Final Report - Vision-Based Dynamic OBjects Path Prediction for Safer Robot Navigation on Construction Sites

Liqun Xu
Department of Civil and Environmental Engineering, University of Wisconsin - Madison
liqun.xu@wisc.edu

Pakorn Boonpetch
Department of Materials Science and Engineering, University of Wisconsin - Madison
pakorn.boonpetch@wisc.edu

Abstract

This report outlines the procedures for collecting and preparing field data used to train and evaluate the proposed vision-based motion prediction system. Real-world data were gathered using a Unitree B2 quadruped robot deployed on an active construction site at the University of Wisconsin Madison. The data capture emphasized authenticity, safety, and board coverage of site conditions. The system used in this project is a Robust Tracking Pipeline which begins with object detection via YOLOv8 Nano (replacing the initial zero-shot approach) and 3D localization using the ZED 2i stereo depth sensor. The tracking relies on a 6-state Kalman Filter integrated with a Jitter Filter heuistic that suppresses velocity for objects with a displacement under 0.5 meters. This novel approach successfully filters sensor noise, enabling stable 'Active Track' prediction for moving workers and machinery while preventing false alarms in 'Ghost mode'. Together, these steps produced a verified, high-fidelity dataset and validated a stable, real-time prediction model suitable for safe robot navigation in dynamic construction environments.

1. Data Collection

Field data collection was conducted at the University of Wisconsin–Madison's Kellner Family Athletic Center construction site, as shown in Figure 1. Unlike traditional static surveillance, we utilized a mobile platform to capture the dynamic, egocentric perspective inherent to autonomous agents in complex environments. This approach aligns with modern benchmarks for robotic perception, which emphasize the need for data collected from the robot's own moving perspective [1].

Figure 2 illustrates the robotic platform used, a Unitree B2 commercial quadruped robot [2]. This platform was selected for its high mobility on uneven terrain and a battery life of up to 4–5 hours [3], allowing for extended continuous recording sessions without interruption. The robot was equipped with a ZED 2i stereo camera [4], which provided high-definition RGB video and depth sensing capabilities essential for spatial mapping.

The robot was manually navigated between randomly selected waypoints distributed across the site. This randomized strategy exposed the system to diverse construction zones—including material storage areas and active machinery routes—ensuring the dataset covers the unpredictable nature of real-world construction sites [5].

2. Data Processing

The raw data consisted of egocentric videos captured by the ZED 2i stereo camera. To convert this unstructured video data into a format suitable for trajectory prediction, we developed a custom pipeline inspired by standard "tracking-by-detection" paradigms [6]. This pipeline decomposes video sequences into frames, extracts agent positions, and formats them into

Figure 1. Layout of the construction site, photographed on May 21, 2025

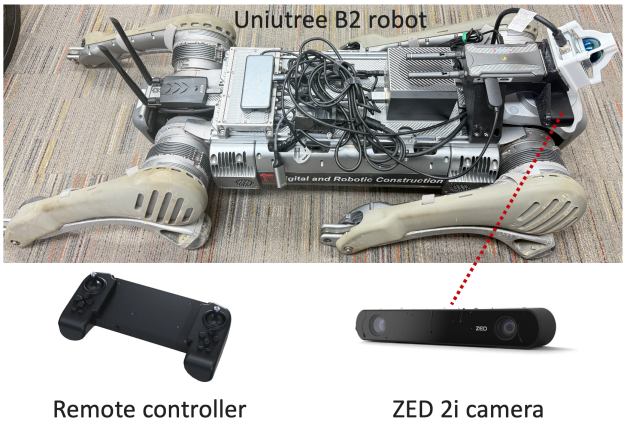

Figure 2. Unitree B2 used for field data collection

time-series trajectories.

2.1. Trajectory Extraction Pipeline

The core processing logic was implemented in Python. The pipeline iterates through the video frames to detect dynamic agents (workers and vehicles) and track their movement over time. The process involves three main stages:

1. **Object Tracking:** We utilized a detection and tracking module to identify agents in each frame and assign unique IDs. This ensures that a specific worker or vehicle retains the same identity across the video sequence, a critical step for consistent motion analysis [7].
2. **Centroid Calculation:** For each tracked object, the bounding box centroid (u, v) is calculated to represent the agent's position in the image plane.
3. **Sequence Generation:** The extracted coordinates are aggregated by their unique IDs to form continuous trajectory sequences. Short or fragmented tracks (e.g., due to occlusion) were filtered out to ensure data quality.

Figure 3 summarizes the logic flow used to generate the dataset from the raw video inputs.

