Modeling, Simulation & Gaming Working Group
May 2023

The Evolving Role of Synthetic Data in GEOINT Tradecraft

Abstract

The objective of this USGIF Working Group White Paper is to educate and inform the GEOINT community on the evolving role of synthetic data. Advances in artificial intelligence (AI) methods, such as Deep Learning (DL) and Generative AI, pose new opportunities and challenges in geospatial intelligence tradecraft. At the top of the challenges list is the need for massive amounts of labeled data to feed into AI systems. Synthetic data generation has emerged as a keystone technology to address this need. In this White Paper we seek to address the following questions:

- What is synthetic data for GEOINT?
- How is synthetic data being used across the GEOINT community?
- Why is synthetic data important for AI applications?
- What future trends in synthetic data will influence GEOINT tradecraft?

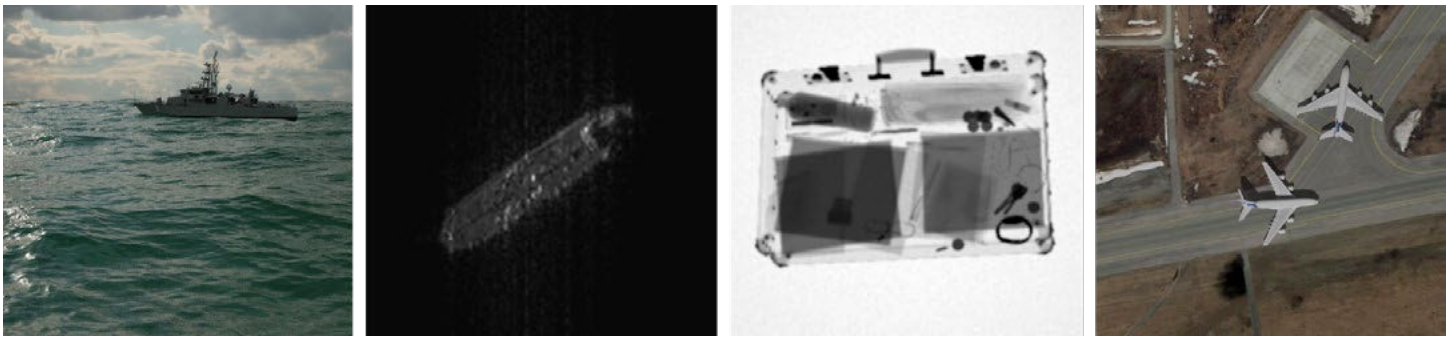


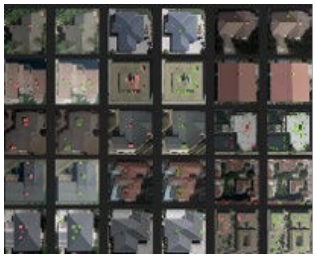
Figure 1. Examples of synthetic images used for training AI detection and segmentation algorithms across multiple sensor domains and platforms. ***Images courtesy of Rendered.ai.***

1.0 Introduction

GEOINT analysis techniques increasingly use AI methods that depend upon algorithms trained by accurate input data. Reliance on "real labeled training data" carries burdens of cost, time, and accuracy. In turn, data acquisition costs and timelines can impact mission critical applications. In addition, labeled data can also have security or privacy issues that limit distribution as well as biases inherent in data collection techniques. Engineered or simulated data, often called synthetic data, is increasingly being used to augment sparse datasets when training AI systems (Figure 2). Synthetic data has multiple patterns of creation, each carrying different benefits and drawbacks. For example, Generative AI techniques have garnered interest in creating "AI-driven synthetic data" because they automatically capture the underlying complex statistical data distribution of real data.

Generative AI techniques are complex because they require expertise in AI methods coupled with advanced computer software and hardware architectures to process massive amounts of data. In recent years, the Generative AI algorithm landscape has evolved mainly through breakthrough inventions in Generative Adversarial Networks algorithms (GANs), Variational Encoders (VAEs) and diffusion models. High-quality Generative AI models must address three important goals: (1) deliver at speed, (2) provide high quality samples and (3) generate good diversity. In practice however, GANs are weak on the diversity property, VAEs are weak on quality, and diffusion models are weak on the speed requirement. Accordingly, in order to address this "trilemma," a combination of techniques is frequently used to generate synthetic data. Generative AI models can also be used for training other AI algorithms by sharing the same statistical properties of real data samples. Generative AI models used for domain adaptation reduce the simulation to reality gap by translating synthetic data created manually to pseudo-real data.

(a) Real data versus synthetic.



(b) Synthetic data for rooftop clutter.

(c) Synthetic satellite imagery data for training algorithms.

