Scaling for Monocular Depth Estimation in the Reconstruction of 3D Environments

Eric Guenther, Anakin Martinez, Amy Neuenschwander, and Jeff Perry
Center for Space Research, University of Texas at Austin
Austin, TX
eguenther@utexas.edu

ABSTRACT

The need to dynamically update 3D virtual environments for defense training and operations often outpaces traditional methods like full sUAS re-surveys. Monocular Depth Estimation (MDE) presents a promising alternative, yet its inherent lack of metric scale has been a primary barrier to its use in quantitative volumetric analysis. This paper introduces and validates a complete pipeline that successfully scales MDE output by leveraging existing, a priori 3D models. Our approach first performs a robust camera localization to precisely register a new image to the baseline 3D environment. It then uses this registered pose to generate a reference metric depth map, which provides the ground truth necessary to convert a relative depth map from the Depth AnythingV2 model into absolute, metric units. On a comprehensive dataset collected at the Geronimo CACTF, our method reconstructed new features from single images with an average geometric RMS error of 0.58 m and a geolocational accuracy of 0.96 m when validated against terrestrial laser scans. The pipeline's effectiveness was demonstrated for both additive construction and reductive scenarios simulating battle damage, confirming its utility for a range of operational applications, including potential utility to updating and maintaiing digital twins. This research validates a practical methodology for creating persistent, dynamically updated 3D environments from sparse imagery.

ABOUT THE AUTHORS

Eric Guenther is an Engineering Scientist at the University of Texas at Austin Center for Space Research and a PhD student in the Jackson School of Geoscience at the University of Texas at Austin. He has expertise in machine learning, lidar and sUAS applications. Mr. Guenther holds a B.S. in Geography from Virginia Tech and a Master's degree from Texas A&M University.

Anakin Martinez is an Engineering Scientist Associate at the University of Texas at Austin Center for Space Research. He has a background in aircraft composite fabrication, orbital mechanics, and sUAS applications. Mr. Martinez holds a B.S. in Aerospace Engineering from the University of Texas at Austin.

Amy Neuenschwander is a Senior Research Scientist at the University of Texas at Austin Center for Space Research. She is a past NASA Fellow, and has over 30 years of experience in EO/IR imagery, signal processing and lidar analysis. Dr. Neuenschwander holds a B.S. and a Master's degree in Aerospace Engineering, and a doctorate degree in Geography and the Environment from the University of Texas at Austin.

Jeff Perry is a Senior Engineering Scientist at the University of Texas at Austin Center for Space Research. He is a software developer with over 30 years of experience in digital image processing and 3D algorithm development. Mr. Perry holds a B.S. in Computer Science from Texas A&M University.

Scaling for Monocular Depth Estimation in the Reconstruction of 3D Environments

Eric Guenther, Anakin Martinez, Amy Neuenschwander, and Jeff Perry

Center for Space Research, University of Texas at Austin

Austin, TX

eguenther@utexsa.edu

Introduction

The utilization of high-resolution 3D virtual environments is critical for modern simulation and training efforts. The accurate geospatial and geometric representation of objects within these scenes is essential for a variety of defense-related training and operational applications. These 3D scenes provide realistic scenarios to improve training and readiness of mission planning. The generation of the virtual 3D environment is often derived from sUAS photogrammetric surveys collected over training facilities. These 3D reconstructed models, however, represent one epoch and changes to the environment can present as a challenge to update the 3D space. Changes in the 3D virtual environment require either manually modifying the scene by a human analyst, or by conducting another sUAS scan of the area in a survey pattern. Both of these options are time-consuming and resource intensive and may not be feasible given dynamic operational conditions or time-sensitive training scenarios.

The motivation behind this research is to support the incremental and/or dynamic updating (IDU) of existing 3D virtual content with new content based upon a small set of newly captured images. Recent advancements in artificial intelligence (AI) and deep learning techniques offer a promising alternative to previous options. One class of AI models are Monocular Depth Estimation (MDE) algorithms, which have the ability to generate depth maps from a single image. These algorithms have found great use in the fields of augmented reality and autonomous vehicle navigation. While these MDE techniques are rapidly improving in their capacity to estimate relative depths (i.e., determining if one object is nearer or farther than another), they currently struggle to accurately estimate true, metric depths, that is, the actual distances to objects in real-world units. This limitation prevents their direct application for accurate volumetric representation.

This work proposes a novel approach to detect and dynamically inject 3D volumetric changes into an existing 3D virtual environment over an area of regard (AOR) from a single or sparse-set of images. Our IDU pipeline leverages the *a priori* 3D virtual environment information to 1) refine the geospatial position and orientation of a new, single image capturing new changes within the AOR, 2) compute the relative depths from the new image, 3) scale the relative depths into metric depths, 4) automatically detect change based upon the depth maps, and 5) project the localized change into the full scene with correct geolocation and geometric attributes. Figure 1 illustrates the baseline 3D virtual environment (top left) and a single photograph from the same perspective collected by a sUAS with a change object added (bottom left). In this image, the change object is a field constructed sandbag pillbox bunker. The image on the right illustrates the pre-existing 3D virtual environment after the change object has been added to the model.

