A Semi-Synthetic Approach to Generating Datasets for Training Highly Targeted Computer Vision Models

Executive Summary

Proposing a new approach to generating large datasets for training computer vision models that can detect specific objects and vehicles in overhead surveillance imagery by merging highly accurate synthetic renders of target objects with real world data from high fidelity 3D scans.

Background

Training machine learning models that can identify and track objects in top down imagery and video is necessary for the deployment of aerial surveillance platforms of all types, ranging from low cost drones up to optical surveillance satellites. These platforms generate enormous quantities of data that must be analyzed and interpreted, and the size of the output has long outstripped human analysts' ability to keep up. There has therefore been an ever increasing reliance on machine learning based computer vision techniques to make up the difference.[1] This is a trend that is about to grow exponentially due to the consumer electronics and private space industries fundamentally reshaping the cost structure of fielding these assets.[2] Optical surveillance satellite constellations will number in the thousands instead of dozens, and low cost drone platforms will be everywhere at all times. The pace of progress e need to get better at developing tools to automatically process these data streams.

Problem

The methods that are currently used to develop these tools are too slow, inefficient, and un-targeted to handle the task at hand. Training vision models to detect specific objects generally requires acquiring and manually searching through large datasets of domain relevant imagery for individual examples of the objects we wish to detect, each of those examples must then be carefully traced out and labeled, each label must be verified by a second annotator to avoid false positives, and then the overall dataset must be structured and packaged for training. Much of this work is still done manually, which is labor intensive, expensive, and time consuming.[3] And of course it is fundamentally limited to objects about which sufficient data already exists. As such

---

[1] https://www.nga.mil/news/NGA_releases_new_data_strategy_to_navigate_digital.html

[2] https://www.sda.mil/us-military-places-a-bet-on-leo-for-space-security/

[3] https://zslpublications.onlinelibrary.wiley.com/doi/10.1002/rse2.195

most efforts have been limited to distinguishing between well-established categories of common objects, for instance distinguishing between a building and a truck or a car and a motorcycle.[4] Only rarely is the considerable effort made to gather sufficient targeted data to train a custom computer vision model on a specific instance of an object due to the difficulty of finding enough examples to train on. Furthermore, model efficiency and precision is directly related to the quality of the input dataset.[5] So even in cases where useful models already exist for a given purpose, they were likely trained on limited datasets and model quality can be easily improved with more and better data.

In response to these challenges, many have attempted to use 3D techniques derived from the VFX and video game industries to generate synthetic data with limited success. The advantage of this approach is that a rendered object doesn't need to be found, and it already comes with a perfect segmentation mask that specifies exactly which pixels in the image belong to that object. But in practice these techniques are either low in cost and fidelity, or at best are of middling fidelity and expensive. On the low end, they simply overlay the object they would like to train the models to detect on top of aerial imagery. This approach lacks any 3D information about the environment surrounding the object, which results in the object appearing superimposed and separated from its background.[6] This makes it a poor analogue of what the model will encounter in real world data, and leads to limited improvement detection accuracy. And on the high end there has been limited success with taking the time to also simulate that surrounding context and therefore increase the realism of the object as it is integrated into the scene.[7] But then each individual environment must be laboriously created from scratch by a 3D artist, and then the entire environment must be rendered anew for each frame of training data. The financial and compute resources necessary to accomplish this is non-trivial, quickly approaching the costs associated with full service game development or VFX studios, and even then this approach will never fully replicate the quality and diversity of real world data. Datasets for computer vision model training must match the variety, distribution, and fidelity of the real world data they hope to examine, and building 3D environments that properly encompass this diversity is generally cost and time prohibitive. These 3D methods for generating datasets are therefore typically used to only to supplement real world data that must still be acquired manually, and results in just a small improvements in model accuracy on the order of ten percent.[8]

Solution Criteria

This status quo can be improved along a number of vectors. Of course we can reduce the cost of producing a dataset and/or speed the process. We can also increase the dataset's photorealism, and we can work to create datasets that better reflect the endless variability of the real world.

