3D Terrain Generation for Simulation - An AI based Pipeline for Drone Imagery Processing

Yaniv Minkov, Or Zuriel, Einav Kiperman, Rami Rokach, Yinon Atzmon

Reymark Technologies LTD
Tel-Aviv, Israel
{Yaniv, Or, Einav, Rami, Yinon}@reymark.com

Abstract:
Modern simulation, training and gaming environments increasingly demand high-fidelity, geo-specific photorealistic 3D terrain models. However, traditional terrain generation is time-consuming and costly, requiring extensive manual work and specialized expertise. Advances in AI-driven neural networks now enable faster, more accurate terrain reconstruction from aerial imagery, focusing on point-cloud generation, object and material classification, and segmentation. The proposed approach builds on this body of research, improving upon existing approaches through a holistic pipeline that integrates multiple AI-driven processes. By combining pre-trained neural networks, the authors enhance terrain generation by ensuring meaningful semantic immersive representation, enabling realistic interactions between simulation assets and the environment.

The paper describes an R&D process of a pipeline which integrates a robust AI-driven process to achieve three key objectives: (1) reconstructing accurate 3D models from 2D drone imagery, (2) classifying terrain components such as vegetation, roads, and buildings, (3) optimizing the generated models for simulation engines like Unity and Unreal Engine. Unlike traditional point-cloud-based methods, this approach enhances both geometric accuracy and semantic understanding.

Additionally, physics-aware modeling allows realistic interactions within simulations - such as damaging a building or finding cover under a tree. This makes the technology ideal for defense, disaster response, and training applications, where dynamic interactions with terrain are crucial.

The authors first discuss the required features for the 3D model. Next, they describe the process of collecting and organizing the raw data, including best practices for drone imagery acquisition and preprocessing. Then, they detail how this data is converted into a 3D point-cloud. They then explore the AI techniques and algorithms for feature extraction and generation of this point cloud, such as terrain classification, vegetation and road identification, and physical interaction modeling. Finally, they highlight the challenges encountered in this research and the areas that require further exploration to refine and enhance technology.

ABOUT THE AUTHORS

Yaniv Minkov is a co-founder and CPO at Reyymark Technologies. Yaniv is a system engineer with extensive experience in initiating and managing technology-driven R&D projects, with a strong focus on virtual simulation. He specializes in leading teams, driving organizational innovation programs, and advancing research and development initiatives. Yaniv holds a B.Sc. and M.Sc. in Industrial Engineering (IS & UX) and currently heads the R&D efforts at REYYMARK Technologies.

Or Zuriel is a senior developer at Reyymark Technologies. Or holds a B.Sc. in Physics and pursues an M.Sc. focused on theoretical Physics, which involves designing a dedicated simulation framework. Over the past decade, he has independently developed GPU-accelerated applications for game engines.
Or brings exceptional expertise in GPU-based parallel computing and real-time simulation. He has extensive experience in both private and military R&D settings, including engine development for battle simulations. His work includes the design and implementation of custom rendering engines with integrated physical simulation, advanced shading systems, and ray-based techniques such as ray tracing. His deep familiarity with algorithms and AI frameworks plays a key role in the development of the AI-driven engine presented in this paper, which constructs high-fidelity 3D terrain models from aerial imagery.

Einav Kiperman is a co-founder and CTO of Reyymark Technologies, specializing in advanced simulation solutions for defense and civilian applications. Previously, Einav served as a senior system engineer and simulation team lead at Israel Aerospace Industries and held key leadership positions in the Israel Defense Forces, including Head of the Battle Lab and Head of the Center of Operational Research in the Ground Forces. With a strong background in system engineering and simulation development, Einav has led extensive simulation-based research efforts, leveraging advanced modeling techniques to evaluate operational concepts, optimize defense systems, and enhance decision-making processes. Einav brings decades of expertise in designing and implementing cutting-edge modeling and simulation technologies

Rami Rokach is a co-founder and CEO at Reyymark Technologies. Rami has spent the past two years as a key player in top defense industry companies, following 15 years in elite Israeli defense technology units, where he led multidisciplinary projects in simulations, missile defense, embedded systems, and cybersecurity. He has managed teams of 20+ developers and researchers, integrating cutting-edge technologies with proven success in both operational and commercial settings. Rami holds a B.Sc. in applied mathematics from Bar-Ilan University and an M.Sc. in statistics & operations research from Tel-Aviv University

Yinon Atzmon is a co-founder and CFO at Reyymark Technologies. Previously, he held key leadership roles in the Israel Defense Forces, including Head of the Operational Research and Battle Lab Branches, where he led research in defense systems analysis and real-time simulation development. Yinon has also founded and managed multiple technology companies, focusing on UAV mission management, GIS-based defense applications, and system-of-systems engineering. Yinon holds a bachelor's degree in industrial engineering and management from Tel Aviv University and a master's degree in business administration with a focus on Operations Research from Ben-Gurion University.

Automating 3D Terrain Generation for Simulation - An AI based Pipeline for Drone Imagery Processing

Yaniv Minkov, Or Zuriel, Einav Kiperman, Rami Rokach, Yinon Atzmon

Reymark Technologies LTD
Tel-Aviv, Israel
{Yaniv, Or, Einav, Rami, Yinon}@reymark.com

INTRODUCTION

Modern simulation and gaming environments demand rapid and affordable generation of high-fidelity, geo-specific 3D terrains. These virtual environments are increasingly used for training, mission rehearsal, and serious gaming applications, yet the production of such assets remains slow, expensive, and heavily dependent on expert manual intervention. Recent advances in artificial intelligence, particularly in the domains of image segmentation and object classification, have opened new opportunities for automating terrain reconstruction from drone imagery. These advancements significantly reduce the time and cost associated with traditional (manual) modeling workflows, while simultaneously increasing semantic accuracy and readiness for simulation applications (Lam et al., 2023; Minaee et al., 2020; Trigka & Dritsas, 2025; Wang et al., 2023). Deep learning architectures such as fully convolutional networks and transformer-based segmentation models have demonstrated high performance in complex scene understanding tasks (Minaee et al., 2020; Wang et al., 2023), and the integration of these techniques into UAV-based workflows is already showing practical benefits in synthetic environment generation and real-time geospatial modeling (Lam et al., 2023; Trigka & Dritsas, 2025).