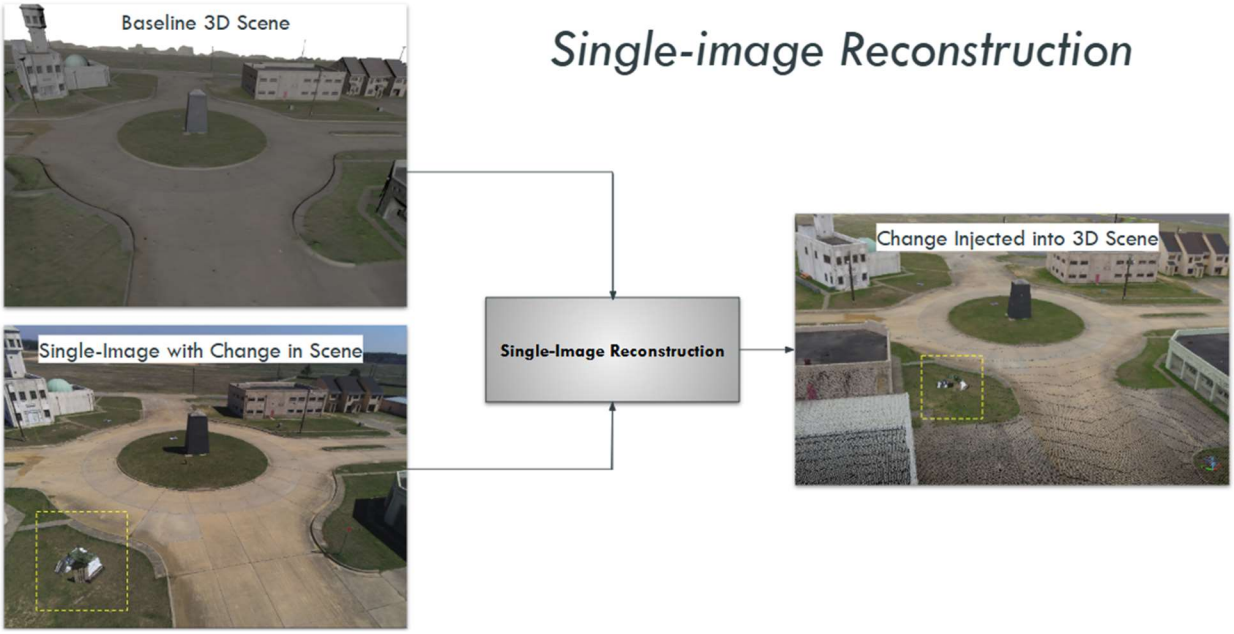

Figure 1. Simplified overview of sparse-view reconstruction pipeline.

Background

Structure-from-motion and its limitations

Structure-from-Motion (SfM) is the primary photogrammetric technique used to create 3D virtual models from 2D images utilized by the DoD community. The SfM process relies upon identifying and matching corresponding common points (or key points) across a large number of images with high overlap. These key point matches are used to simultaneously solve for both the 3D structure within the scene as well as the camera pose for each image during a process known as bundle adjustment (Schonberger & Frahm, 2016).

While highly effective at generating 3D content, the core requirements of SfM present significant limitations for dynamic scene-updating scenarios. SfM requires a dense set (i.e., hundreds to thousands) of highly overlapping images to reliably reconstruct both camera positions and virtual geometry. This necessitates a deliberate, survey-style data collection for optimal results, which is not always possible in time-sensitive or contested environments. In addition, the processing itself is computationally intensive and slow, making it unsuitable for near real-time updates. While core principles of SfM, such as feature matching and geometric verification, are crucial to 3D reconstruction, its reliance on data density makes it ill-suited for the challenge of updating a virtual environment from a single or sparse set of images not collected in a photogrammetric pattern.

Monocular Depth Estimation

Monocular Depth Estimation (MDE) has emerged as a powerful alternative, using deep learning models to infer depth from a single image. Early MDE models often struggled with generalization and produced artifacts. Recent transformer-based architectures, such as Marigold (Ke et al., 2024) and Depth Anything v2 (Yang et al., 2024), have demonstrated remarkable zero-shot performance on a wide variety of images in real-world settings.
For this work, we selected Depth Anything v2, a state-of-the-art MDE model, due to its superior performance in producing high-resolution, clean, and detailed relative depth maps from challenging real-world inputs. It excels at preserving sharp boundaries and capturing fine geometric details, making it ideal for our reconstruction pipeline. Like most leading MDE models, however, its standard output is a relative depth map, where pixel values represent depth on an unscaled, arbitrary gradient.

Some research has explored direct metric depth estimation. The creators of Depth Anything v2, for instance, released a version of their model fine-tuned on the VKITTI2 dataset, which contains synthetic urban driving scenes with metric ground truth (Gaidon et al., 2016). While promising, we found this metric model produced noisier outputs and still required post-processing and scaling for our use case. Other techniques that produce metric depth, such as those leveraging multi-frame photometric consistency (Hu et al., 2024), move back toward requiring dense image sets, thus negating the primary advantage of MDE. Our work therefore focuses on leveraging the high-quality relative depth from the base Depth Anything v2 model and scaling it using external information.

Data and Experimental Design

Field Site and Data Acquisition

Data used in this study were acquired at the Geronimo Combined Arms Combat Training Facility (CACTF) at the Joint Readiness Training Center (JRTC), Fort Johnson, Louisiana from March 8 - 10, 2025. The data collection was conducted over three days and designed to test the viability of updating a high-resolution, high-fidelity 3D virtual model with a sparse number of opportunistically captured images.

