From Real-World Images to Agent-Based Crowd Simulations: An End-to-End Pipeline

Demonstration Track

Helena G. Theodoropoulou
Archimedes/Athena RC
Athens, Greece
hetheod@athenarc.gr

Vasilis Zafeiropoulos
Archimedes/Athena RC
Athens, Greece
v.zafeiropoulos@athenarc.gr

Zoi Lygizou
Archimedes/Athena RC
Athens, Greece
zoi.lygizou@athenarc.gr

Michail Zervas
Archimedes/Athena RC
Athens, Greece
michalis.zervas@athenarc.gr

Dimitris Kalles
Hellenic Open University
Patras, Greece
kalles@eap.gr

Chairi Kiourt
Archimedes/Athena RC
Athens, Greece
chairiq@athenarc.gr

ABSTRACT

This work presents an end-to-end pipeline that transforms a single RGB image into a high-fidelity crowd simulation. The framework integrates multidimensional scene perception and socio-emotional agent intelligence. The platform supports both interactive GUI and GPU-accelerated headless modes for large-scale, complex scenarios. By bridging real-world visual data with autonomous behavioral modeling, this unified, data-driven framework provides a scalable solution for advanced crowd safety research and repeatable experimentation.

KEYWORDS

Crowd Simulation; Multi-agent System; Computer Vision; Agent-Behavior Modeling; Reinforcement Learning

ACM Reference Format:
Helena G. Theodoropoulou, Vasilis Zafeiropoulos, Zoi Lygizou, Michail Zervas, Dimitris Kalles, and Chairi Kiourt. 2026. From Real-World Images to Agent-Based Crowd Simulations: An End-to-End Pipeline: Demonstration Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/TLYB5574

1 INTRODUCTION

Crowd Management (CM) is related to the analysis and prediction of large crowd behavior dynamics in complex environments, at both collective and individual level, including safety concerns [12, 14]. It is central to urban and architectural design [1], as well as large-scale event planning, and its requirements for efficient management dictate the need for highly reliable solutions, where advanced digital tools allow of the exploration of scenarios of natural or human-made disasters, to identify potential emergencies or bottlenecks, with recent studies [6, 12] focusing on the use of Artificial Intelligence and Machine Learning (ML), for real time models. Acknowledging that CSM constitutes a complex process, we present a set of mechanisms forming an integrated simulation framework.

Modern CM faces a critical gap between static visual data and the dynamic behavior required for high-fidelity simulations. Our primary contribution is an end-to-end modular pipeline[1] that automates this transition: we extract multidimensional crowd data and 3D environment models from a single RGB image. Within these spaces, we leverage Deep Reinforcement Learning (DRL) to facilitate sophisticated, human-like decision-making. This unified framework enables repeatable, scalable analysis of complex crowd dynamics across diverse scenarios.

2 PIPELINE OVERVIEW

The proposed end-to-end framework is organized into a three-layered architecture designed to bridge the gap between static visual data and dynamic agent behavior. The process begins with *Raw Scene Analysis*, where Deep Learning (DL) and Computer Vision (CV) models extract spatial and semantic data from a single RGB image. These outputs feed into *Agent Logic* and *Behavior Modeling layer*, where human data is translated into autonomous agents equipped with DRL and cognitive exchange mechanisms. Finally, the *Integrated Simulation and Evaluation layer* brings these components into a unified digital twin [5], enabling the execution of complex scenarios and the extraction of actionable safety metrics. This pipeline is depicted in Figure 1.

2.1 Layer I: Multidimensional Scene Perception

The pipeline's foundational layer transforms a single RGB image into a semantically rich 3D digital twin. In the reconstruction phase, the scene's location is identified via image-based localization [16], with 3D terrain and urban structures rendered through a georeferencing framework, Figure 1 (2a).