Motivation

Traditional modeling pipelines, including those based on photogrammetry, remain constrained by several critical limitations - particularly in terms of processing time, scalability, and semantic consistency. Scianna et al. (2020) highlight that while photogrammetry and terrestrial laser scanning can yield accurate geometric models, these methods require extensive manual post-processing, including cleanup, texturing, and annotation, making them time-consuming and labor-intensive. Chen et al. (2019) propose a framework for segmenting photogrammetric-generated point clouds and extracting object information to create realistic virtual environments for simulations and training. Furthermore, Chen et al. (2022) emphasize that traditional photogrammetric workflows are difficult to scale to large urban areas and often lack consistent semantic labeling, motivating the creation of hybrid synthetic datasets. Recent datasets such as STPLS3D (Chen et al., 2022) provide a valuable benchmark for evaluating semantic segmentation algorithms on large-scale photogrammetric reconstructions. However, these datasets are primarily designed for offline accuracy assessment under controlled conditions and do not directly support the development of operational pipelines for real-time interaction or integration with simulation platforms. Hu et al. (2022) further point out that although photogrammetry provides detailed geometry, it typically lacks rich, reliable semantic information due to the irregularity and noise inherent in manually generated point clouds. Collectively, these findings underscore the need for automated or AI-assisted alternatives that can deliver high-fidelity, semantically rich terrain models at scale.

Training simulations require environments not only to look realistic but also to behave meaningfully - e.g., allowing avatars to take cover behind vegetation or interact physically with infrastructure. To support these needs, the research goal was to develop an **end-to-end AI-enhanced pipeline** that transforms 2D drone imagery into simulation-ready, semantically segmented, and interactable 3D models. Another research goal was to enable an **independent pipeline** that can rely on self-produced drone imagery with no necessary additional GIS information.

Paper Contribution

This paper presents a scalable, modular pipeline that:

- Reconstructs high-resolution 3D terrains from drone-captured imagery using AI-driven classification.
- Automatically segments and classifies physical and man-made terrain features (e.g., vegetation, roads, buildings, windows).
- Integrates semantically segmented data into 3D functional models for simulation and gaming engines.

- Reduces human involvement and processing time by combining pre-trained neural models with custom heuristics.
- Enables physically plausible interactions within training simulations, compatible and optimal for common simulation and gaming engines.
- Produces affordable digital twins with no need for human intervention.

BACKGROUND AND RELATED WORK

The growing demand for immersive, geo-specific simulation environments has driven significant progress in the generation of digital twins - virtual replicas of physical terrains and structures. These systems are now essential for a wide range of applications, including defense training, disaster preparedness, and large-scale simulation-based urban planning. To fulfill these needs, we need to find a way to use novel technologies that might produce more affordable near-real-time geo-specific, photo-realistic, simulation-ready 3d models.

Digital-Twin Generation Pipelines

Traditional digital twin generation methods have relied on traditional 3D modeling or semi-automated photogrammetric workflows. While effective, these approaches require domain expertise, are labor-intensive, and thus, lack scalability. Human effort is still required for object classification and segmentation. Recent research has focused on automating these processes through Structure-from-Motion (SfM), LiDAR fusion, and aerial imagery processing to produce terrain-aligned 3D reconstructions (Arrigoni, 2025; Özyeşil, Voroninski, Basri, & Singer, 2017), but they still lack the semantic classification phase that makes the 3D model ready for simulation.

Several initiatives have demonstrated the feasibility of automating terrain generation for simulation. For instance, the One World Terrain (OWT) initiative (U.S. Army PEO STRI, 2020) leverages drone and satellite imagery to rapidly produce simulation-ready terrain tiles. Case studies from past I/ITSEC conferences have described semi-automated urban model creation using UAS footage and object detection. For example, Spicer et al. (2016) demonstrated a process for generating georeferenced 3D terrain and building models from commercial multirotor drone footage, including point cloud segmentation and export to simulation-ready formats. These efforts established foundational techniques for extracting terrain geometry and structural footprints directly from aerial data. Similar methodologies continue to evolve, leveraging AI-based object detection frameworks for extracting semantic elements from aerial imagery for various applications including traffic analytics and understanding (Benjdira et al., 2023).

AI-Based Aerial-Image Segmentation and Classification

Artificial intelligence has become a central enabler of scalable terrain modeling, particularly through advances in semantic segmentation of overhead imagery. Deep convolutional neural networks (CNNs) have demonstrated high accuracy in extracting land features such as roads, buildings, vegetation, and water bodies. For example, Liu et al. (2024) provide a comprehensive review showing that CNN-based models like U-Net and DeepLabV3+ have achieved state-of-the-art performance in road extraction tasks using high-resolution remote sensing data. Similarly, Elgamily et al. (2024) introduced a novel CNN architecture (W13) capable of effectively segmenting multiple land cover types, including built structures and natural features. Expanding on these approaches, Yamazaki et al. (2023) proposed a transformer-based architecture – Aerial Former - that leverages global spatial relationships to improve semantic segmentation in complex aerial scenes, outperforming traditional CNNs in detecting structural and vegetative elements. Together, these developments demonstrate how AI-driven methods—spanning CNNs to transformers—are instrumental in producing semantically rich terrain models suitable for simulation and analysis at scale.

