on macroscopic flow equations and agent-based simulations. More recently, data-driven methods using recurrent neural networks, transformers, and graph neural networks have demonstrated superior performance in trajectory prediction and anomaly detection [3], [4]. Video-based approaches further extend these capabilities by integrating raw pixel-level information. Among these, multi-stream video architectures have been shown to capture distinct motion modalities—such as appearance, optical flow, and temporal gradients—that are complementary for understanding crowd behaviors [5], [6]. However, most existing models operate at a single spatio-temporal scale or fuse streams in a shallow manner, limiting their ability to represent hierarchical interactions such as individual movements within group flows and group flows within global crowd dynamics. The concept of hierarchical multi-stream representation has been explored in the context of action recognition and long video understanding, but its application to crowd dynamics and digital twin integration remains nascent. A notable contribution is a recent technical report that introduces hierarchical interleaved multi-stream motion encoding for long video understanding, demonstrating that interleaving features across scales improves temporal coherence and reduces information loss [13]. Another relevant line of work uses attentive radiate graphs for pedestrian trajectory prediction in disconnected manifolds, showing how graph structures can encode spatial relationships in non-Euclidean spaces [7]. These foundations motivate the design of a digital twin framework that leverages similar principles but extends them to meet the real-time, scalable, and governance-aware requirements of socio-technical systems.

3. Hierarchical Multi-Stream Video Representations for Crowd Dynamics

The core of the proposed digital twin architecture is a hierarchical multi-stream video encoder that processes input video frames through multiple branches, each specialized for a particular temporal granularity or visual modality. At the lowest level, a fast stream captures high-frequency motion features—such as sudden acceleration, directional changes, and micro-interactions—using lightweight convolutional modules that operate on consecutive frame differences. At the intermediate level, a medium stream integrates optical flow estimates over short windows, encoding smooth motion trajectories and local group formations. At the

highest level, a slow stream processes temporally subsampled frames with deeper networks to infer long-term semantic context, such as crowd density trends, event boundaries, and environmental influences [8]. These streams are interleaved through a series of cross-attention mechanisms that allow information to flow bidirectionally, ensuring that fine-grained motion details inform high-level reasoning and vice versa. The resulting multi-scale representation is then fed into a temporal aggregation module that produces spatio-temporal embeddings suitable for downstream tasks like prediction, simulation, or anomaly detection [13].

This hierarchical structure introduces important design trade-offs. Increasing the number of streams or the depth of each stream improves representational capacity but at the cost of higher computational latency and memory footprint, which may be prohibitive for real-time digital twin updates. Conversely, reducing stream complexity may degrade the model's ability to capture subtle crowd behaviors, particularly in heterogeneous environments with mixed pedestrian-vehicle interactions. A balanced architecture must therefore be tuned based on the operational requirements of the specific deployment context, such as the density of the crowd, the availability of edge computing resources, and the acceptable prediction horizon. Furthermore, the choice of fusion strategy—whether early, late, or hierarchical interleaving—affects the stability of training and the robustness to missing or corrupted video feeds. Hierarchical interleaving, as demonstrated in recent work [13], offers a favorable compromise by maintaining separate stream identities while enabling adaptive feature recombination at multiple levels of abstraction.

4. System Architecture and Integration with Digital Twin Infrastructure

Deploying hierarchical multi-stream models within a digital twin ecosystem requires careful orchestration across three layers: perception, modeling, and actuation. The perception layer consists of a distributed network of cameras and edge devices that perform preliminary video processing and stream compression to reduce bandwidth consumption. These devices run lightweight variants of the encoder that extract local motion features and transmit them to a central fusion server or a federated aggregation node. The modeling layer hosts the full hierarchical multi-stream encoder along with learning-based dynamics models that use the embeddings to predict future crowd states. This layer also maintains the digital twin's state representation—a structured virtual model of the physical space that includes static geometry, dynamic entities, and environmental conditions [9]. The actuation layer translates predicted outcomes into actionable interventions, such as rerouting pedestrian flows, adjusting traffic signals, or dispatching security personnel. Feedback loops between these layers allow the system to adapt to changing conditions, for example by reallocating computational resources to regions with high uncertainty.

A critical consideration for infrastructure sustainability is the trade-off between centralized and distributed processing. Centralized architectures offer simplified data management and model consistency but suffer from single points of failure and high latency, especially when video streams originate from diverse sensors across a city. Distributed approaches, on the other hand, improve resilience and reduce communication overhead but introduce challenges in maintaining a coherent global digital twin state across asynchronous updates. Federated learning has emerged as a promising middle ground, enabling edge nodes to collaboratively train shared model parameters without transferring raw video data, thus preserving privacy and reducing network load [10]. In the context of crowd dynamics, federated learning can also help mitigate demographic biases by ensuring that models are trained on heterogeneous data distributions from different neighborhoods, camera viewpoints, and time periods.

5. Trade-offs, Robustness, and Fairness

The effectiveness of a digital twin-oriented crowd modeling system depends not only on technical accuracy but also on its resilience to adversarial inputs, sensor failures, and distributional shifts. Hierarchical multi-stream representations, by their nature, encode redundant motion cues across streams, which can improve robustness against occlusions or temporary camera outages. For example, if the fast stream fails due to a dropped frame, the medium and slow streams may still provide sufficient motion context to maintain prediction quality. However, this redundancy increases the system's vulnerability to coordinated attacks that simultaneously corrupt multiple streams, such as subtle perturbations designed to alter optical flow estimates. Robustness can be further enhanced by incorporating uncertainty quantification mechanisms that assign confidence scores to each stream's output and dynamically down-weight unreliable streams [11].