Concurrently, a perception-driven stack extracts granular agent data using DL and CV approaches to detect individuals [15] and estimate their age, gender, and emotional state [10]. To refine these outputs, a Vision Language Model (VLM) [3, 13] is utilized, validating initial predictions against the broader environmental context

[1]Video: https://artcogs.github.io/crowd-sim

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/TLYB5574

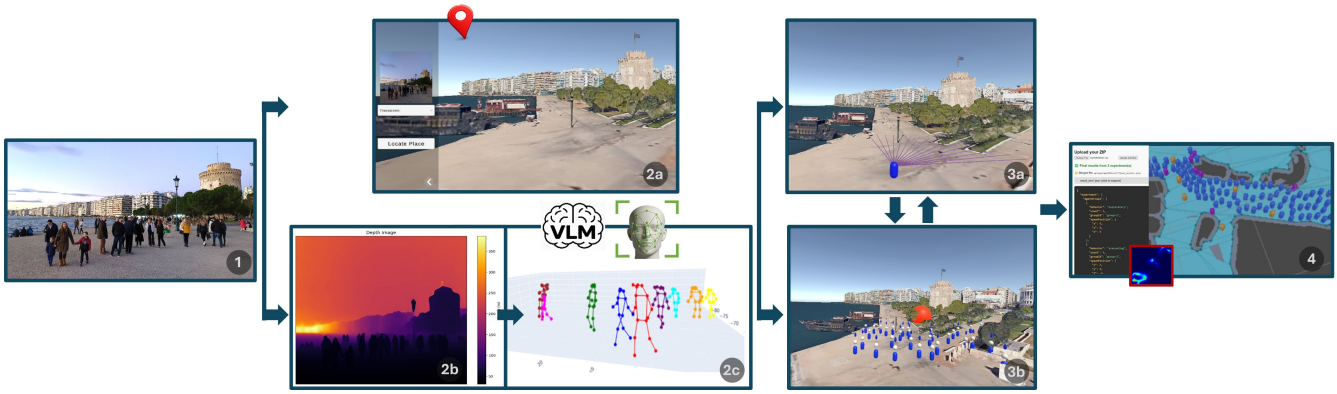

Figure 1: The end-to-end pipeline: (1) Input RGB image, *Layer I:* (2a) 3D scene reconstruction, (2b) depth estimation, and (2c) multidimensional perception (pose and emotion recognition). *Layer II:* (3a) autonomous navigation and (3b) socio-emotional behavior modeling. *Layer III:* (4) Crowd simulation scenario execution.

and providing semantic explanations for the agents' emotional states[2]. Finally, by calculating 3D pose coordinates and depth estimation [8], we determine the precise spatial positioning of each person for measuring distances between individuals, ensuring that the initial agent placement in the 3D environment accurately mirrors the real-world density and social distribution (Figure 1 (2b, 2c)).

2.2 Layer II: Agent Intelligence and Social Dynamics

In this layer, the pipeline translates perceived crowd data into a multi-agent system governed by socio-emotional dynamics [4]. Individuals are replaced by autonomous agents whose adaptable behavior is based on fear and situational knowledge. Each agent maintains an *Event Certainty Level* $[0, 1]$, which increases through environmental awareness and a social contagion mechanism that models the flow of information between neighboring agents. We model spatial knowledge based on agent profiles (e.g., staff vs. first-time visitors), while a Fear Level dynamically degrades navigation accuracy to simulate realistic panic states and evacuation dynamics [7], Figure 1 (3b). This interplay ensures that information diffusion directly impacts collective crowd behavior.

We utilize DRL as the core motion controller, specifically employing the Proximal Policy Optimization (PPO) [11] algorithm for its stability and reliability in continuous action spaces. Agents navigate via a 120° field-of-view (FOV) simulated through ray casting, as seen in Figure 1 (3a). We implement two architectures:

- **Flat RL Model:** A single end-to-end policy that maps sensory inputs directly to navigation actions for obstacle avoidance and goal-reaching.
- **Hierarchical RL Model:** A multi-level policy structure where a *low-level controller* is trained for reactive collision avoidance, while a *high-level controller* learns strategic path selection and decision-making.

This hierarchical structure enables a human-like combination of reactive maneuvering and long-term strategic planning, allowing agents to navigate uncertainty with high fidelity.

2.3 Layer III: Simulation Execution and Scenario Analysis

The final layer of the pipeline is a data-driven platform designed for complex, repeatable simulations. Developed on a state-of-the-art game engine, it exploits physics engine to ensure high-fidelity interactions and realistic environmental conditions. All simulation parameters—including the reconstructed environment, crowd composition, and agent profiles—are defined in JSON configuration files. The platform supports two distinct operational modes:

- **Interactive GUI Mode:** Real-time visualization for scenario debugging, parameter tuning, and direct behavioral observation.
- **GPU Headless Mode:** Optimized background execution for large-scale, complex scenarios by eliminating graphical overhead [9].