Benchmarks such as ISPRS Vaihingen and Inria Aerial Image Labeling datasets have catalyzed progress in pixel-wise labeling accuracy (Audebert et al., 2018). Yet, high classification accuracy alone is insufficient for simulation contexts. Models must support consistent geometric representation, object continuity, and physical plausibility to integrate with real-time engines and allow interaction between entities and their environment.

Recent pipelines have addressed the longstanding disconnect between perceptual understanding and simulation logic by fusing semantic outputs with procedural modeling rules. This integration allows for the automated instantiation of simulation-ready assets that exhibit meaningful behavior in context, thereby reducing manual workload and enhancing realism in interactive virtual worlds. Somanath et al. (2023) describe a semi-automated workflow for urban digital twin generation that leverages semantic building data and extends it procedurally to include terrain, vegetation, and infrastructure components within game engines such as Unreal Engine. Similarly, Yaghi et al. (2025) introduce

3DGENie, a framework that combines procedural layout generation and AI-driven semantic labeling to synthesize realistic virtual environments in the form of annotated 3D point clouds, enabling scalable simulation asset creation.

REQUIREMENTS DEFINITION

The development of a terrain-generation pipeline for simulation purposes must begin with a clear understanding of end-user needs, operational contexts, and system constraints. This section outlines the fidelity requirements, feature prioritization, and performance targets that guided the design of the proposed automated solution.

End-User Training Scenarios and Fidelity Needs

Simulation users - particularly in military, emergency response, and urban planning domains - require environments that are not only visually realistic but also structurally and behaviorally accurate. For example, a soldier in a tactical training simulator must be able to take cover behind a tree or enter a building whose geometry reflects the real-world counterpart. This demands consistent geometric resolution, accurate object placement, and materials that support physical interaction (e.g., damage, occlusion, cover). High-fidelity environments are especially critical in multiple training contexts that demand spatial realism and situational detail:

- Urban operations training benefits from 3D models of buildings, alleys, and road networks that reflect the structural complexity and tactical constraints of real-world environments (Champney et al., 2017; Calian, 2022).
- In rural and forested terrains, the ability to distinguish between vegetation types, terrain morphology, and elevation gradients is essential and field navigation (Zürcher et al., 2023; Le et al., 2024).
- For disaster response rehearsal, virtual environments must include realistic representations of collapsible structures, blocked or flooded roads, and unstable terrain to effectively prepare trainees for chaotic and high-risk scenarios (Alshowair et al., 2024; Le et al., 2024).

Reference scenario

To rigorously evaluate the fidelity and operational utility of the algorithmically generated 3D models, we established a comprehensive reference scenario. This scenario is designed to represent a challenging and realistic urban environment, spanning an area of approximately one square kilometer. The terrain features a heterogeneous mix of structures, with building heights ranging from one to five stories, simulating a typical blend of residential and commercial districts. Furthermore, the environment is realistically populated with diverse vegetation, including dense tree canopies, bushes, and grassy areas, which present significant challenges for line-of-sight (LOS) calculations, sensor modeling, and path-finding algorithms. To replicate a living urban ecosystem, the scenario is populated with both static and dynamic entities. This includes parked and moving vehicles, as well as pedestrian models exhibiting standard patterns of movement consistent with daily city life. This high-fidelity reference environment serves as a critical benchmark for validating the semantic classification and overall fitness of the generated 3D models for advanced training, mission planning, and simulation applications.

Feature Set and Prioritization

Based on stakeholder input and operational analysis, the following feature classes were prioritized for automated extraction and modeling:

- **Terrain (DTM)**: high-resolution elevation models.
- **Vegetation**: tree species height, canopy size, density maps for grass and shrubs.
- **Buildings**: volumetric structure, texture, number of floors, and windows position and placement.
- **Windows**: explicit geometry enabling line-of-sight queries and tactical behavior in relation to indoor-outdoor interaction.
- **Roads**: navigable surfaces, intersections, roads and their characteristics such as width, lanes etc.
- **Vehicles and obstacles**: detection and model positioning for vehicles and other man-made items.

Each feature type must be detected, segmented, and instantiated with enough precision to enable procedural behavior, pathfinding, and line-of-sight calculations in the simulation environment.

Performance and Affordability Targets

To ensure scalability and adoption, the pipeline was designed with three critical performance goals:

- **Minimal manual intervention**: Fully automated processing from raw imagery to simulation-ready tiles.
- **Short turnaround time**: Complete generation of a 1 km² area within a few hours on a high-end workstation (Processor: Intel Core i9 14900, RAM: 32GB, Storage: 1TB, GPU: Nvidia RTX 4070)

- **Cost-effective operation:** Use of commercially available drones and open-source or in-house processing tools where possible during the production process.

These objectives reflect the practical constraints faced by end-users who must deploy and adapt simulation environments rapidly - sometimes in theater or field conditions - without access to large modeling teams or cloud infrastructure.

METHODOLOGY

The proposed pipeline is designed as a modular and fully automated process that converts 2D drone imagery into semantically rich, simulation-ready 3D environments. The pipeline consists of six main stages (see Figure 1): (1) data acquisition and preprocessing, (2) point-cloud generation (3) DTM generation, (4) semantic segmentation and feature classification, (5) procedural model instantiation, and (6) export to real-time simulation engines. Figure 1 presents an overview of the process.

Data Acquisition and Preprocessing

The pipeline begins with drone flights conducted under a standardized protocol to ensure sufficient coverage, angle and point-of-view variety, overlapping, and lighting conditions for photogrammetric reconstruction. The acquired imagery is aligned and georeferenced using off-the-shelf software and self-developed scripts, generating a dense point cloud. In order to optimize performance and flexibility, this raw data is then organized into uniformly sized tiles (e.g., 128m x 128m), each containing both imagery and geometric content.

