Synthetic Computer Vision Data helps Overcome AI Training Challenges

Chris Andrews, Mark Hogsett
Rendered.ai
Seattle, WA
chris@rendered.ai, mark@rendered.ai

ABSTRACT

Artificial Intelligence and Machine Learning systems have become enterprise capabilities integrated in data collection and knowledge extraction pipelines for real-world sensor data exploitation. As enterprise tools, AI/ML systems need to constantly adapt to changing real world scenarios and information. Yet, the efficacy of AI/ML directly depends on the data used for training and validation, meaning that it is impossible to build high performance systems for future scenarios before real world conditions change.

Synthetic computer vision data offers the opportunity to proactively design, train, and test AI/ML systems in circumstances when real sensor data acquisition is impossible, expensive, or difficult. With emerging tools, developers and data scientists can create physics-based synthetic, or engineered, datasets that emulate sensors, platforms, and scene content for unusual, hard to detect, or future circumstances. As synthetic data is 100% accurately labeled, engineers can develop synthetic datasets for non-human-interpretable sensor domains with reliable annotation for training AI/ML.

Through collaboration with commercial and academic partners, we have often seen synthetic data integration oversimplified as a drop-in replacement for data collection as part of an AI strategy. While this can work, domain and technical knowledge used for stand-alone synthetic data efforts often isn't transferrable to future projects. Synthetic data implementation shines in an iterative workflow that enables experimentation with real-world scenarios and in which performance is probed with multiple engineered datasets.

ABOUT THE AUTHORS

Chris Andrews is COO and Head of Product at Rendered.ai, helping customers overcome the costs and limitations of using real-world data to train computer vision AI systems. Chris previously led a team at Esri responsible for 3D, Defense, Urban Planning, and AEC products. Prior to Esri, Chris was the lead product manager for Autodesk's InfraWorks. Chris is currently co-chair of the USGIF Modeling, Simulation, and Gaming working group.

Mark Hogsett is Rendered.ai's Head of Growth. He brings 30 years of National and Homeland Security experience and is a veteran of the United States Marine Corps. Mark is responsible for Rendered.ai's growth within federal, state, local, tribal, and territorial government markets.

Synthetic Computer Vision Data helps Overcome AI Training Challenges

Chris Andrews, Mark Hogsett
Rendered.ai
Seattle, WA
chris@rendered.ai, mark@rendered.ai

PERVASIVE USE OF AI

As forecast in many reports, artificial intelligence (AI) and machine learning (ML) are together becoming an essential tool for operational competitiveness by US commercial and government organizations (Bacastow 2021). Whether on the battlefield or in the supermarket, AI enables the rapid processing of increasing streams of data that would otherwise be impossible to support with human labor with adequate time and cost to facilitate informed decision-making. From the ground into space, the rate and volume of sensor-based data collection and telemetry about the entire world around us is accelerating at a pace that makes it essential for us to deepen AI adoption.

Typical AI Applications

Today, AI has become defined by a set of building blocks that, when used with other computational tools, can be combined into complex analytical and inferential systems that help humans process, analyze, and extract information from data. Typical primary applications include clustering or classification of specific data points, identifying and counting instances of occurrences of a specific data value, or identifying relationships between individual data instances (Szeliski 2011). AI techniques are useful for identifying relationships across complex, high-volume datasets, sometimes in ways that are nearly impenetrable to human analysis.

Analyzing Pixels and Text (Symbols)

Generally, AI techniques are applied to two different types of data inputs. Analysis of symbolic information, sometimes referred to as 'structured' data, is defined by inherent information in and between symbols that obey common patterns of usage, such as for words in a book or numbers in medical laboratory results. Data may be referred to as 'unstructured' when the specific data values have limited inherent meaning out of context of the spatial arrangement of values around them. This typically occurs in sensor-based data capture, especially for imagery or lidar. Sensor-captured data that is composed of pixels or similar spatially structured data storage may be referred to as computer vision (CV) or machine vision in different industrial domains.

This discussion will mostly relate to CV analysis of sensor-based imagery data.