---

[4] https://www.kaggle.com/datasets/evilspirit05/visdrone

[5] https://ncsu-las.org/2023/12/eyeglass-improving-the-quality-and-efficiency-of-computer-vision-model-development-for-out-of-domain-datasets/

[6] https://www.iqt.org/library/the-rareplanes-dataset

[7] https://www.themoonlight.io/fr/review/improving-object-detector-training-on-synthetic-data-by-starting-with-a-strong-baseline-methodology

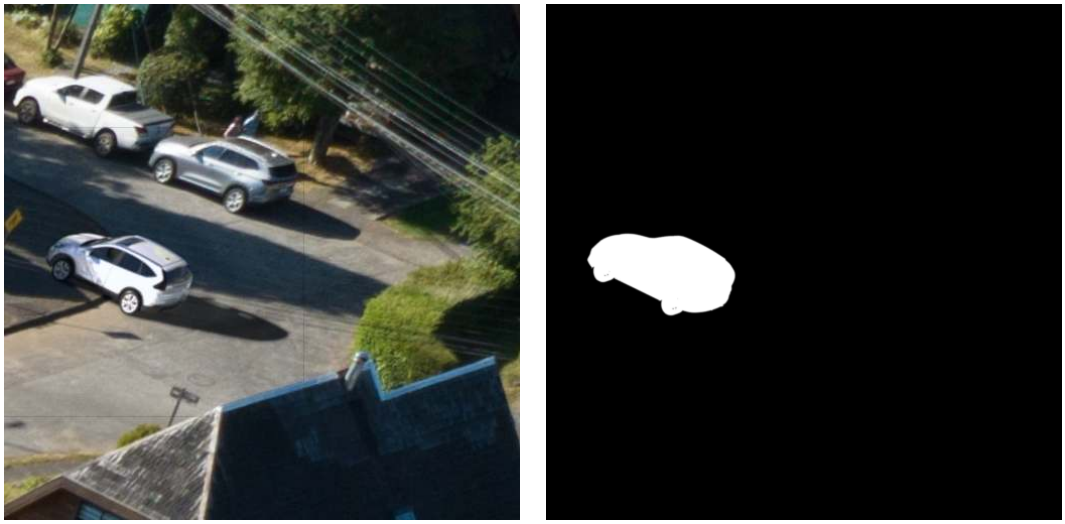
[8] https://arxiv.org/pdf/2411.05633

Finally we can also improve the process by reducing dependence on finding large numbers of physical examples of the objects to be detected. Sometimes you want to find something that is rare.

Solution

We propose a semi-synthetic approach to generating these datasets. Replicating the diversity of real world environments is achieved by actually going places and using drones to capture a large number of 3D scans of real landscapes, aggregating a diverse library of photorealistic 3D scenes that are representative real world distributions. And then we build a highly automated procedural VFX pipeline to batch render instances of the target object into those 3D scans. This VFX pipeline can use the scan data to reconstruct the 3D environment immediately around each rendered instance, which allows us to simulate photorealistic light interactions on the object. And recent advances in real-time radiance field rendering such as Gaussian Splatting allow us to render large scale 3D scanned data with unprecedented efficiency, often at frame-rates well in excess of 100 frames per second.

This semi-synthetic approach to acquiring datasets for training computer vision models allows us to automatically generate large numbers of photorealistic training images, each paired with a pixel perfect segmentation mask. And it improves performance over current state of the art methods along every possible vector.



Our photorealistic render of a vehicle into a 3D scanned scene along with its associated segmentation mask

It reduces cost and is considerably faster than traditional photorealistic 3D pipelines by eliminating the need for VFX environment artists. And relying on Gaussian Splatting to render the environments reduces both time and cost of compute. Of course it is also far better at replicating the realism and diversity of real world data because it is in fact data from the real

world. And finally it eliminates the need to accumulate a large number of real life examples of the target object, because any 3D can be rendered into the scene. Physically accurate 3D models of any object can be created from as little as one or two reference images, and then those 3D models can form the basis of a custom training dataset.

Already Accomplished

Capturing 3D scans of landscapes is easy, and rendering 3D models into a captured environment also easy. These are basic VFX tasks that any beginner should be able to accomplish in a matter of hours. The difficulty is in doing these things automatically and with a great deal of efficiency, and accomplishing them at scale.

Using synthetic data to train computer vision models that can track and identify specific objects in diverse environments requires generating a training dataset with a minimum of thousands and ideally millions of individual renders. -- each in a unique environment with a unique background. Accomplishing these numbers requires an automated system that has an understanding of the scene and is therefore capable of realistically distributing the objects to be rendered throughout the scene, and must also glean information from the immediate environment around each rendered instance so that the renderer can use physics based rendering to insert it into the scene in a convincing manner.

We have already built an end to end VFX pipeline that accomplishes this task using industry standard tools such as QGIS, UgCS, Reality Capture, and Houdini.

We have developed an easy to use QGIS plugin that can instantly generate drone flight patterns that were custom designed for the needs of this workflow, and center them on the operator's current location. And then UgCS is used to execute those flights. It currently takes about 15 minutes to capture 20 acres in high resolution and a square mile in lower resolution. Both high and low resolution captures are required for the workflow, and they are captured simultaneously using two drones. For comparison, generating a single photorealistic scene of equivalent size using traditional methods requires at least a few weeks of work, and as mentioned before requires much more compute resources to render.

