Cross-Modal Scene Semantics and Graph Attention Networks for Human Motion Intention Prediction

Cody C. Hansen
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
hansen86@uc.edu

Zhen Ding
Department of Computer Science, University of North Texas, Denton, TX, USA.
zhen.ding513@unt.edu

Tejas Mishra
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
tejasmail@missouri.edu

Abstract

Human motion intention prediction is a fundamental capability for autonomous systems operating in shared environments, such as autonomous vehicles, service robots, and intelligent surveillance. Traditional trajectory forecasting approaches primarily rely on observed motion history and simple spatial interactions, often neglecting the rich semantic information embedded in the surrounding scene and the complex relational structure among multiple agents. This paper proposes a comprehensive framework that integrates cross-modal scene semantics with graph attention networks to predict human motion intentions. The architecture fuses visual, depth, and semantic segmentation streams to construct a high-dimensional scene representation, which is then processed through a graph attention network that models dynamic inter-agent and agent-scene relationships. We discuss the structural trade-offs inherent in designing such a system, including the balance between computational latency and prediction accuracy, the fusion strategies for heterogeneous sensor modalities, and the scalability of graph attention mechanisms to dense crowds. Deployment considerations such as real-time inference on edge devices, robustness to sensor degradation, and sustainability of training data pipelines are examined. Furthermore, we address governance and fairness implications, particularly regarding biases in scene semantics and the equitable treatment of diverse pedestrian populations. Through a systems-oriented analysis, this paper highlights how cross-modal scene understanding and relational graph modeling can together enhance the reliability and interpretability of motion intention prediction, while also outlining open challenges for large-scale deployment in socio-technical infrastructures.

Keywords

human motion prediction, cross-modal fusion, graph attention networks, scene semantics, autonomous systems, socio-technical infrastructure, fairness.

1. Introduction

The ability to anticipate where a pedestrian or other human agent will move in the near future is critical for safe and efficient human-robot interaction, autonomous driving, and smart city management. Early methods relied on simple extrapolation of past trajectories using linear

models or recurrent neural networks that encode motion history but ignore the surrounding environment [1]. As autonomous systems become embedded in complex, unstructured environments, the need for richer contextual reasoning has become apparent. Scene semantics — the understanding of objects, surfaces, affordances, and social zones — provides essential cues about where a person is likely to walk, such as crosswalks, sidewalks, doorways, or open plazas [2]. Meanwhile, the interactions among multiple agents, whether cooperative, competitive, or indifferent, impose relational constraints that are naturally captured by graph-based representations [3].

Graph attention networks have emerged as a powerful tool for modeling non-Euclidean relational data, allowing each agent to attend to relevant neighbors and scene elements with learned importance weights [4]. When combined with cross-modal scene input — for example, RGB images, depth maps, and semantic segmentation masks — the resulting system can fuse heterogeneous sensory data into a unified representation that drives intention prediction. This paper takes a systems-level view of such an architecture, examining the design decisions, trade-offs, and governance implications that arise when deploying cross-modal graph attention predictors in real-world socio-technical contexts.

2. Related Work

Human trajectory prediction has been extensively studied using recurrent and generative models. The Social LSTM introduced a pooling layer to capture interactions among nearby pedestrians [1], while later work employed social attention mechanisms [5] and generative adversarial networks [6] to produce socially plausible trajectories. These models, however, typically ignore static scene context. Concurrently, the computer vision community developed deep networks for semantic segmentation and scene parsing, enabling extraction of high-level semantic maps [7]. Fusing trajectory data with scene semantics has been shown to improve prediction accuracy, particularly at intersections and junctions [8].

Graph-based methods offer a natural representation of multi-agent systems. Social GAN utilized a pooling network, but subsequent work formalized agent interactions as graphs with message passing [9]. Graph attention networks (GATs) introduced learnable attention coefficients that allow each node to weigh the influence of its neighbors [4]. Several studies have extended GATs to trajectory prediction by encoding both spatial and temporal dependencies [10]. Meanwhile, cross-modal fusion has been explored in autonomous driving perception, where LiDAR, camera, and radar data are combined for object detection [11]. For pedestrian intention prediction, the integration of depth and semantic cues with graph attention remains an emerging area.