During runtime, the platform synthesizes the multi-layered data-integrating the 3D environmental models and agent profiles from Layer I with the cognitive and DRL-based behaviors from Layer II. Agents are instantiated and grouped by specific scenarios (e.g., emergency evacuation), see Figure 1 (4). Throughout execution, the platform records high-fidelity agent state trajectories and computes aggregated safety metrics. These results are exported as structured JSON data, facilitating both quantitative post-simulation analysis and qualitative visual validation.

3 CONCLUSIONS

This work demonstrates a modular, three-layer pipeline that significantly enhances the fidelity and scalability of agent-based crowd simulations. By integrating multidimensional scene perception with socio-emotional and DRL-based agent intelligence, we bridge the gap between static imagery and dynamic behavioral modeling. Our data-driven simulation platform exploits state-of-the-art game engine physics and a GPU-accelerated headless mode to execute complex, large-scale scenarios with minimal overhead. Ultimately, this framework provides a scalable, high-performance solution for identifying and managing simulation contexts, offering a robust tool for advanced crowd safety and urban planning research.

ACKNOWLEDGMENTS

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

REFERENCES

[1] Gideon Aschwanden, Jan Halatsch, and Gerhard Schmitt. 2008. Crowd simulation for urban planning. In *Architecture in Computro–26th eCAADe Conference Proceedings*. 493–500.

[2] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual Emotion Recognition using Large Vision Language Models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 4769–4776. https://doi.org/10.1109/IROS58592.2024.10802538

[4] Saba Khan and Zhigang Deng. 2024. Agent-based crowd simulation: an in-depth survey of determining factors for heterogeneous behavior. *The Visual Computer* 40 (06 2024), 1–12. https://doi.org/10.1007/s00371-024-03503-2

[5] Chairi Kiourt, Zoi Lygizou, Anestis Koutsoudis, and Dimitris Kalles. 2026. *A Framework for Autonomous Crowd Management Through Reinforcement Learning and Digital Twins*. Springer Nature Switzerland, Cham, 257–289. https://doi.org/10.1007/978-3-032-06364-9_11

[6] Yihao Li, Yuting Chen, Junyu Liu, and Tianyu Huang. 2025. Efficient crowd simulation in complex environment using deep reinforcement learning. *Scientific Reports* 15, 1 (2025), 5403.

[7] Mariusz Pecio. 2025. Issues of Crowd Evacuation in Panic Conditions. *Urban Science* 9, 7 (2025). https://doi.org/10.3390/urbansci9070258

[8] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. 2025. UniDepthV2: Universal Monocular Metric Depth Estimation Made Simpler. arXiv:2502.20110 [cs.CV] https://arxiv.org/abs/2502.20110

[9] Mirella Santos Pessoa de Melo, José Gomes da Silva Neto, Pedro Jorge Lima da Silva, João Marcelo Xavier Natario Teixeira, and Veronica Teichrieb. 2019. Analysis and Comparison of Robotics 3D Simulators. In *2019 21st Symposium on Virtual and Augmented Reality (SVR)*. 242–251. https://doi.org/10.1109/SVR.2019.00049

[10] Andrey Savchenko and Egor Churaev. 2025. EmotiEffLib: Library for Efficient Emotion Analysis and Facial Expression Recognition. https://github.com/sb-ai-lab/EmotiEffLib GitHub repository; accessed 2025-12-21.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[12] George Sidiropoulos, Chairi Kiourt, and Lefteris Moussiades. 2020. Crowd simulation for crisis management: The outcomes of the last decade. *Machine learning with applications* 2 (2020), 100009.

[13] Gemma Team and et al. 2025. Gemma 3 Technical Report. arXiv:2503.19786 [cs.CL] https://arxiv.org/abs/2503.19786

[14] Daniel Thalmann and Soraia Raupp Musse. 2013. *Crowd Simulation* (2nd ed.). Springer, London. https://doi.org/10.1007/978-1-4471-4450-2

[15] Ultralytics. [n.d.]. Ultralytics YOLO11 Pose Estimation Documentation. https://docs.ultralytics.com/tasks/pose/. YOLO11x-pose model; accessed 2025-12-21.

[16] Yihong Wu, Fulin Tang, and Heping Li. 2018. Image-based camera localization: an overview. *Visual Computing for Industry, Biomedicine, and Art* 1 (12 2018). https://doi.org/10.1186/s42492-018-0008-z