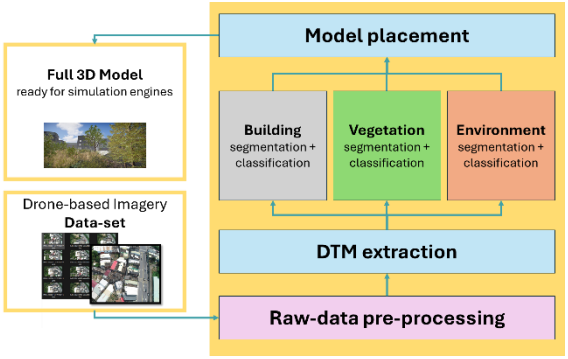

Figure 1. High-level pipeline flow-chart

Point-Cloud and DTM Generation

Next, the point cloud is classified into ground and non-ground segments using Reality Capture AI classification tool (Reality Capture insights ai classifier, n.d.). Ground points are processed to extract a high-resolution digital terrain model (DTM) using custom GPU kernels that perform elevation rendering, spike filtering, and interpolation over gaps. Simultaneously, an orthographic color image is rendered from the top-down view of the point cloud, using a dedicated C# Unity algorithm. These outputs form the basis for subsequent semantic analysis.

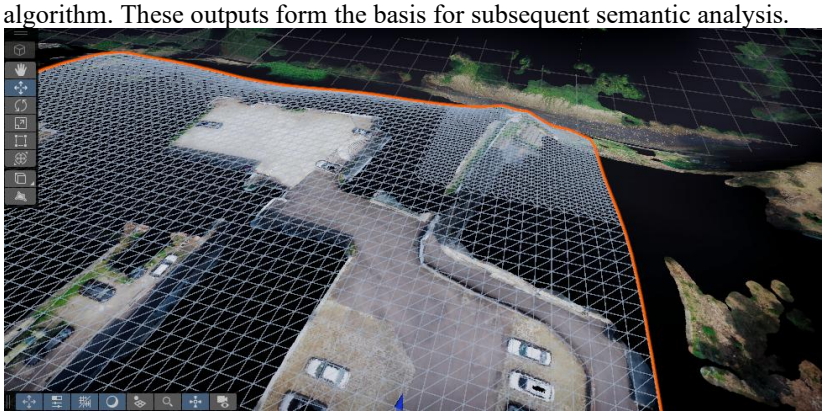

Figure 2. DTM (e.g. Unity terrain) was extracted and optimized for Unity 3d.

Semantic Segmentation and Classification Pipeline

Semantic segmentation is applied to identify key terrain components - vegetation, buildings, roads, and windows—through a combination of AI models and heuristic filters. The main features are listed here while a detailed example is presented next.

- **Vegetation**: A vegetation height map is derived from the non-ground point cloud and color orthoimage, enabling the differentiation of trees, thickets, and grasslands.

Figure 3. Vegetation was planted according to a "planting map" that was derived from the point cloud using AI models.

- **Buildings**: A heat map classifier and object segmentation model isolate building structures. Contours are extracted and regularized, then triangulated into 3D meshes. These are textured via reprojection of the original raw images and further refined to support cutouts for windows.

Figure 4. building structures before being textured based on raw images.

- **Windows**: Virtual façade renderings of buildings are processed by a specialized model that detects window boundaries (for more details see below). These are used to insert holes into the mesh, followed by placement of window instances. Identifying windows enables not only to place window models but also to estimate the building floors.

Figure 5. Building models, with "window-holes" that were "dug" out of the original building model, were populated with window structures and marked by floor with red lines.

- **Roads**: Large-area orthoimages are processed by a road-detection model to extract a road graph (nodes and edges), which is trimmed and merged across tile boundaries to ensure continuity.
- **Cars**: Private vehicles displayed in the raw images are classified, segmented, color-tagged, and then, using a vehicle "planting map" planted as 3D-colored models according to the analyzed color.

Process Summary

Figure 6 presents an example of a full 3D terrain outcome derived by the suggested technology in comparison to a reference Google Street View image. Table 1 summarizes the automated processing pipeline for generating the various features of the 3D model. It details how each feature is derived from raw input data and converted into a functional object within the simulation environment.

Figure 6. Full scene comparison:
3D generated model (left) vs. drone image provided by Google Street View (right)

Table 1. Feature Processing Summary

Feature Type	Input	Processing Steps	Output	Simulation Function
Terrain (DTM)	Point cloud	Depth rendering → Gap filling → Elevation rasterization	elevation map (e.g. Unity terrain)	Terrain collision, navigation
Vegetation	Color ortho + non-ground point cloud	AI Semantic classification → Height map → Canopy detection	Tree models with position, height, species ID	Visual cover, occlusion, physics
Buildings	Orthographic image + point normal + raw images	AI Semantic classification → Footprint detection → Mesh generation → Texture projection	Textured 3D models (.fbx)	Enterable structures, line-of-sight calculations
Windows	Rendered façade views	AI window detection → Boolean mesh cutting	Building models with embedded window geometry	Entry points, aiming visibility
Roads	Large area ortho image	AI road graph detection → Vector connectivity → Integration with DTM	Road graph (.json) + road model	Pathfinding, vehicle mobility

Detailed example - window reconstruction

As part of the ongoing effort within the urban modeling ecosystem, one of the key innovations in the proposed technology is the development of an automated pipeline for detecting and placing windows within 3D building models, using synthetically rendered façade images generated from photogrammetric mesh tiles. Figure 7 shows a flowchart of the process that begins with virtual planning of optimal façade viewpoints for each building tile. A proprietary algorithm, developed in Unity (C#), generates a virtual "photo shoot" plan, including orthographic perspective matrices and transformation matrices between camera space and world coordinates. Each tile's photogrammetric mesh is then imported from EPIC Reality Capture into Unity, where the pipeline renders façade images from each of the predefined virtual viewpoints. These rendered images are subsequently processed by a window detection AI model

(based on lck1201/win_det_heatmaps), which produces annotation files (in JSON format) containing the pixel-level coordinates of each detected window. The AI model is enhanced through retraining to support the detection of shaded or occluded windows. Next, the 2D annotations are projected back onto the 3D mesh to identify the corresponding 3D locations of the windows. Using the CGAL 3D Boolean Operation library (integrated into Unity via a wrapper), boolean operations are performed to cut accurate openings in the original 3D building geometry. Finally, modular window models are placed into these cutouts based on the location, orientation, and size derived from the detection phase.