THE AI TRAINING AND THE DATA PROBLEM

The process of developing an AI algorithm, referred to as training, involves the processing of large datasets to establish patterns between different instances of tokens or values as they occur in the dataset. Additional datasets are often required to validate training performance.

For some types of analysis, such as when searching for specific objects or scenarios in previously unanalyzed datasets, algorithms are trained through a process called supervised learning. Supervised learning requires training and validation on datasets that have been preprocessed to have additional information, typically referred to as labels or annotation. Labels are used to indicate, for example, what pixels in an image might be an instance of a car, how many cars are present in an image, and even whether specific scenarios of different cars may be present.

Training algorithms to analyze imagery, may require preprocessing millions of instances of data, especially when attempting to detect rare objects or scenarios. Data labeling, as this preprocessing is referred to, may required a combination of human and machine analysis. This type of acquisition and processing of training datasets presents challenges.

Cost and Time to Acquire Real Datasets

Real datasets can be expensive to acquire and time consuming to collect. Cost of data can be impacted by the technical complexity of creating or collecting the data and the specificity of curated data required to build specific types of algorithms. Dataset collection time can differ widely depending on sensor availability and frequency with which a sensor is in the right place at the right time to capture required data. Recollection or reacquisition, in the event of insufficient data or discovery of bias, will add significant expense to most training budgets.

Inaccurate and Expensive Data Labelling

Labeling of real datasets can be imprecise and expensive. Some types of datasets, such as x-ray and radar, can require specialist expertise for data labeling which can be both costly and time intensive. Both human and automated data labeling carry inherent inaccuracies leading to degradation in value of resources invested in data collection and acquisition.

Bias

All datasets have bias. Rare and unusual objects and events can be difficult or impossible to capture in real sensor data at quantities and qualities necessary to train algorithms. Scarcity of specific objects or scenarios of interest in a training dataset can lead to consistent error, referred to as bias, when algorithms are applied to real datasets. Overcoming scarcity with real datasets can require multiplying the size of datasets required to have reduced bias.

Innovation Challenges

Some sensors can be difficult to acquire or may not yet be deployed in the field. Without real data, developing analysis tools or processing pipelines on new sources of data can be impossible.

Privacy or Security Limitations

Real datasets may have privacy, security, or safety limitations. These issues carry inherent risk that often increases with the size of the datasets required to train algorithms.

INTRODUCING SYNTHETIC DATA

Simulated or synthetic data has become an alternative and supplement to the use of real sensor data for training and validating algorithms in some cases. Widely used in industries such as the autonomy industry, synthetic data can be engineered to have properties or characteristics intended to produce specific training and validation outcomes. Synthetic data has the potential to overcome many of the challenges of collecting and using only real datasets to train and validate algorithms.

Generally, there are two types of synthetic data produced for training AI, simulation-based synthetic data and algorithmically produced synthetic data. Each carries different costs and benefits.

Physics-based Simulated Training Data

Modern techniques for 3D modeling, material definition, fluid and electromagnetic physics modeling, and animation can be combined to create dynamic 3D and 4D environments that closely model real-world scenarios. Combined with digital models of physical sensors, these environments enable the simulation of physics-based sensor data capture, including modalities such as visible light, hyperspectral imaging (HSI), multispectral imaging (MSI), x-ray, and radar.

Synthetic data produced through physically accurate digital simulation can be inexpensive to create in large quantities once simulations are configured and deployed. Often, synthetic data applications can be set up and running in hours or weeks, faster than real dataset acquisition in many cases.

Given that 3D environment modeling is explicitly under the control of the software engineer, synthetic data can be generated with any variety and types of objects, scenarios, and environmental conditions required to simulate real imagery. Because of the explicit knowledge built into the process of simulating systems, synthetic data produced through simulation can be 100% accurately labelled and can be produced with a wide range of metadata useful for improving algorithm training and validation, including characteristics such as precise geospatial positioning.

In circumstances in which real data are unavailable because of privacy or security risk, synthetic data is simulated and can be created to fit scenarios where customers are unable or unwilling to work with sensitive or difficult to collect information in fields such as medical, insurance, or defense. Similarly, when developing algorithms to interpret data from future sensor types or configurations, synthetic data can substitute for real data, enabling testing of analytics and infrastructure configuration for anticipated data processing.