During the first two days, a comprehensive high resolution 3D baseline dataset was collected using survey-grade photogrammetric techniques. The sUAS data used in this study were collected from a Skydio X2D system which is equipped with only a pseudorange code GPS; thus, differential correction through either a Real Time Kinematic (RTK) or Post Processing Kinematic (PPK) was not possible. The Skydio X2D is equipped with a 12 M Pixel camera and has obstacle avoidance capabilities that allows for easier piloting at lower altitudes for more detailed collection. The pose information from each Skydio X2D photo includes a position (X, Y, Z) where Z is height above the reference ellipsoid and orientation angles of omega (rotation about the x-axis), phi (rotation about the y-axis) and kappa (rotation about the z-axis). The AgiSoft MetaShape software was used to construct a 3D reference mesh from the raw photos using Structure from Motion (SfM). The data were georeferenced with ground control points measured by a real-time kinematic (RTK) GPS such that the final 3D geolocation accuracy of the resulting baseline 3D virtual model is better than 5 cm.

TLS data were collected at the Geronimo CACTF site using a Trimble X7 laser scanner. The Trimble X7 utilizes a self-registration from each scanning position creating high accuracy 3D point clouds. When collecting TLS data of building exteriors, the goal is to obtain multiple views such that occlusions are minimized or eliminated. Data from the TLS were placed into absolute geolocation positions (UTM projection, WGS84 Epoch 2010 ellipsoid) using the same GCPs targets as were used to anchor the input sUAS data with an overall RMS error of 2.3 cm. By placing the TLS data into absolute real-world coordinates, the TLS data serves as a very high spatial resolution validation data source. The TLS collected data for the full three-day period including capturing scans of all the buildings as well as the change objects.

Following the baseline collection, deliberate physical changes were introduced to the environment, which are referred to as ‘change objects’ in this study. These included the construction of a sandbag pillbox bunker, a blue pop-up tent, and the strategic placement of replica unexploded ordnance. On Day 3, the modifications to the AOR were captured using collection patterns designed to simulate opportunistic, non-survey data acquisition. This involved fast fly-by maneuvers with sUAS platforms and on-the-ground data collection using handheld cameras, mimicking the movements of a squad navigating the facility. The key objective of this experimental design was to evaluate if a metrically accurate 3D model could be successfully updated using sparse, erratically-captured imagery from different sensor perspectives.

Figure 2. Example change objects in CACTF site

Methodology

Our IDU pipeline considers two types of change; additive or reductive. For the additive case, something new is observed in the sUAS photograph that is not reflected in the baseline 3D virtual environment. The reductive case represents when something has been removed from the baseline 3D virtual environment and is no longer present in the sUAS photograph. Our IDU pipeline consists of a seven-step process (highlighted in Figure 3) designed to ingest two inputs: 1) Baseline 3D virtual environment: an existing 3D model that represents the AOR in high-detail and serves as the a priori, and 2) sUAS collected change image(s): a single or sparse-set of images with some change in the image that is not represented in the baseline 3D scene. Though not explicitly required, this version of the IDU pipeline requires image EXIF tags to include their estimated position and camera pose. The output of the IDU pipeline is an updated 3D model including the change in the scene as observed in the image set for additive changes or a modification of the baseline 3D virtual environment for reductive changes. The seven-step process is outlined as follows:

1. **Camera Localization**: The initial camera pose of the change image is refined using a camera localization algorithm.
2. **Baseline Depth Extraction**: A ground-truth metric depth map is rendered from the baseline 3D virtual environment using the refined camera pose information.
3. **Monocular Depth Estimation (MDE)**: Depth Anything v2 is run on the change image to produce a relative depth map.
4. **Apply Scaling Function to MDE output**: A novel scaling algorithm uses the baseline depth map as a reference to convert the relative MDE depth map to a metric depth map.
5. **Change Detection**: The scaled MDE depth map and the baseline depth map are differenced to identify volumetric changes in the scene.
6. **Depth Map Fusion (for multiple photos)**: If multiple sparse images are available, their scaled depth maps are fused using a Truncated Signed Distance Function (TSDF).
7. **Meshing**: The final depth map of the change is projected into 3D space as a 3D point cloud and subsequently meshed to create the final volumetric update.

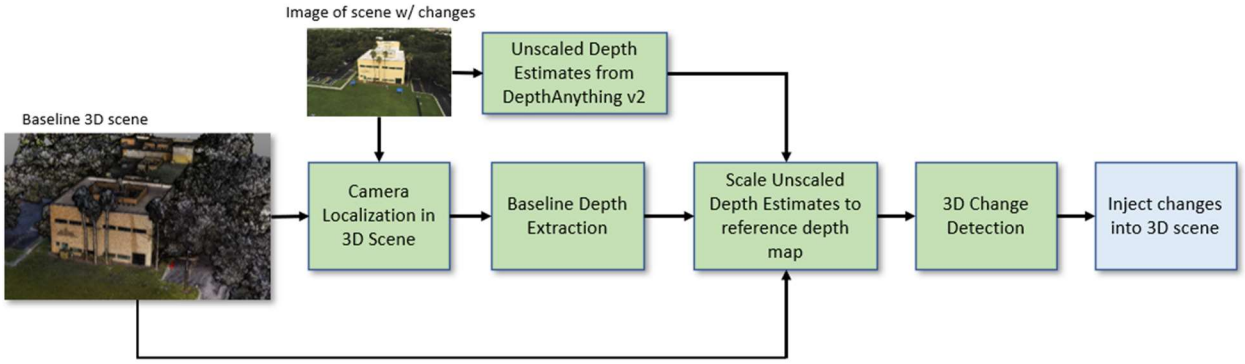

Figure 3. Flow chart showing processing pipeline for sprase-view reconstruction.