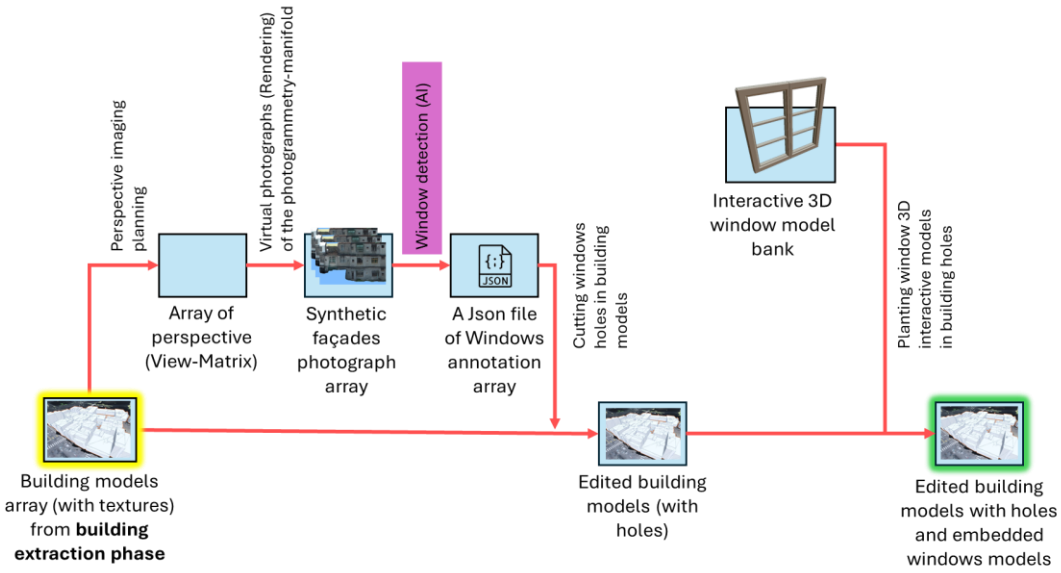

Figure 7. A flowchart describing the windows reconstruction process. Yellow shaded block marks the beginning of the process while a green shaded block marks its end.

Model adjustment for Window Detection in Heterogeneous Environments

In the effort to develop an autonomous object detection capability, the starting point was a pre-trained computer vision model designed to identify windows in buildings (see previous paragraph). An initial evaluation of this model on a primary dataset (hereafter "the Original" Dataset), which comprised images of conventional structures, yielded satisfactory performance with a Detection Rate of approximately 80%. However, when the model was challenged with a second dataset (hereafter "Dataset B"), collected in a different operational environment with distinct visual characteristics, it experienced a dramatic and unacceptable decline in performance (~40%).

Figure 8. Heterogeneous Environment datasets: original (left) and B (right)

An in-depth analysis of the failures revealed that the visual characteristics of the "windows" in Dataset B were fundamentally different from those in the Original Dataset. In fact, the structures in this dataset did not contain windows in the classic sense (with frames, glass, and reflections), but rather empty openings in the building's walls.

The primary visual cue for their detection was the deep shadow cast within the structure - a feature the original model was not trained to identify effectively.
To address this challenge and improve the model's generalization ability, we retrained it using an extended dataset comprising both the original training data ("zju_facade_jcst2020") and additional annotated images collected from our maps. The first step involved a comprehensive manual annotation process for Dataset B. Subsequently, the model was retrained using this new data, allocating approximately 80% of the annotated images for training and the remainder for validation.
The results, detailed in Figure 9, demonstrated the effectiveness of the training. The Retrained model achieved a detection rate of approximately 80% on Dataset B, in comparison to 40% before the retraining, thereby successfully overcoming the initial challenge. Furthermore, this additional training process also improved the model's performance on the original validation set, increasing its accuracy from 80% to approximately 82%. This improvement indicates that the model not only learned to identify a new type of "window" but also strengthened and generalized its understanding of the essential features that define openings in structures overall.

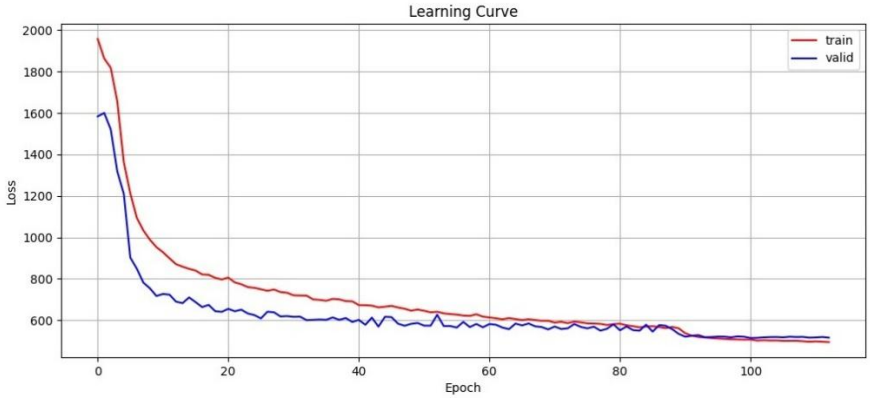

Figure 9. The learning process results over time for the full dataset ("Original" + "B")

EVALUATION & RESULTS

To evaluate the effectiveness of the proposed terrain generation pipeline, both qualitative and quantitative assessments were conducted focusing on three key metrics: (1) runtime performance, (2) semantic and geometric fidelity, and (3) operational efficiency compared to traditional workflows.