Using Generative AI for Training Data

The use of techniques such as Large Language Models and Diffusion Models to generate text and imagery are collectively referred to as Generative AI (Taulli 2023). Generative AI models are in the early stages of being explored as tools for generating synthetic datasets for AI training with more current adoption in the text-based AI training space.

Generative AI models are developed by ingesting large datasets into algorithms that, after training, can be used to create representations of data instances or even entire datasets through a descriptive input, typically called a prompt. The results of using a generative model are determined by the diversity and distribution of data that was used for model training. That is, generative models cannot usually create novel data instances or datasets that fall outside the statistical combinations of input instances in training data.

As a result, Generative AI models can be extremely good at generating new datasets that emulate source data, which is highly useful in cases in which security and privacy need to be preserved. There is also potential for Generative AI to be used to produce content for use in 3D simulations. Research in this area is limited but shows promise.

Generative AI vs. Simulation-based Synthetic Data

There are a few differences between Generative AI created data and physics-based simulated data that should be considered when presented with the option to use one or the other. Table 1 captures some of these differences.

Table 1. Characteristics of Synthetic Computer Vision Data Created through Generative AI vs. Physics-based Simulation

	Variations of Synthetic Data Creation Mechanism	
	Physics-based data simulation	Generative AI derived data
Diversity	Unlimited diversity including ability to introduce scenarios and objects that are impossible to capture in real sensor data	Limited to generating datasets with diversity that falls within the statistical population characteristics of input training data
Physics-based accuracy	Constrained by limitations of physics-based mathematical models and compute. Demonstrated capability across many domains such as HSI, MSI, and radar.	Limited research has been done to quantify how physically accurate Diffusion-generated data may be.
Dataset size limitations	Limited only by device memory to perform calculations and store data which tend to be scalable across time and instances. Authors are not aware of research comparing dataset size limitations of these approaches.	

Simulated image capture size limitations	Conceptually limited only by ability to parallelize compute of partial simulated collections which is possible across most sensor domains.	Diffusion models, typically focused on known output sizes that mimic input data sizes, are known to require exponentially more compute for large images, potentially limiting size of generated output (McKeag 2023).
Sensor domain applications	Possible to simulate any sensor domain for which there are physics-based models. Effort will depend on complexity of the domain.	Relies on adequate training data with the implication that training datasets may be unavailable or costly for unusual or rare sensor types and applications.
Traceability / Explainability	Explicit control over simulation and scene content conveys perfect ability to trace and know output data provenance at the pixel-level	Limited ability to trace and explain generative output past the training data input stage. Risks of adversarial data injection into very large training datasets or through parent LLM and Diffusion model training sets.
Configurability	Initial setup requires knowledge of 3D, simulation, and physics which can be an obstacle. Subsequent configuration can be done by less technical users through commercial off-the-shelf tools (COTS).	After initial model training and refinement, which requires data science skills, data may be generated through Prompt Engineering which is a newly developing field that has the potential to be accessible to users with domain expertise, but less technical ability.

EXAMPLES OF SYNTHETIC DATA APPLICATION

Physics-based synthetic data has been in wide use in some industries, such as the automotive industry, for several years (Toews 2022). In our experience as a commercial provider of synthetic data tools, we have gathered evidence of synthetic data application in some of the following cases.

Rare Object Detection in Satellite Imagery

In a Phase I and Phase II SBIR project with the NGA, Rendered.ai supported Orbital Insight, Inc. by creating synthetic data applications that generated imagery for purposes of investigating dataset combinations that could improve the detection of rare objects in the xView dataset, such as construction cranes. Orbital Insight was able to generate hundreds of datasets with thousands of simulated images and demonstrated that use of synthetic imagery could increase detection capability with synthetic data. As shown in Figure 1, use of synthetic data alone increased algorithm performance over use of real data alone, however an even bigger boost was seen when real and synthetic data were combined together (Weber 2021).

Another conclusion of the study was that use of synthetic data has the potential to significantly reduce the number of real data instances required for training. Given the acquisition and labeling costs for use of real data, any opportunity to reduce the total quantity of real data used per training instance has the potential for considerable additive impact on detection capacity when confronted with limited budgets.