Camera Localization

Camera localization is the process to precisely determine the camera's position and orientation (pose) which are critical for accurate 3D reconstruction. Since the GPS available on sUAS systems do not include real-time differential correction, the camera position (XYZ) can be off by several meters. Our process for camera localization uses the baseline 3D virtual environment as the reference. Using the pose and position information provided in the sUAS image metadata, the frustum (or 2D representation from the same perspective) of the baseline 3D model is generated. Next, both images are converted to RMS contrast images to highlight high contrast areas such as edges, in addition to mitigating differences in color space between real images and synthetic renderings of the same area. A correlation error is computed between each contrast image. The iterative process then recomputes the synthetic image while a semi-stochastic search is performed across six degrees of freedom (i.e., X, Y, Z, roll, pitch, yaw) with replacement within a predefined range to minimize the correlation error between the RGB image and synthetic image of the 3D scene. The minimum residual error yields the optimal camera pose. A key feature of this algorithm is its ability to operate on both RGB-images as well as unscaled depth maps and match features between real-images and synthetic images from 3D models.

Baseline Depth Map Extraction

Once the refined camera pose is determined, it is used to render a view of the baseline 3D scene from that exact perspective. From this vantage point, we compute the 3D Euclidean distance from the virtual camera origin to every point on the model's surface, generating the reference metric depth map. This map serves as our ground truth for the unchanged portions of the scene and is the crucial reference for the scaling process. The quality and accuracy of the extracted reference depth map depends on the alignment of the camera pose from the prior step, as well as the accuracy, quality, and detail of the baseline 3D scene. Depending on the source of the 3D environment may have degraded quality in obscured areas.

Monocular Depth Estimation with Depth Anything v2

To estimate initial relative map depths for the change images, we use the monocular depth estimation model Depth Anything v2. Depth Anything v2 is an image transformer built on the DINO v2 encoder (Caron et al., 2021), and produces very-high detail depth maps. A critical property of these depth maps is that the depths have a high degree of internal scale consistency or relative accuracy. This consistency in depths allows for the relative depth estimates from Depth Anything v2 to be converted into metric depths via a scaling function. To further enhance the fidelity and precision of the depth estimation for a more detailed reconstruction, we modified the standard model output to generate 16-bit depth images, providing a much greater dynamic range than standard 8-bit images. An example of the depth estimates from Depth Anythingv2 compared to those generated by calculating the true distance from the baseline 3D virtual scene are shown in Figure 4.

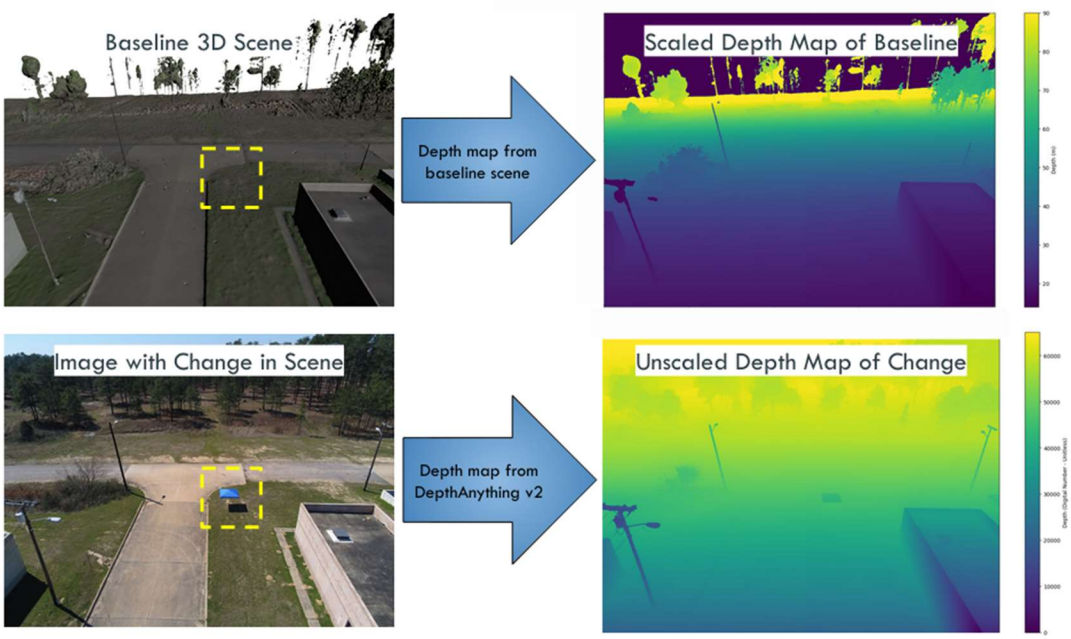

Figure 4. Examples of images and corresponding output depths. Top row represents the baseline 3D scene from the same perspective as the sUAS photo was taken. The corresponding depth map from the baseline 3D scene was determined by computing the Euclidean distance between each point in the image to the virtual camera position. Bottom row represents the sUAS image and corresponding depth map with a change object (blue tent) observable. The output depths from the sUAS image are relative depths predicted by Depth Anythingv2 and are unscaled.