Runtime Performance

Benchmark tests were conducted on a graphic workstation (Intel Core i9 14900 CPU, 32GB RAM, RTX 4070 GPU). Processing time for a 1 km² area - including point-cloud generation, segmentation, classification, and model instantiation - ranged from 3 to 4 hours, depending on scene area, building area complexity and vegetation density.

Table 2. Feature and total Processing Summary (using the proposed pipeline)

Task Stage	Average Runtime (1 km²)
Drone data alignment & point cloud	30–45 minutes
Terrain classification (DTM)	5 minutes
Vegetation reconstruction*	15-20 minutes
Building shape reconstruction*	15-20 minutes
Building texture reconstruction*	30-40 minutes
Window reconstruction*	30-40 minutes
Road reconstruction*	10-15 minutes
Vehicle reconstruction*	5-10 minutes
Total**	~2.5–4 hours

* Each reconstruction phase includes 2 steps: Semantic segmentation and classification phase and model generation and planting.
** The total predicted processing time for 1 km^2 using traditional pipeline is at least 1 week.

Evaluation Metrics for Building Identification Accuracy
A key feature of the pipeline is the generation of 3D building models based on drone imagery. To assess the performance of the building identification process, we compared the automatically produced model with a manually annotated ground truth dataset. Two evaluation metrics were used to quantify detection accuracy. The first is the well-established F1-score (Müller et al., 2022), computed at the pixel level by comparing binary building masks of the model and the reference, all on an orthographic image. The second is a custom-designed metric we propose, which compares the alignment between building-related masks produced by the SAM (Segment Anything Model) segmentation model (Kirillov et al., 2023) and those derived from the classification output and from the ground truth data. This mask-level metric is computed as one minus the normalized symmetric difference between the building-related SAM masks in the result and the ground truth. This approach captures structural misalignment beyond pixel-wise overlap. In a representative test case using a U-Net classifier, the F1-score reached 0.95 and the SAM mask compatibility score reached 0.884. These quantitative results were complemented by a manual evaluation of key dataset attributes - such as building color contrast, spatial density, average size, surface coverage, and occlusions - which can help contextualize classification performance.

While the Pixel-Level F1-Score is widely described and used, a brief explanation for SAM Mask Compatibility Score is described together with a comparison table between these two. The purpose of the SAM Mask Compatibility Score is to evaluate how well the segmented building masks from the model (e.g., from a SAM-based classifier) align with the annotated building masks in the ground truth, focusing on object-level agreement rather than pixel overlap. The formulas are described in Figure 11, where:

- A stands for a set of predicted building-related masks (a SAM mask that overlaps with the classification result by more than 50% at the pixel level, (see Figure 12) predicted by the model, and
- B stands for a set of building-related masks (SAM masks that overlap with the ground truth). This represents one minus the normalized symmetric difference between the sets of masks, capturing mismatches in detected structures rather than pixel alignment.

$$\text{Mask Compatibility} = 1 - \frac{|A \setminus B| + |B \setminus A|}{\max(|A|, |B|)}$$

Figure 11.
SAM Mask Compatibility Score formula

Figure 12.
Building-related SAM mask (left), ground truth masks (middle) and classification masks (right)

Table 3. Key Differences Between the Metrics

Aspect	Pixel-Level F1-Score	SAM Mask Compatibility Score
Unit of Analysis	Individual pixels	Whole object masks (instances)
Sensitivity to outlines inaccuracies and continuity artifacts	High – every pixel counts	Lower – only mask presence/absence is considered
Best suited for	Measuring spatial precision	Assessing object-level detection consistency
Comparison Type	Binary mask overlap	Set difference between detected and expected masks

Semantic Fidelity
The outputs were evaluated using visual inspection and reference ground truths. Key observations:

- **Road extraction feature** maintained full topological connectivity across adjacent tiles with minimal false positives (less than 20%).
- **Window detection** feature correctly identified >70% of visible windows in complex facades using AI models trained on synthetic datasets.

Operational Efficiency
Compared to traditional 3D modeling or commercial photogrammetry workflows that were discussed with design partners, the proposed pipeline achieved:

- **Time savings** of up to 90% per km² (see Table 2.)
- **Dramatically Cost reduction** by minimizing the need for skilled 3D artists

These improvements make the pipeline viable for use in scenarios requiring rapid environment generation, such as live exercise rehearsal, disaster response planning, or dynamic mission briefing simulations.

DISCUSSION

The development of an AI-enhanced terrain generation pipeline introduced several technical and operational challenges. The following section discusses the key lessons learned, the limitations observed, and considerations for future improvements.