The study also pioneered the use of CycleGANs (Zhu 2020) to post-process real remote sensing imagery. The CycleGAN is trained on real imagery and then used as a filter to modify synthetic data, making the synthetic data behave more as if it is real data to an algorithm. CycleGAN usage in this manner is an example of combining physics-based simulation and generative tools to arrive at an optimum outcome for creating synthetic training data.

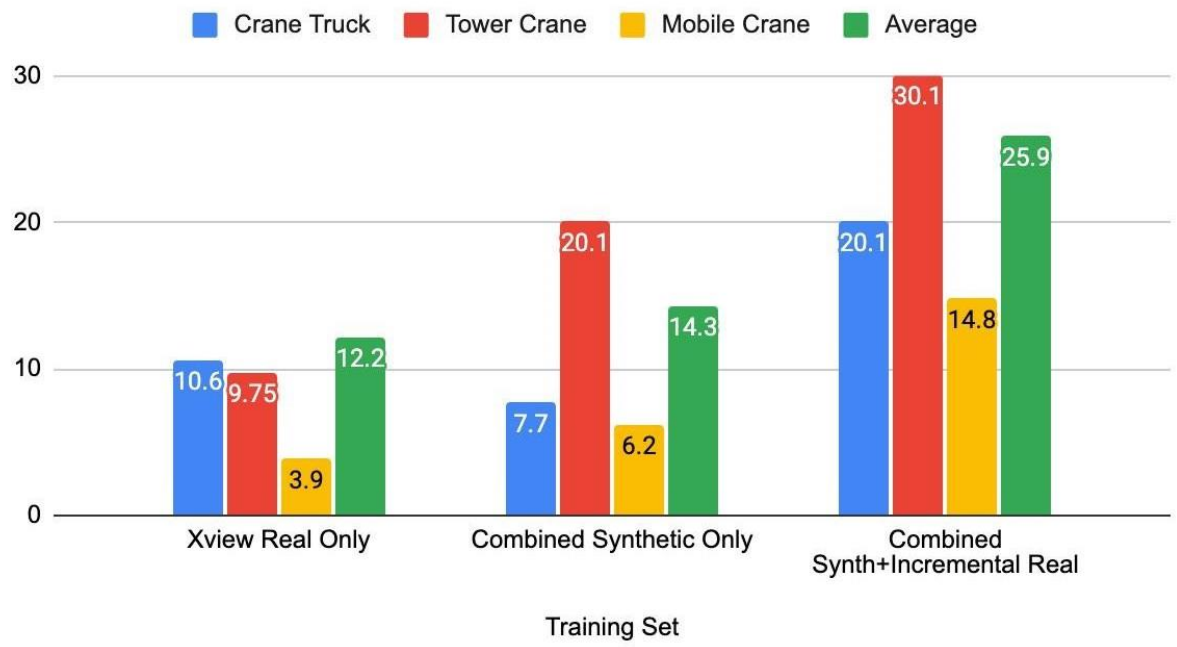

Figure 1. Average Model Precision Per Training Set Showing Improvement of Rare Object Detection Over Real Data with Synthetic Data Only and with Synthetic and Real Combined. (Weber 2021)

Improved Defect Detection in Manufacturing

Eigen Innovations, a Toronto-based manufacturing monitoring company, used Rendered.ai to build a synthetic data application to expand their idealized digital part imaging workflow with randomized variation in simulated imagery (Eigen Innovations 2023). The Eigen team built the capability to introduce both pose variation, by changing simulated camera position relative to manufactured parts, and random defects, including simulated scratches and paint smudges.

The team discovered that they could increase defect detection by approximately 21% with the introduction of synthetic data. However, the team also discovered that introducing variation in synthetic data that is not encountered in real data can degrade algorithm performance. In this case, the addition of pose variation was detrimental to training performance which was later attributed to the fact that their physical system never, in fact, has any variation in pose in captured imagery. This offers a direct example of how synthetic data needs to be precisely designed to address desired algorithm training outcomes.