Scaling Function

With a metric reference depth map (from the baseline 3D scene) and a relative MDE depth map (from the change image) determined, we employ a best-fit scaling algorithm. The core assumption of the scaling function is that the majority of the scene is unchanged. The scaling algorithm analyzes the relationship between the two depth maps in these static regions, identifying the most common metric depth associated with each relative depth value. Figure 5 illustrates the relationship used to construct a scaling function mapping the relative MDE digital numbers to metric depths. It is important to note that the variability in predicted depths increases as the reference depths increase. The scaling relationship between the predicted depths and true depths are likely to be a function of the camera pointing angle, camera height above ground, and camera resolution. This scaling function is then applied to the entire MDE depth map, ensuring that new objects (the "change") are scaled in a way that is consistent with the real-world proportions of the scene. Because the relative depths predicted from Depth Anythingv2 are largely self-consistent, the scaling function can be applied globally to the entire image.

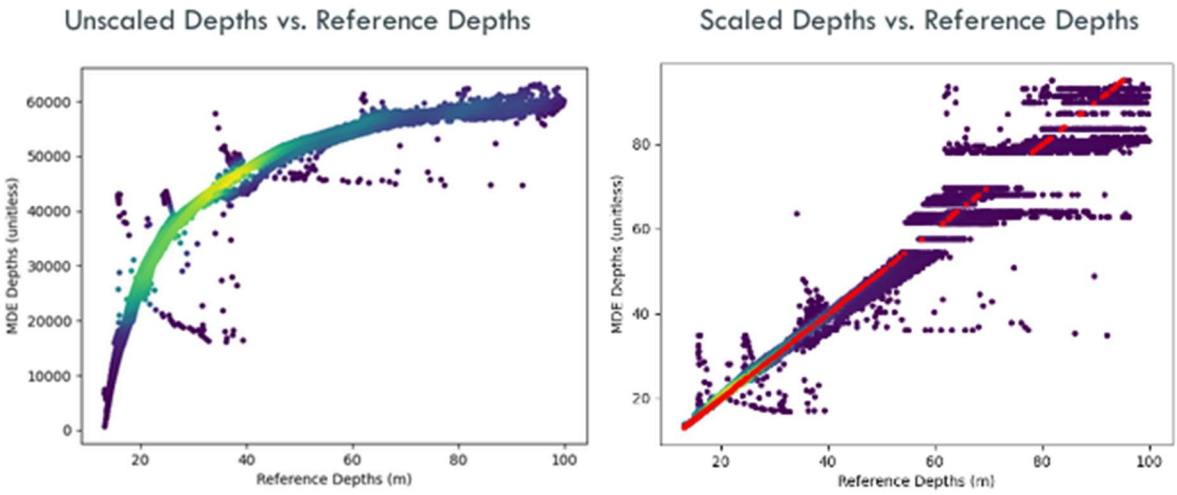

Figure 5. Left, chart showing the true metric depths compared to the monocular depth estimation for the sUAS photograph shown in Figure 4. Right, true metric depths compared to the scaled monocular depth estimation for the same scene.

Change Detection

For our pipeline, change detection is performed in the depth domain by calculating the difference between the reference depth map and the newly scaled MDE depth map. This difference map highlights pixels where there is a significant discrepancy, corresponding to either added or removed volume. This process effectively isolates the true volumetric change while also helping to identify potential errors from camera misalignment. The resulting difference map is segmented to create a mask that isolates the change object for the final meshing step.

Fusing Multiple-Depth Maps (if available)

For a sparse set of photos that observe a change object from multiple vantage points, we fuse the multiple scaled and masked depth maps using a Truncated Signed Distance Function (TSDF) (Werner et al., 2014). TSDF is a volumetric fusion technique that integrates depth data from multiple views into a single voxel grid. By averaging distance measurements within each voxel, it effectively reduces noise and fills gaps from occlusions, making it robust to slight misalignments and ideal for generating a cohesive model from sparse data.

Projecting 3D Changes and Meshing

The final depth map of the segmented change object is converted into a 3D point cloud by projecting each pixel's depth value into 3D space using the known camera intrinsic and pose. In cases where an object is added, the depth map for the added object is projected into the 3D scene using the depth map along with the known camera parameters including the camera intrinsic and extrinsic matrices, creating a dense point cloud ready to be meshed. For instances where an object is removed, the corresponding geometry must also be removed from the baseline 3D environment. Using the camera parameters and the change mask, the vertices that compose the removed object are identified from the perspective of the camera and then deleted from the baseline 3D model. The new scaled depths within the change mask are then projected into a 3D dense point cloud, as is the case with additive changes.

The RGB color from the original image is applied to each point providing a photo quality representation of the 3D object. A Ball-Pivoting algorithm (Bernardini et al., 1999) is then used to construct a continuous 3D mesh from the change object point cloud. Finally, vertex normals are calculated and a statistical outlier removal filter is applied to improve visual fidelity. This new mesh can then be added to the baseline 3D virtual scene. Our pipeline also

supports destructive changes; if the depth difference indicates an object has been removed, the corresponding vertices and faces are deleted from the baseline model.