Observed Challenges
The Tension Between High-Fidelity Reconstruction and Perceptual Realism - A central challenge in the development of the proposed pipeline is navigating the inherent tension between achieving high-fidelity photorealism and ensuring a compelling, immersive user experience. The primary objective of photorealism is to produce a model that is a geometrically and texturally accurate representation of the real-world environment. This requires faithfully replicating not only macro-features such as vegetation, buildings, and road networks, but also fine-grained sub-features like specific window and door types, facade textures, solar water heaters, stairways, and balconies.
Conversely, achieving a high degree of immersion - the subjective sense of presence and authenticity—presents a different set of challenges. While modern Generative AI (GAI) techniques excel at creating visually plausible and aesthetically pleasing generic assets (Ramesh et al., 2022), their capability to reconstruct specific, real-world objects with high fidelity is still limited and computationally intensive. This creates a critical decision point within the pipeline: for each feature class, we must strategically prioritize between absolute faithfulness to the source data (photorealism) and the perceived realism that contributes to an effective simulation (immersion). This balancing act is a foundational aspect of the proposed methodology, guiding the development and integration of each component in the 3D environment.
Balancing AI with Heuristics - While neural networks excel in feature recognition, their outputs are often noisy, unexpected or overfitted when applied to novel terrains or extreme lighting conditions. For example, using an AI model for building classification did a great job on certain datasets while revealing very poor results on others. Integrating heuristic post-processing - such as normal vector analysis or shape regularization - was crucial for improving structural consistency, especially for building outlines and road graphs.
Data Sparsity and Occlusion - Vegetation-heavy areas and narrow urban alleys frequently yielded sparse or incomplete point clouds. This impacted both elevation, accuracy and object detection. GPU-based gap-filling and interpolation shaders provided partial remedies, but performance still degraded in highly occluding scenes.
Cross-Tile Continuity - Ensuring geometric and semantic continuity across tile boundaries (e.g., roads, rivers, terrain slopes) was non-trivial. Without careful stitching and graph merging, visual artifacts and simulation logic breaks could occur. This required developing specialized logic to merge adjacent tile metadata during export.
Window Geometry Insertion - Embedding windows into 3D building meshes demanded robust operations - often brittle when applied to complex or irregular geometry. Achieving clean cuts and alignment required careful coordination between AI-detected window bounding boxes and the 3D geometry pipeline.

Lessons Learned on AI/Heuristic Synergy
The most stable results were obtained through hybrid pipelines: using AI for coarse classification and detection, and heuristics for refinement, regularization, and final placement. For example, vegetation detection combined a trained classifier with rule-based filtering based on canopy height and normal orientation.
This synergy also supports failover mechanisms: if the users understand that an AI model underperforms in a certain area, they can set key parameters of the dataset into the pipeline which will lead to an automated choice of the optimal logic to be used.

Limitations

- **Model Generalization**: Pretrained models, especially for roofs, windows and cars, were sensitive to geographic or architectural domain shifts. Fine-tuning on domain-specific datasets is still required.
- **Small Vegetation and Surface Types**: Current classification models do not distinguish between fine-grained ground covers (e.g., dry grass, sand, pavement), limiting realism for close-up scenarios. Model training sessions could be helpful here and are part of the project's future roadmap.
- **Cloud/Weather Conditions**: Shadows and lighting variations introduce noise to photogrammetric reconstructions and image-based segmentation, requiring future compensation or correction techniques.

Strategic Insight: Automation Enables Rapid Operational Flexibility

Perhaps the most significant takeaway is that automated generation not only reduces cost but transforms how simulation assets can be used. Environments that once required fixed locations and long lead times can now be generated in-theater, tailored to mission-specific needs, or even adapted in near-real-time to reflect evolving intelligence. This flexibility is especially critical for domains such as:

- **Disaster response** (e.g., simulating flooding in a newly affected region)
- **Military rehearsal** (e.g., An operation model or simulation prior to field operations)
- **Civil planning** (e.g., stakeholder engagement with realistic urban models)

CONCLUSION AND FUTURE WORK

Conclusions

This paper presented a fully automated, AI-enhanced pipeline for generating photorealistic, simulation-ready 3D terrain environments from drone imagery. The proposed technology addresses the growing need for rapid, affordable, and semantically rich virtual environments for training, planning, and operational rehearsal. By combining photogrammetry, semantic segmentation, procedural modeling, and integration with real-time engines, the pipeline significantly reduces the time and expertise required to produce high-fidelity digital twins. Through hybrid use of AI models and geometric heuristics it enables robust performance across diverse terrain types.

Evaluation results demonstrated substantial runtime reduction, reasonable classification accuracy, and seamless engine compatibility with common visual engines. The pipeline’s modular architecture supports future model integration, scalable deployment, and its roadmap outlines future directions in surface classification, interactive logic, and operational integration.

Ultimately, this work transforms 3D environment generation from a static, labor-intensive task into a dynamic and responsive capability - empowering simulation users to create mission-specific worlds at unprecedented speed and realism.

Future work

Future development of the proposed pipeline will advance along several key vectors.
First, algorithmic enhancements will focus on improving the precision of object classification and increasing the number and granularity of detected features - including finer distinctions between vegetation types, building sub-components, and surface materials. Second, the team will continue the effort to expand and refine the evaluation framework by incorporating additional performance metrics for both semantic and geometric accuracy, enabling more rigorous model validation and benchmarking. Third, the pipeline will be extended to support seamless integration with additional simulation and gaming platforms such as Unreal Engine, VBS, and others, broadening its operational applicability. In parallel, field-based experimentation will be conducted to evaluate real-world performance under diverse environmental and operational conditions. Lastly, the integration of complementary geospatial inputs — including GIS layers, vector maps, and additional imagery sources - will further enrich the pipeline’s input space and support more robust terrain modeling across varied domains.

ACKNOWLEDGEMENTS

We wish to thank Yoram Bentzur and Oded Zelinger from B-Design3D for working with us on defining the requirements and the current semi-automated pipelines.