The work in [12] proposes an attentive radiate graph for pedestrian trajectory prediction in disconnected manifolds, addressing scenarios where spatial regions are separated by obstacles, and the graph structure must handle non-uniform adjacency. This approach underscores the importance of scene structure in shaping relational graphs. Our framework builds upon these foundations by explicitly incorporating cross-modal scene semantics into the graph attention pipeline, while also considering the broader system-level implications.

3. System Architecture and Design Trade-offs

A cross-modal scene semantic graph attention network for motion intention prediction comprises several key modules: sensor acquisition, modality-specific encoders, a cross-modal fusion unit, a graph construction and attention module, and a prediction decoder. Each module

introduces design choices that affect overall system performance, latency, and resource consumption.

The sensor suite typically includes a color camera, a depth sensor (e.g., LiDAR or stereo camera), and possibly a thermal or event camera for low-light conditions. RGB images provide texture and color information essential for semantic segmentation, while depth data offers geometric cues about obstacles and surfaces. Fusing these modalities at an early stage — concatenating feature maps before graph construction — enables the model to learn joint representations but increases computational cost and requires careful alignment and synchronization [13]. Late fusion, where each modality undergoes separate graph attention and the outputs are combined, preserves modality-specific signals but may miss cross-modal interactions that occur at the feature level. Mid-level fusion, where a shared attention mechanism operates on a combined modality embedding, strikes a balance but introduces additional hyperparameters. In practice, the choice depends on the available compute budget and the temporal coherence of sensor streams.

The graph attention module constructs a node for each agent and optionally for static scene elements such as obstacles, crosswalks, or traffic signs. Edges are defined based on spatial proximity or semantic similarity. Attention weights are computed using a learned function that takes node features and edge attributes as input, allowing the model to attend more to relevant neighbors while ignoring distant or irrelevant ones. The complexity of the graph grows quadratically with the number of agents in dense crowds, forcing trade-offs between full connectivity and sampling strategies. Hierarchical graph attention or spatial windowing can reduce computational load while retaining critical interactions [14]. Furthermore, the temporal dimension can be incorporated by using recurrent graph networks or spatiotemporal attention across time steps.

Another critical trade-off is between prediction horizon and accuracy. Shorter horizons (one to two seconds) can be predicted with higher confidence using local motion patterns, while longer horizons (four to six seconds) require deeper reasoning about scene affordances and social norms. The architecture must be trained with a loss function that balances imitation learning (minimizing displacement error) with adversarial training to produce diverse, realistic trajectories [6]. System designers must also consider the frequency of prediction updates: high-frequency updates (e.g., every 100 ms) reduce latency but increase compute load, whereas lower frequencies may miss sudden changes in intention.

4. Cross-Modal Fusion and Scene Semantics

Scene semantics provide a structured understanding of the environment that goes beyond raw geometry. Semantic segmentation networks such as DeepLab [7] assign a class label to each pixel (e.g., sidewalk, road, grass, building, door). The resulting semantic map can be used to compute affordance scores: regions where walking is likely, permissible, or forbidden. For example, a pedestrian standing near a crosswalk is likely to wait for a signal and then cross, whereas a person on a sidewalk may continue straight or turn into a building entrance. Scene semantics also capture social norms, such as queuing areas or no-go zones around construction sites.

Cross-modal fusion aligns the semantic map with depth and RGB features. One effective approach is to project the semantic map into the same coordinate frame as the trajectory history, using camera intrinsic and extrinsic parameters, and then concatenate the semantic features with agent position and velocity encodings [2]. Alternatively, attention mechanisms

can learn to select which semantic cues are most relevant for each agent based on its current location and heading. This is especially important in cluttered scenes where many semantic labels compete for attention. The scene can also be represented as a graph of semantic regions, where edges encode connectivity (e.g., a sidewalk leads to a crosswalk) and obstacles act as disconnected manifolds, as highlighted in [12]. Such a graph can be seamlessly integrated with the agent graph, allowing the model to reason about how scene structure influences motion intentions.

Depth information adds a third dimension, crucial for understanding occlusions, elevations, and volume. A pedestrian partially occluded by a parked car may intend to emerge from behind it, and depth can disambiguate the occluder from the pedestrian. Moreover, depth aids in estimating the distance to scene elements, which modulates the urgency of avoidance maneuvers. Cross-modal fusion must handle asynchronous sensor streams, differences in resolution, and calibration errors. Robustness techniques such as dropout of random modalities during training can mitigate reliance on a single sensor and improve fault tolerance [15].