Results and Discussion

Results

The sparse-view reconstruction was able to recreate changes in the 3D virtual environment as represented in the photos at the same location and with similar geometric characteristics. For this analysis, we focused on three distinct additions to the AOR: a sandbag pillbox, a blue tent, and a police kiosk. Each of these change objects were reconstructed from a single photo. These single-image reconstructions, while positionally and geometrically accurate, notably only represent the portion of the object visible to the camera. As a result, surfaces occluded from the camera's viewpoint are not reconstructed. This limitation due to obscuration can be mitigated if the object is viewed from multiple perspectives.

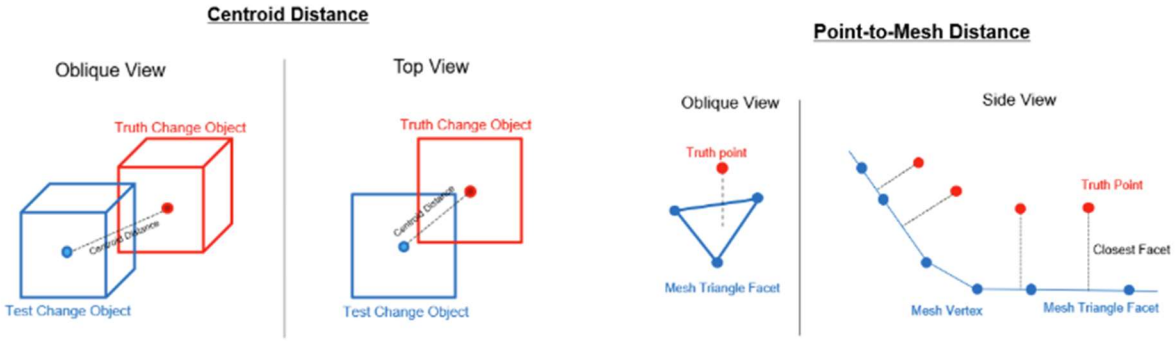

Figure 6. Diagrams showing how the Centroid Distance and Point-to-Mesh Distance calculations are made to assess the reconstructed meshes compared to the TLS data.

To quantitatively assess the accuracy of the 3D reconstruction of change objects, the 3D objects were compared against coincident TLS validation data that was collected of the change objects, capturing centimeter-level geolocation accuracy, and sub-centimeter geometric accuracy of these objects. Two metrics (as illustrated in Figure 6), centroid distance and point-to-mesh distance, were computed to assess the geolocation and geometric accuracy, respectively. Geolocation accuracy was computed by differencing the centroid position of the change object against the TLS validation data. The geolocation accuracy represents how precisely the change object's reconstructed position in a 3D virtual environment matches its actual position within the AOR. The geometric accuracy measures how closely the 3D reconstruction matches the true shape and form of the true object. To calculate the point-to-mesh distance, the meshed change object is aligned via interactive closest point (ICP) for the best 3D fit between the reconstructed mesh and the TLS scan of the same object. Next, the 3D Euclidean distance between each point in the TLS scan and the closest triangle facet in the reconstructed mesh is calculated and the distances are aggregated to calculate the mean absolute error (MAE) and the RMS Error.

Table 1. Geolocation and Geometric Error Metrics for Reconstructed Change Objects

Change Object	Geolocation Accuracy	Geometric Accuracy	
-	Centroid Distance (m)	Point-to-mesh RMSE (m)	Point-to-mesh MAE (m)

Bunker	0.61	0.31	0.23
Tent	0.41	0.12	0.09
Kiosk	1.85	1.33	1.25

The geolocation errors for the bunker, tent, and kiosk were 0.61 m, 0.41 m, and 1.85 m, respectively, yielding an average error of 0.96 m. The average geometric RMS Error for the 3D reconstructions **is** 0.58 m, indicating that this approach can represent an object's geometry well within about a half-a-meter of detail.

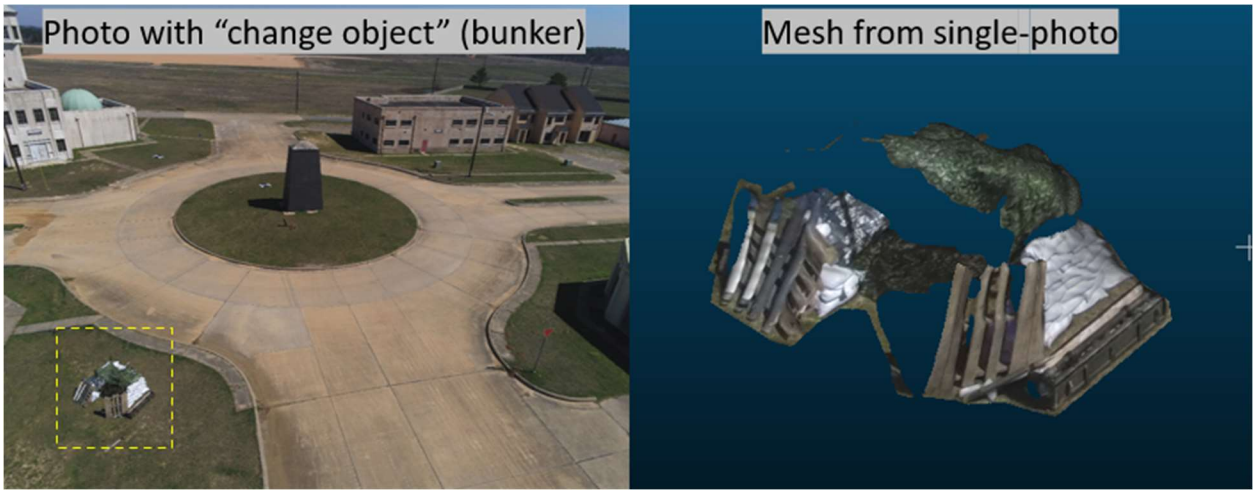

Figure 7. On the left, the image of a new change (the sandbox pillbox bunker). This image was used as the input. On the right is the 3D mesh model output of the bunker. Occluded areas of the bunker were not reconstructed from the single-photo.