REFERENCES

Alshowair, A., Bail, J., AlSuwailem, F., Mostafa, A., & Abdel-Azeem, A. (2024). Use of virtual reality exercises in disaster preparedness training: A scoping review. *Journal of Emergency Nursing*. https://pubmed.ncbi.nlm.nih.gov/38623475/

Arrigoni, F. (2025). *A taxonomy of Structure from Motion methods*. arXiv. https://arxiv.org/abs/2505.15814

Audebert, N., Le Saux, B., & Lefèvre, S. (2016). Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks. arXiv. https://arxiv.org/abs/1609.06846

Benjdira, B., Bazi, Y., Koubaa, A., & Alajlan, N. (2023). TAU: A framework for video-based traffic analytics leveraging AI and UAS. *Journal of Intelligent & Robotic Systems, 108*(1), Article 24. https://doi.org/10.1007/s10846-023-01923-6

Calian. (2022). *Training for urban warfare—Multi-dimension complexity*. https://www.calian.com/resources/blogs/training-for-urban-warfare-multi-dimension-complexity/

Capturing Reality. (n.d.). *RealityCapture tools: AI classifier*. Retrieved June 15, 2025, from https://www.realitycapture-training.com/en/2021/09/01/realitycapture_insights_ai_classifier/

Champney, R., Stanney, K. M., Milham, L., & Carroll, M. (2017). *An examination of virtual environment training fidelity on training effectiveness*. Design Interactive, Inc. https://www.researchgate.net/publication/316612165

Chen, M., Feng, A., McCullough, K., Prasad, P. B., McAlinden, R., Soibelman, L., & Enloe, M. (2019 a). Fully automated photogrammetric data segmentation and object information extraction approach for creating simulation terrain. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2019&AbID=27816&CID=48

Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., McCullough, K., Ren, F., & Soibelman, L. (2022). *STPLS3D: A large-scale synthetic and real aerial photogrammetry 3D point cloud dataset*. arXiv. https://arxiv.org/abs/2203.09065

Elgamily, K. M., Mohamed, M. A., Abou-Taleb, A. M., & Ata, M. M. (2024). A novel W13 deep CNN structure for improved semantic segmentation of multiple objects in remote sensing imagery. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-024-10765-3

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., & Markham, A. (2022). *SensatUrban: Learning semantics from urban-scale photogrammetric point clouds*. arXiv. https://arxiv.org/abs/2201.04494

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment anything*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 4015–4026).

Lam, T., Reilly, M., Ramos, P., York, H., Burford, C., Shiflett, S., & Larrieu, A. (2023). Analyzing, preparing, and processing input geospatial data for high-resolution terrain generation. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2023&AbID=121255&CID=1001#View

Le, F., Guo, X., & Wang, Y. (2024). Advancing safety and efficiency training goals: The development of a virtual reality rescue training system for forest fires. *Simulation & Gaming*. https://doi.org/10.1177/14727978241299703

Li, C.-K., Zhang, H.-X., Liu, J.-X., Zhang, Y.-Q., Zou, S.-C., & Fang, Y.-T. (2020). Window detection in facades using heatmap fusion. *Journal of Computer Science and Technology, 35*(4), 900–912. https://doi.org/10.1007/s11390-020-0253-4

Liu, R., Wu, J., Lu, W., Miao, Q., Zhang, H., Liu, X., Lu, Z., & Li, L. (2024). A review of deep learning-based methods for road extraction from high-resolution remote sensing images. *Remote Sensing, 16*(12), 2056. https://www.mdpi.com/2072-4292/16/12/2056

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2020). *Image segmentation using deep learning: A survey*. arXiv. https://arxiv.org/abs/2001.05566

Müller, D., Soto-Rey, I., & Kramer, F. (2022). *Towards a Guideline for Evaluation Metrics in Medical Image Segmentation*. arXiv.

OpenDroneMap. (n.d.). *The open-source photogrammetry toolkit*. Retrieved May 2025, from https://www.opendronemap.org/

Özyeşil, O., Voroninski, V., Basri, R., & Singer, A. (2017). A survey of structure from motion. *Acta Numerica, 26*, 305–364. https://arxiv.org/abs/1701.08493

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP latents*. arXiv. https://doi.org/10.48550/arXiv.2204.06125

Somanath, S., Naserentin, V., Eleftheriou, O., Sjölie, D., Stahre Wästberg, B., & Logg, A. (2023). *On procedural urban digital twin generation and visualization of large scale data*. arXiv. https://arxiv.org/abs/2305.02242

Spicer, T., Hamlett, D., Evans, T., & Chambers, J. (2016). Producing usable simulation terrain data from UAS-collected imagery. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. https://www.iitsec.org/-/media/sites/iitsec/Proceedings/2016/2016-IITSEC-11216.ashx

Trigka, M., & Dritsas, E. (2025). A comprehensive survey of machine learning techniques and models for object detection. *Sensors, 25*(1), 214. https://www.mdpi.com/1424-8220/25/1/214U.S. Army PEO STRI. (2020). *One World Terrain (OWT) Overview*. U.S. Army Program Executive Office for Simulation, Training and Instrumentation. https://www.peostri.army.mil/one-world-terrain

Wang, Y., Ahsan, U., Li, H., & Hagen, M. (2023). *A comprehensive review of modern object segmentation approaches*. arXiv. https://arxiv.org/abs/2301.07499

Yaghi, A., Tekli, J., Kamradt, M., & Couturier, R. (2025). 3DGENie: Synthetic point clouds for semantic segmentation in realistic virtual environments. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-025-20973-1

Yamazaki, K., Hanyu, T., Tran, M., de Luis, A., McCann, R., Liao, H., Rainwater, C., Adkins, M., Cothren, J., & Le, N. (2023). *AerialFormer: Multi-resolution transformer for aerial image segmentation*. arXiv. https://arxiv.org/abs/2306.06842

Zürcher, R., Zhao, J., Lau Sarmiento, A., Brede, B., & Klippel, A. (2023). Advancing forest monitoring and assessment through immersive virtual reality. *AGILE: GIScience Series, 4*, 15. https://agile-giss.copernicus.org/articles/4/15/2023/