Figure 2. Generative AI can produce massive amounts of synthetic satellite imagery for algorithm training and deployment. ***Images provided courtesy of Presagis and the VELOCITY 5D digital twin production platform.***

2.0 What is Synthetic Data and How is it Used?

Pixel-Based GEOINT Synthetic Data

Pixel-based data, such as satellite imagery, is what most people in the GEOINT community think of when considering synthetic data. Synthetic imagery data is used with computer vision algorithms to extend GEOINT exploitation applications such as searching for rare objects and unusual scenarios in scenes. In addition, synthetic data can also be used to test and evaluate new sensor specifications in order to assess AI efficacy in analyzing difficult to interpret features in complex geospatial environments.

Non-Pixel GEOINT Synthetic Data

Non-pixel GEOINT data includes structured observations from imagery, metadata of telecommunications, transponder data, and georeferenced human reporting. Synthetic data derived from high-fidelity simulations of human activity provides a valuable source of mission training data including: (1) Agent-based simulation of large populations for statistical testing of hypotheses for information transfer; and (2) Training AI algorithms to look for patterns of population interaction and communication. Digital twins are virtualized models of real-world systems, typically intended to solve specific business or operational problems, and support both pixel and non-pixel GEOINT data. Synthetic data of both these source types are essential to the training and validation process before a digital twin can be used for decision support applications.

How is Synthetic Data Used Today in GEOINT Analysis?

Labeled object examples can be used to train and test AI for counter-detection techniques, such as by generating different camouflage patterns, synthetic jammed signal examples, and novel scenarios used to mask certain observable patterns of activity. Synthetic non-pixel data is already being used to train GEOINT analysts on advanced tradecraft such as Activity Based Intelligence (ABI) skills. AI algorithms trained on synthetic data can be rapidly deployed to support edge capabilities or as a data mining tools for existing sensor data libraries to discover historical events and occurrences.

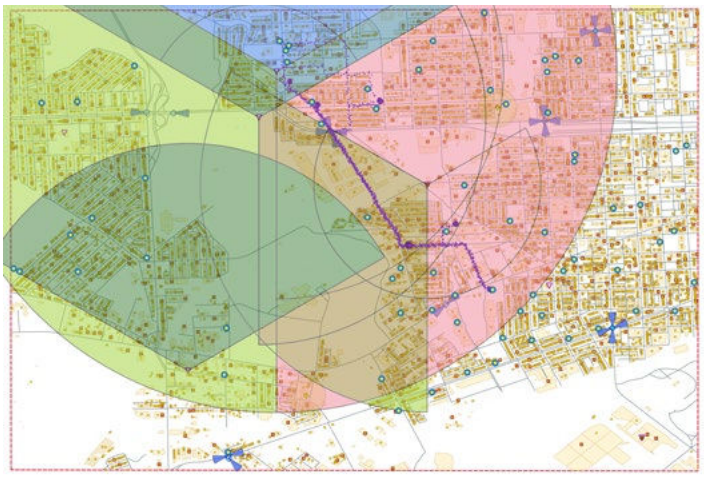


Figure 3. Example of Synthetic Non-Pixel Data, used in Whitespace's Worldline data in an urban geographic setting for training purposes. *Image courtesy of Whitespace.*

3.0 Recommendations for Use of Synthetic Data in GEOINT

- In practice, we recommend analysts create an iterative Synthetic Data Generation (SDG) feedback loop to improve analytical results for each problem set. SDG practices supplement large quantities of expensive real sensor data, thereby reducing the overall cost of AI/ML training and validation.
- Organizations should focus on capturing institutional knowledge regarding effective techniques for creating and using synthetic data for enterprise application approaches.
- Analysts should view SDG as an ongoing, updated enterprise capability that is essential for integration into their AI/ML training and validation pipelines.
- Synthetic data can also be used to test future sensors, training algorithms to detect rare objects or novel scenarios that do not frequently occur in operational settings.
- An important next step for synthetic GEOINT data will be exploring ways to use pixel and non-pixel data together to advance GEOINT tradecraft. The growing popularity of digital twins provides a realistic geospatial foundation for human activity simulations, which in turn creates more entity-based synthetic data.

Acknowledgments

Authors:
Chris Andrews, Rendered.ai
Steven Fleming, Ph.D., Institute for Environmental and Spatial Analysis, University of North Georgia
Patrick Kenney, Whitespace
Shehzan Mohammed, Cesium
Don Widener, BAE Systems
Sacha Lepretre, Presagis

Editors:
Stuart Blundell, MDA GeoIntelligence USA
Barry Tilton, Maxar

USGIF Staff
Ronda Schrenk, Chief Executive Officer
Christy Monaco, Vice President of Programs
Brad Causey, Senior Director of Communications
Madeline Rouse, Senior Intern

About USGIF

The United States Geospatial Intelligence Foundation (USGIF) is a 501(c)(3) nonprofit educational foundation dedicated to promoting the geospatial intelligence tradecraft and developing a stronger GEOINT community with government, industry, academia, professional organizations, and individuals who develop and apply geospatial intelligence to address national security challenges. USGIF achieves its mission through various programs and events and by building the community, advancing the tradecraft, and accelerating innovation.

White Paper Sponsor

SPACE CAPITAL®