In addition to additive reconstruction, our pipeline is equipped to model large-scale reductive changes, a critical feature for applications like Battle Damage Assessment (BDA). While our field test did not involve destructive events, we validated this capability with a synthetic scenario. We digitally removed a building from a source image of the Geronimo CACTF to simulate the aftermath of a kinetic strike (Figure 8).

Figure 8. To test the reductive case, an image was altered to make it appear that the building was removed. This was then used as an input for sparse-view reconstruction.

Our change detection algorithm registered a large-scale negative depth discrepancy where the structure was formerly located. This triggered a reductive modification process: the pipeline identified and deleted all vertices and faces

associated with the building in the 3D baseline virtual environment. Significantly, this included geometry that was not visible in the 2D input image (e.g., the rear and interior sections of the building), confirming a true 3D update rather than a simple projection. The algorithm then used the visual information from the altered photo to reconstruct the newly visible ground plane. This example illustrates that our approach can provide rapid, accurate updates to 3D environments to reflect major destructive events, a crucial requirement for modern military operations.

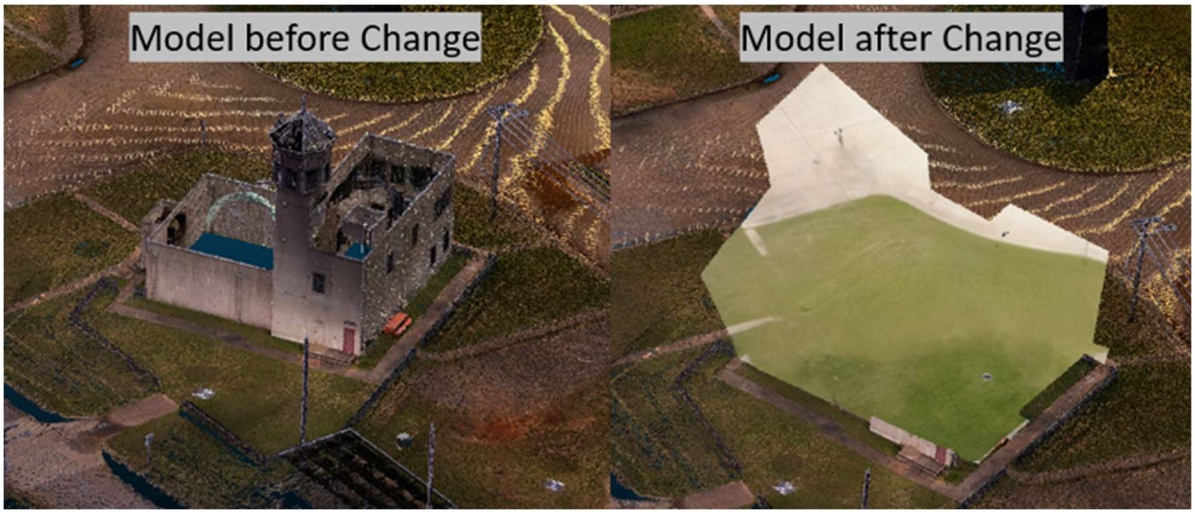

Figure 9. Demonstrating the removal of a building using the sparse-view reconstruction pipeline.

Discussion

This work demonstrates a significant departure from traditional scene SfM reconstruction workflows, which typically demand dense, survey-pattern image sets. Our results show that it is feasible to dynamically update high-fidelity 3D virtual scenes from a minimal set of images; including a single photograph. The core of our methodology is to expedite the standard SfM pipeline by leveraging MDE as a powerful front-end depth predictor. This discriminative approach, which predicts depths directly, is fundamentally different from is fundamentally different from generative methods. Computing and calculating true metric depths unlocks the ability to adopt powerful tools from the wider SfM ecosystem for rapid 3D reconstruction. Furthermore, once a metric depth is established, we can employ standard and robust techniques like TSDF for fusing multiple sparse views; a task that would be difficult with unscaled, relative depth maps.

This work could be crucial for many training and operational contexts related to DoD activities. For instance, in training scenarios, updates to a 3D virtual environment can be made with only a few reference photos of those changes, especially in high-tempo training environments. This also does not require any particular training in new systems as would be required in the case of a repeat sUAS survey-mission or manual manipulation of the 3D virtual environment. Similarly, in operational use-cases where access is limited, updates to a reference 3D virtual environment could be made with opportunistic photos.

In addition, this work provides a framework that could be used to provide near real-time updates to digital twins. A key function of digital twins is that they mirror their real-world counter parts. Our approach provides a mechanism to accurately update the visual and geometric characteristics of a digital twin from a sparse or opportunistically captured imagery. This allows for digital twins to be updated with minimal operational footprint and without the need for additional hardware such as laser scanners. As a result, this proposed approach could significantly lower the barrier to keeping digital twins in near-constant synchronicity with their physical environment, enhancing their value for monitoring, simulation, and planning.