5. Graph Attention Mechanisms for Motion Prediction

Graph attention networks assign attention weights to edges based on the features of the connected nodes. In the context of motion intention prediction, each node represents an agent with state features (position, velocity, orientation, and optionally body pose), and edges are defined between agents that are within a certain distance or that have a high likelihood of interaction. The attention mechanism computes a scalar weight for each neighbor, and these weights are used to aggregate neighbor features into an updated node representation. Multiple attention heads capture different interaction patterns (e.g., collision avoidance, group walking, leader-follower) [4].

Beyond agent-agent edges, agent-scene edges connect each agent to the nearest or most salient scene nodes. Scene nodes can be derived from the semantic map, representing key points or regions such as crosswalks, doors, or benches. The graph attention can then learn to attend to scene nodes that are relevant for the agent's future motion. For instance, a pedestrian approaching a crosswalk may assign high attention to the crosswalk node, while a person standing near a building entrance may attend to the door node. This enables the model to predict not only where the agent will move but also why — a form of interpretability.

The temporal evolution of the graph is handled by stacking graph attention layers with recurrent connections or by using spatiotemporal graph networks [10]. Each time step, node positions change, edges may be added or removed, and attention weights are recomputed. Computational efficiency can be improved by using dynamic graph pruning, where edges with low attention are dropped, or by employing a fixed k-nearest neighbor graph updated only periodically. The trade-off is between losing potentially important long-range interactions and maintaining real-time performance. In highly dynamic environments, such as busy intersections, the graph structure must be updated at each inference step to capture sudden changes in mutual positions.

6. Deployment, Robustness, and Sustainability Considerations

Deploying a cross-modal scene semantic graph attention network in a real-world autonomous system presents numerous engineering challenges. First, real-time inference on resource-constrained platforms such as embedded GPUs or edge TPUs requires model compression and quantization. Knowledge distillation from a larger teacher model to a smaller student

network can retain accuracy while reducing memory and latency [16]. Second, sensor failures must be handled gracefully. If the camera is obscured or the depth sensor malfunctions, the system should degrade to a mode that uses only available modalities, potentially with reduced accuracy. Training with simulated sensor dropout and adversarial perturbations can improve robustness [15].

Sustainability of the data pipeline is another concern. Training such models requires large, diverse datasets with annotated trajectories and synchronized scene semantics. Collecting and labeling such data is labor-intensive and raises privacy issues, especially in public spaces. Synthetic data generation using game engines or digital twins can augment real data, but the domain gap must be carefully addressed [17]. Moreover, the model must be updated over time as infrastructure changes — new buildings, road layouts, or temporary obstacles — necessitating a continuous learning pipeline. Federated learning approaches could allow models to be updated using data from multiple vehicles or robots without centralizing sensitive trajectory data, though communication and privacy costs must be managed [18].

Energy consumption is a growing concern for large-scale deployment. A single autonomous vehicle may run multiple deep learning models simultaneously, and the cumulative power draw impacts battery life and thermal management. Graph attention networks, especially with multiple heads and large graphs, can be computationally intensive. Hardware accelerators designed for sparse computations and graph processing can reduce energy footprint. Additionally, the frequency of prediction updates can be dynamically adjusted based on the complexity of the scene — for example, reducing update rate in empty straight roads and increasing it in crowded intersections — saving energy without compromising safety.

7. Governance, Fairness, and Policy Implications

The deployment of motion intention prediction systems has significant governance and fairness dimensions. One major concern is algorithmic bias. If training data is collected predominantly from certain demographic groups or geographic regions, the model may underperform for underrepresented populations. For instance, pedestrian walking styles, group formations, and adherence to traffic rules vary across cultures. A system trained on data from North American cities may not generalize well to Asian cities where jaywalking is more common or where social norms differ [19]. Scene semantics themselves can encode biases: a semantic map may fail to recognize informal crosswalks or temporary pathways used by marginalized communities, leading to incorrect predictions and potentially unsafe decisions.