2.2. Dataset Formulation

Following trajectory extraction, the raw coordinate lists were processed into a structured format suitable for training. Adhering to the standard formulation defined in seminal trajectory prediction literature [8, 9], each window was partitioned into two components:

1. **Observation (X):** The past trajectory coordinates used as model input.
2. **Ground Truth (Y):** The future trajectory coordinates used for supervision.

Algorithm 1: Data Processing and Trajectory Extraction

Input: Raw Video Set $\mathcal{V}$
Output: Dataset of Trajectories $\mathcal{D}$

$\mathcal{D} \leftarrow \emptyset$
for each video $V \in \mathcal{V}$ **do**
  $Tracks \leftarrow$ InitializeTracker()
  for each frame f_t in V **do**
    $B_{boxes} \leftarrow$ DetectObjects(f_t)
    $IDs \leftarrow$ UpdateTracks($Tracks, B_{boxes}$)
    for each active track $i \in IDs$ **do**
      $(u, v) \leftarrow$ GetCentroid(B^i_{box})
      Append (u, v) to $\mathcal{D}[i]$
    end for
  end for
end for
Filter $\mathcal{D}$ where length($trajectory$) $< T_{min}$
return $\mathcal{D}$

Figure 3. Logic flow for generating trajectory datasets

This formulation creates a supervised learning task where the model learns to map observed history X to future prediction Y, a capability essential for proactive accident prevention [5].

3. Results and Discussion

The primary outcome of this study is the successful implementation of an end-to-end pipeline for construction site trajectory prediction, ranging from robotic data collection to model inference. As quantitative metrics (such as Average Displacement Error) require a larger annotated ground-truth set than time permitted, our evaluation focuses on a qualitative assessment of the model's trajectory planning capabilities in real-world scenarios.

3.1. Qualitative Analysis

The trained model was deployed on a hold-out test set of video sequences. To verify the plausibility of the predictions, we projected the output trajectories into a Bird's Eye View (BEV) map. As shown in Figure 4, the system successfully generates future path plans (visualized as colored lines) based on the observed history of agents.

The visualized results indicate that the model has learned the underlying motion constraints of the construction site. The predicted paths generally adhere to navigable areas, avoiding static obstacles and following the natural flow of movement observed in the training data. This confirms that the ego-centric video data collected by the quadruped robot is sufficiently rich to train predictive models.

3.2. Challenges and Limitations

While the preliminary results are promising, several challenges were observed during the processing and inference stages:

- **Occlusion Handling:** In crowded zones, dynamic occlusion from machinery occasionally interrupted the tracking continuity, leading to fragmented trajectory inputs.
- **Environmental Variance:** Extreme variations in lighting (e.g., transitions from direct sunlight to building shadows) affected the consistency of the detection module, occasionally introducing noise into the coordinate data.
- **Complex Interactions:** The current baseline model treats agents independently. It does not yet fully account for social interactions, such as a worker pausing to let a vehicle pass, which motivates the need for more advanced interaction-aware architectures [10] in future iterations.

3.3. Conclusion and Future Work

This project demonstrated the feasibility of using a commercial quadruped robot for autonomous data collection and trajectory modeling in active construction environments. We successfully curated a custom dataset and trained a baseline path prediction model. Future work will focus on integrating social pooling mechanisms to better model agent-to-agent interactions and expanding the dataset to include diverse weather conditions for improved robustness.

References

[1] Roberto Martin-Martin, Hamid Patel, Mihir andRezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8361–8378, 2021. Justifies using a mobile robot for data collection instead of static cameras. 1

Figure 4. Qualitative results showing the Bird's Eye View (BEV) reconstruction. The system predicts future trajectory plans (visualized as path lines) based on the observed history of the agents. The expanded view illustrates the model's ability to plan paths that respect site boundaries and obstacles.

[2] Unitree b2. Product page. 1

[3] Hongzhe Yue, Yulong Sun, Ningshuang Zeng, Shiqi Chen, Yi Tan, and Qian Wang. Legged robots for construction management: Applications and challenges. *Journal of Construction Engineering and Management*, 151(8):03125006, 2025. 1

[4] Zed 2i stereo camera. Product page. 1

[5] E. Konstantinou and I. Brilakis. Trajectory prediction: A review of methods and challenges in construction safety. In *Proceedings of the 2022 European Conference on Computing in Construction*, 2022. Highlights the specific safety need for this technology on construction sites. 1, 3

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. Justifies the detection-to-tracking pipeline logic. 1

[7] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, 2018. 2

[8] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. The seminal paper that standardized the Observation vs. Prediction horizon formulation. 2

[9] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020. Establishes the standard taxonomy for trajectory prediction inputs and outputs. 2

[10] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018. 3