Our development process identified two primary limitations of the current MDE-based approach: effective range and minimum object size. In testing the sparse-reconstruction at long range, Depth Anythingv2 struggled with differentiating distances between far away objects. As a result, it was difficult to determine if changes in depths were

a result of errors with the monocular depth estimation or from true volumetric change. Currently, our IDU reconstruction pipeline works best with objects within 100 m of the camera, however, improvements in each of the fundamental steps should further improve the accuracy and range capabilities. In addition to a distance limitation, small objects are difficult to discern within our IDU pipeline. One reason for this limitation is that the change detection algorithm relies solely on changes in depth (volumetric or 3D change) and smaller objects are more difficult to identify. Specifically objects smaller than 1 cubic meter are difficult to determine if they result from true change or noise. Similar to the current range limitations, this limitation could be mitigated with improvement of the change detection and volumetric approaches. In addition, incorporating an image-based change detection step could allow for the identification of smaller objects that could then be reconstructed volumetrically.

The primary source of error in the final 3D reconstruction of the change objects was traced back to minor inaccuracies in the initial camera localization step. While our camera localization algorithm proved effective, any residual pose error propagates through the pipeline, slightly distorting the final scaled geometry. Future work will prioritize improving this crucial first step. Improvements include not only refining the camera localization algorithm itself but also investigating its replacement with more advanced, state-of-the-art localization frameworks such as Dust3r (Wang et al., 2024). Furthermore, the quality of the baseline3D virtual mesh impacts the accuracy of the localization. Future implementations testing operationally-representative 3D virtual reference models will be essential for validating the robustness of the camera localization methods.

Improvements to the MDE component itself present another clear path for enhancement. The Depth Anythingv2 model was trained primarily using ground-based photos rather than from the vantage point of a sUAS as its was developed for vehicular navigation. Retraining the Depth Anything v2 model on a custom dataset more aligned with our aerial, defense-centric use case could significantly improve depth prediction quality. Moreover, further investigation into directly predicting metric MDE models, rather than relative depths, remains a high priority. A reliable method to predict metric depth without a scaling step would be transformative, especially in scenarios where dramatic changes to the environment might invalidate the core assumption of our scaling algorithm—that most of the scene remains static.

Ultimately, improvements to both camera localization and monocular depth estimation will enable more accurate reconstructions with minimal effort. This opens the possibility of iterative updates, allowing a 3D virtual model to be refreshed continuously as an area experiences ongoing change. Such a capability is a critical step toward creating persistent, “living” 3D virtual models for mission-critical applications.

Conclusion

This work presents a novel pipeline for updating existing 3D environments using single or sparse images. We addressed the primary limitation of Monocular Depth Estimation—its inability to produce metric-accurate depth—by leveraging a priori scene data as a ground-truth reference. Our method successfully refines camera pose, scales the MDE output to metric units, and reconstructs new or altered geometry with a high degree of fidelity. The results, validated on a comprehensive field-collected dataset, demonstrate that this approach is a viable solution for the dynamic updating of 3D scenes in resource-constrained environments. This work represents a significant step toward creating persistent, dynamically updating 3D models for critical defense and training applications.

ACKNOWLEDGEMENTS

This work is sponsored by the U.S. Army Simulation and Training Technology Center (STTC). Many thanks are given to the STTC team of Mr. Clayton Burford, Mr. Gage Jenners, and Ms. Shahira Castellano. Additional thanks are provided to Leidos Corporation: Mr. Tu Lam, Mr. Farid Mamaghani, and Ms. Amanda Larrieu. Final thanks are given to GeoAcuity for the collection and usage of the sUAS data used in this study.

REFERENCES

Bernardini, F., Mittleman, J., Rushmeier, H., Silva, C., & Taubin, G. (1999). The ball-pivoting algorithm for surface reconstruction. IEEE Transactions on Visualization and Computer Graphics, 5(4), 349–359. https://doi.org/10.1109/2945.817351

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers (No. arXiv:2104.14294). arXiv. https://doi.org/10.48550/arXiv.2104.14294
Gaidon, A., Wang, Q., Cabon, Y., & Vig, E. (2016). Virtual Worlds as Proxy for Multi-Object Tracking Analysis (No. arXiv:1605.06457). arXiv. https://doi.org/10.48550/arXiv.1605.06457

Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., & Shan, Y. (2024, September 3). DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos. arXiv.Org. https://arxiv.org/abs/2409.02095v2

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., & Schindler, K. (2024). Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation (No. arXiv:2312.02145). arXiv. https://doi.org/10.48550/arXiv.2312.02145

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-Motion Revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4104–4113. https://doi.org/10.1109/CVPR.2016.445
Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., & Revaud, J. (2024). DUSt3R: Geometric 3D Vision Made Easy (No. arXiv:2312.14132). arXiv. https://doi.org/10.48550/arXiv.2312.14132

Werner, D., Al-Hamadi, A., & Werner, P. (2014). Truncated Signed Distance Function: Experiments on Voxel Size. In A. Campilho & M. Kamel (Eds.), Image Analysis and Recognition (Vol. 8815, pp. 357–364). Springer International Publishing. https://doi.org/10.1007/978-3-319-11755-3_40

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth Anything V2 (No. arXiv:2406.09414). arXiv. https://doi.org/10.48550/arXiv.2406.09414