Fairness also relates to the allocation of risk. In autonomous driving, the prediction system influences which trajectories are considered plausible and thus which evasive actions are taken. If the system systematically overestimates the unpredictability of certain pedestrian groups (e.g., children or elderly), the vehicle may react overly cautiously, causing traffic disruptions. Conversely, underestimating unpredictability may lead to accidents. Auditing frameworks should be developed to test prediction models across demographic subgroups using synthetic and real-world scenarios [20].

Governance structures must address accountability when prediction errors lead to harm. Is the manufacturer, the software developer, or the fleet operator responsible? Transparent model interpretability can help: attention weights and scene attributes that contributed to a prediction can be logged for post-hoc analysis. Regulatory bodies may require that prediction systems adhere to minimum accuracy standards across diverse operating conditions, and that they be re-certified after software updates. Furthermore, privacy regulations such as GDPR impose

constraints on collecting and storing trajectory data, requiring anonymization and consent mechanisms [21]. The design of the system should incorporate privacy-by-design principles, such as processing data locally and discarding raw sensor feeds after inference.

Finally, the integration of scene semantics raises questions about the representation of public space. Semantic maps that include commercial premises, advertising, or surveillance infrastructure may inadvertently encode biases about which areas are considered desirable or safe. Policymakers and urban planners should collaborate with technologists to ensure that scene semantic models reflect equitable public space design, rather than reinforcing existing inequalities. Open standards for scene semantic representation and public benchmarking of motion prediction systems can foster accountability and continuous improvement.

8. Conclusion

This paper presented a systems-oriented exploration of cross-modal scene semantics and graph attention networks for human motion intention prediction. We examined the architectural design choices that balance computational efficiency, prediction accuracy, and robustness in real-world deployment. The fusion of RGB, depth, and semantic modalities enriches the contextual understanding of agent motion, while graph attention mechanisms enable flexible modeling of inter-agent and agent-scene relationships. Critical trade-offs regarding sensor fusion strategy, graph construction, and temporal resolution were discussed. Deployment challenges such as real-time inference, sensor failure tolerance, and sustainable data pipelines were highlighted. Furthermore, we addressed governance and fairness considerations, emphasizing the need for equitable model performance, transparent accountability, and privacy-preserving design. As autonomous systems become increasingly embedded in socio-technical infrastructures, the cross-modal graph attention framework offers a promising path toward safer and more interpretable motion prediction. Future work should focus on large-scale empirical validation, adaptive resource management, and the development of regulatory frameworks that ensure these systems serve all members of society equitably.

References

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

2. Bartoli, F., Lisanti, G., Ballan, L., & Del Bimbo, A. (2018). Context-aware trajectory prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4681–4690).

3. Vemula, A., Muelling, K., & Oh, J. (2018). Social attention: Modeling attention in human crowds. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 4601–4607).

4. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations.

5. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., & Savarese, S. (2019). SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1349–1358).

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

7. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848.

8. Kitani, K. M., Ziebart, B. D., Bagnell, J. A., & Hebert, M. (2012). Activity forecasting. In Proceedings of the European Conference on Computer Vision (pp. 201–214).

9. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., & Savarese, S. (2019). Social-BiGAT: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In Advances in Neural Information Processing Systems (pp. 137–146).

10. Li, J., Ma, H., & Tomizuka, M. (2019). Conditional generative neural system for probabilistic trajectory prediction. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 6150–6156).

11. Liang, M., Yang, B., Chen, Y., Hu, R., & Urtasun, R. (2019). Multi-task multi-sensor fusion for 3D object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7345–7353).

12. 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.

13. Meyer, G. P., Charland, J., Hegde, D., Laddha, A., & Vallespi-Gonzalez, C. (2019). Sensor fusion for joint 3D object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1230–1237).

14. Shi, W., & Rajkumar, R. (2020). Point-GNN: Graph neural network for 3D object detection in a point cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1711–1719).

15. De Lange, M., et al. (2021). A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3366–3385.

16. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

17. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In Proceedings of the Conference on Robot Learning (pp. 1–16).

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

19. Liu, Y., & Wen, J. (2021). Cultural differences in pedestrian behavior: A cross-national study of crossing tendencies. Journal of Safety Research, 77, 152–161.

20. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 77–91).

21. European Parliament. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Official Journal of the European Union, L 119, 1–88.