We have used this to capture about 1000 acres of diverse landscapes across Argentina and Chile in high resolution. These countries were chosen due to their unique combination of permissive drone laws, and the nearly comprehensive diversity of their landscapes as you move from North to South. This is a seed dataset for the proof of concept phase of this effort. It covers everything from desert to agriculture to forest, and industrial to suburb to urban. It has been processed such that any object of interest can be placed into any place within those thousand acres of high resolution data and rendered from any perspective up 1000 meters altitude with a high degree of photorealism. And then those renders can be used to train custom computer vision models as previously described.

We have also built the Houdini pipeline that can take an unstructured image scan dataset and automatically organize, parse, and process it into 3D models and associated metadata such that Houdini can accurately place the objects of interest into these diverse scenes. And then we have

built the part of the pipeline that allows Houdini to manage the batch rendering of those objects in a way that allows the user to design the composition of the training dataset.

Furthermore, Houdini allows for the development of custom interfaces that have full access to all of its capabilities, so we are building all of the above functionality into a unified interface that does not require deep technical understanding of what is happening under the hood in order to generate these datasets. Automation and efficiency is key here, and the goal is for anyone who can work with 3D models to simply load in examples of the object/s that they would like the computer vision model to detect, choose between various settings that change the structure and attributes of the resulting dataset in order to tune the output to a specific desired purpose, wait for it all to render, and then use that dataset to train a highly targeted and efficient custom computer vision model.

Next Steps

- Having built the system that generates the datasets, the next step is to proceed to train custom models, demonstrate that it can accomplish real world tasks, and quantify improvements over current state of the art methods.

- Incorporate lessons learned from capturing the first batch of drone scans in order to increase capture efficiency and thus enable us to grow the database of 3D environments. For the initial dataset we intentionally captured more than was necessary, and now we can significantly speed up capture process by capturing only what we need. Between these efficiencies, and more resources onsite, we could easily increase the speed of capture by an order of magnitude or more. The first priority is to increase diversity by capturing more types of landscapes and to capture environments in different seasons in order to better match real world distribution.

- Expand dataset generation capabilities to video to allow us to train models specifically to track and navigate based on visual input. We are working to incorporate AI methods to better label the 3D models so that the target objects can be animated and realistically navigate within the scene, which will allow us to generate animated segmentation masks that can be used to train interactive computer vision models.

Conclusion

A higher quality and more targeted computer vision dataset allows a model to be trained with fewer parameters while maintaining accuracy. Fewer parameters means that it can either run faster on a powerful computer and thus search through more data or else it can run more efficiently on low power edge devices such as the onboard compute available to drones, allowing them to independently interpret their environments. Creating highly targeted datasets is currently prohibitively expensive in terms of both cost and time. A semi-synthetic approach that combines the benefits of real world data with the flexibility of computer generated imagery improves on the status quo by allowing for the efficient generation of large, custom, photorealistic training datasets.

Glossary of Terms

A **Computer Vision Model** is an array of machine learning weights that were acquired by training an Artificial Intelligence algorithm to accomplish a particular task in computer vision. Among other things, these tasks can include identification of specific objects, segmentation of elements within a scene, and vision based navigation within an environment.

A **Training Dataset** for computer vision is a collection of image tiles paired with information that the trainer would like the system to learn. The paired information can take a number of forms ranging from heat maps to squares drawn around the object, but the method we use due to increased performance is a **Segmentation Mask**, which is a simple black and white image that overlays the training image tile and marks exactly which parts of the image make up the target object. It is more difficult and expensive to get segmentation masks because they are usually drawn by hand, but training on them typically results in a more capable and efficient computer vision model.

GSD (ground sample distance) is a measure of resolution in remote sensing, specifically indicating the real world distance between the center of two adjacent pixels in an image, with the distance being measured at the surface of the object that is being imaged. Commercially available satellite imagery typically achieves GSD values of around 0.5, and drone imagery is generally less than 0.1 GSD. Notably, there are efforts underway to increase the resolution of commercially available satellite imagery by flying a satellite constellation in ultra low orbit, much closer to the surface of the earth, targeting GSDs of around 0.1 from an orbital platform with global coverage. If they succeed, this will drastically increase the range of objects that can be detected from space.