Fairness is another critical dimension that must be addressed explicitly in the design and deployment of crowd dynamics models. Research has shown that video-based pedestrian detection and tracking systems can exhibit performance disparities across demographic groups, often due to biased training data or camera calibration artifacts [12]. Hierarchical multi-stream models that rely on handcrafted features like optical flow may inadvertently amplify these biases if the underlying sensor characteristics vary across locations. To mitigate such risks, the digital twin must incorporate fairness-aware evaluation protocols that measure prediction accuracy not only globally but also across subgroups defined by spatial zones, time of day, and observed crowd composition. Governance mechanisms, such as audit trails and human-in-the-loop review for high-stakes decisions, can further ensure that the system operates within ethical boundaries.

6. Deployment, Governance, and Policy Implications

The transition from prototype to large-scale deployment of digital twin-oriented crowd modeling systems raises several governance challenges. One fundamental issue is data sovereignty: video streams often contain identifiable individuals, and their collection, storage, and analysis must comply with privacy regulations such as the General Data Protection Regulation and similar frameworks. Hierarchical multi-stream processing can reduce privacy risks by extracting motion features locally on edge devices and transmitting only abstract embeddings rather than raw frames. However, even aggregated motion patterns can be reverse-engineered to infer sensitive behaviors, necessitating differential privacy guarantees or anonymization techniques [14]. Policy makers must therefore establish clear guidelines for the minimum level of abstraction that constitutes acceptable privacy protection without undermining the utility of the digital twin.

Another policy dimension concerns accountability for decisions made or supported by the digital twin. If a crowd management intervention recommended by the system leads to a hazardous situation, who bears responsibility—the system operators, the data provider, or the model developer? Establishing a liability framework requires transparent documentation of model limitations, uncertainty estimates, and the rationale behind each recommendation. Hierarchical multi-stream representations offer some advantage here because their modular structure allows for increased interpretability: the contribution of each stream to a given prediction can be traced, enabling post-hoc explanations that are crucial for both auditing and trust-building [15]. Additionally, the digital twin should be subject to periodic third-party audits that assess its fairness, robustness, and alignment with societal values.

From a sustainability perspective, the computational energy required for continuous video processing and model inference must be balanced against the benefits of improved crowd safety and efficiency. Hierarchical models, by design, can allocate more resources to streams that capture critical motion dynamics while throttling less informative streams, thereby reducing power consumption during low-activity periods. Adopting green computing practices, such as using energy-efficient hardware accelerators and scheduling heavy computations during off-peak electricity hours, can further mitigate the environmental footprint [16]. Finally, the long-term viability of such systems depends on their ability to adapt to emerging threats, evolving urban layouts, and shifts in crowd behavior patterns—a requirement that calls for continuous learning pipelines and versioned model governance.

7. Future Directions and Open Research Questions

Despite the promising capabilities of hierarchical multi-stream video representations, several open research questions remain. First, the integration of these models with reinforcement learning for closed-loop control of digital twins is largely unexplored. While predictive models can forecast crowd states, optimal intervention policies require learning sequential decision-making under uncertainty, which may benefit from the rich spatio-temporal embeddings provided by the encoder. Second, the scalability of hierarchical multi-stream architectures to city-wide, multi-camera networks has not been thoroughly evaluated; real-world deployments will need to handle massive data volumes with strict latency constraints. Third, cross-modal fusion that combines video streams with other sensor modalities—such as LiDAR, Wi-Fi signal strength, or social media feeds—could yield more robust crowd estimates, especially in scenarios where visual occlusions are common [17]. Fourth, the development of standardized benchmarks and evaluation metrics for digital twin-oriented crowd dynamics is essential for fair comparison and reproducibility. Current benchmarks often focus on either pure video understanding or trajectory prediction in isolation, without capturing the closed-loop, uncertainty-aware nature of digital twin applications.

From a governance perspective, future research should explore participatory design approaches that involve community stakeholders in defining the objectives and constraints of the digital twin. This could help mitigate concerns about surveillance overreach and ensure that the system serves the public good rather than narrow commercial or security interests. Moreover, international coordination on technical standards for privacy-preserving motion data representation would facilitate cross-border collaborations and interoperability of emergency response systems [18].

8. Conclusion

This paper has presented a comprehensive framework for digital twin-oriented spatio-temporal modeling of crowd dynamics using hierarchical multi-stream video representations. We have argued that the hierarchical, multi-scale nature of such representations is essential for capturing the complexity of human movement while maintaining system efficiency and interpretability. Through a detailed analysis of architectural trade-offs, deployment challenges, robustness, fairness, and governance, we have situated the technical contributions within a broader socio-technical context. The proposed framework offers a pathway toward resilient, equitable, and sustainable crowd-aware digital twins that can enhance urban safety and efficiency without compromising individual rights. As video understanding and digital twin technologies continue to mature, the integration of hierarchical multi-stream principles will be pivotal in ensuring that these systems are not only accurate but also trustworthy and aligned with societal values.

References

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

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

3. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 961-971.

4. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2255-2264.

5. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27.

6. 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. European Conference on Computer Vision, 20-36.

7. Zhu, P., Zhao, S., Deng, H., & Han, F. (2025). Attentive radiate graph for pedestrian trajectory prediction in disconnected manifolds. IEEE Transactions on Intelligent Transportation Systems.

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

9. Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. Transdisciplinary Perspectives on Complex Systems, 85-113.

10. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273-1282.

11. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning, 1050-1059.

12. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77-91.

13. 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.

14. Dwork, C. (2008). Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, 1-19.

15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

16. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63.

17. Lv, Z., Li, X., & Li, J. (2021). Multi-modal crowd counting via cross-modal fusion. IEEE Transactions on Multimedia, 24, 1023-1034.

18. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).