Houdini is the Visual Effects industry standard software for managing all major aspects of 3D production, with a particular emphasis on 3D data pipelines, and physically accurate simulations. Since its official release nearly 30 years ago, it has grown to play an integral role in the data management workflow of nearly every top tier VFX studio ranging including Disney, Pixar, Weta, Lucasfilm and Marvel Studios, as well as in the pipelines of countless video game studios. Unlike most other 3D softwares that are typically designed to be a tool for an artist to accomplish a specific range of tasks, Houdini is a generalized procedural software that allows the user to use visual programming tools and scripting to design custom tools that accomplish repeatable tasks, and integrate them into highly automated workflows that manage any form of data for a particular purpose. It is a foundational software in the industry, and using it as the basis of our data generation pipeline is no different than an engineering firm choosing to use something like AutoCAD or Solidworks.

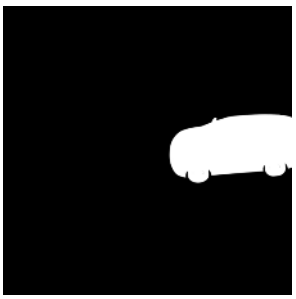
QGIS is an open source Geographic Information System management software that we use to automatically generate efficient flight paths for the drone to follow when capturing the 3D scan datasets used as a basis for our pipeline.

UgCS is a drone management software solution that can act as a base station for controlling a wide variety of proprietary and open source drones for 3D scanning and agricultural purposes. It can also control large numbers of drones simultaneously and is frequently used to conduct professional drone light shows.

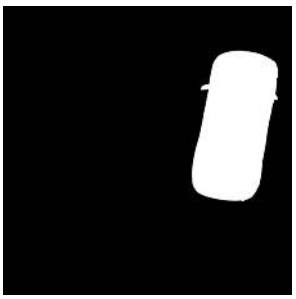
Appendix A: Computer Vision Model Training Target Object Examples

Physically accurate 3D models of a wide variety of civilian and military vehicles and military are readily available for purchase on commercial websites, and each can be used to generate custom datasets used to train computer vision models that can target those objects. In cases where such models do not exist, they can be created from a small number of reference photos.

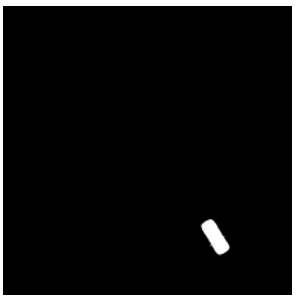
Appendix B: Example Training Data

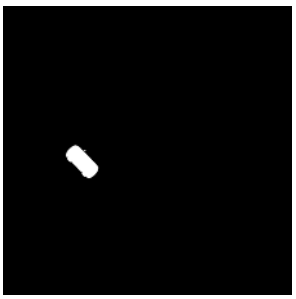
Each render utilizes 3D data about the object's immediate surroundings to give the object realistic occlusion, environmental reflections, as well as allow it to cast and receive shadows. This added realism helps the model more accurately learn what the object looks like in a wide variety of realistic circumstances.

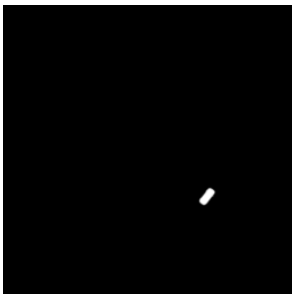
These effects work even with fine details such as overhead telephone wires, which are brought forward both of the rendered object and the shadow it casts on the ground.

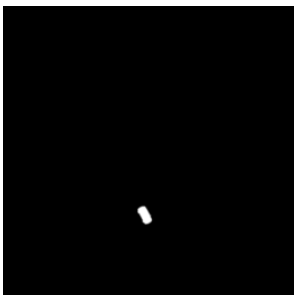
These interactions between the rendered object and its environment function at any arbitrary distance and GSD.

To better integrate the object into the scene, the pipeline uses real world data including image timestamps and gps coordinates to automatically align shadows and lighting conditions with the scan data in the background.

The pipeline also connects to historical weather data to drive the parameters on a volumetric simulation of the atmosphere in order to match diffusion levels for renders captured at a distance.

The maximum distance above ground captured in the 3D scans is about 1000 meters. The GSD at this altitude is between 0.3 and 0.5 meters per pixel, which is roughly equivalent to commercially available satellite imagery and allows us to create vision models that detect objects even at extremely low resolutions.

Appendix C: Comparison with Other Synthetic Dataset Efforts

The RarePlanes dataset used satellite imagery of known airports and rendered in 3D models of a variety of airplanes, as well as various augmentations to the landscape in order to improve airplane detection. These images were used to supplement a dataset of actual airplanes, and resulted in modest improvements to detection accuracy.

https://www.iqt.org/library/the-rareplanes-dataset

SkyScenes is a synthetic dataset for aerial scene understanding that uses traditional video game graphics pipelines to generate diverse landscapes and realistic scene interactions, but with limited realism.

https://arxiv.org/pdf/2312.06719