Transfer Learning

Transfer Learning, a technique for augmenting a pre-trained AI model with additional training steps using data not in the original training data, offers significant opportunity for use of synthetic data to both focus the detection scope of an algorithm and to also overcome gaps in source data that may introduce detection bias. In the Open Geospatial Consortium's (OGC) Testbed-19, Rendered.ai was able to demonstrate that synthetic computer vision data can be used to effectively improve detection capability when starting with a generic, unspecialized model (Lavander 2024).

In this exercise, the focus was on 'zero shot' training or training with no actual instance of real data for training. The team experimented with techniques such as CycleGAN domain adaptation for purposes of improving synthetic data performance (Figure 2). The effort also demonstrated that synthetic data can be used to test the boundaries of detection capability, such as by introducing unexpected variation not found in real datasets along with predictable failures in detection results.

Figure 2. Left: Original Simulated Synthetic Image. Right: Resulting Image After Modifying the Original Image with GAN-based Domain Adaptation, a Post-processing Technique used to Enhance Domain Match (image from Lavander 2024).

Dataset Standardization for Analysis Ready Data

In another component of the OGC's Testbed-19, Rendered.ai used emerging OGC standards to create a synthetic data application that could produce idealized packages of Analysis Ready Data (ARD) (Li 2023). The application generates data that models existing commercial satellites with correct georeferencing and physical characteristics, metadata, and annotation for purposes of testing and validating AI training and detection pipelines. The effort demonstrated that synthetic data applications can be produced that allow data scientists to use standardized interfaces to develop baseline datasets for training and validation. Baseline ARD datasets in this case can include variation for physically precise atmospheric variation, ground cover materials, time of day, geospatial position, and sensor bands and sampling distance (Table 2).

Table 2. Specifications for Commercial Satellite Platforms Simulated in the Testbed-19 ARD Project. (Li 2023)

PLATFORM TYPE	SENSOR APPROXIMATION	GSD	SPECTRAL BANDS	ARRAY SIZE
Medium resolution EO	Maxar WorldView-3	~1.24 m	9-channel VIS+NIR	640 x 480
High resolution EO	Planet SkySat 16-21	~0.75 m	5-channel PAN+VIS+NIR	1024 x 768

Validating AI Pipelines for Future Satellite Sensor Data Collection

Planet Labs PBC is planning the launch of a new hyperspectral satellite constellation, Tanager, in 2024. Planet's new sensors will collect hundreds of channels of imagery across a broad swath of visible and invisible light spectrum, resulting in massive images with novel information that has not yet been available for robust analysis. Using Rendered.ai and the DIRSIG simulator from the Rochester Institute of Technology, Planet was able to build a synthetic data channel, or application, that simulates full Tanager collects (Andrews 2023). The DIRSIG model was specifically built with the intent to enable simulation of accurate light wavelength collection for experimental purposes (Goodenough 2017).

In this project, the Planet mission team was able to simulate raw examples of imagery as if it was directly collected from a sensor deployed in orbit. The team used content from the developed application to build and test algorithms

for analyzing pixel and channel-level variation in imagery, representing small variations in photon collection across massive images that would have been otherwise unavailable until a vehicle was deployed into space.

Figure 3. Simplified Examples of Synthetic Hyperspectral Data using Rendered.ai and the DIRSIG Model Showing Variation in Scene Content, Spectrum, and Atmospheric Conditions. (Andrews 2023)

TECHNIQUES AND PROCESS FOR APPLYING SYNTHETIC DATA

For any technology adopted for enterprise use, the implementation and capabilities must meet the operational and maintenance requirements of the organization deploying the technology. Synthetic data for training AI systems is a relatively new application of technologies and tools that span multiple decades and continue to rapidly change. The basic implementation tends to be best informed by understanding core concepts such as requirements, problem type, and tolerance for experimentation and support.

Understanding the Problem

The authors and their colleagues have worked with multiple organizations and agencies in the emerging computer vision space. In early-stage markets, understanding of new tools and their applications can be heterogeneous. Often the intended beneficiary of a new technology is not an expert in that technology and will be focused on the problem already at hand. In the computer vision market, this typically manifests as data scientists who are habituated to working with the data that they can acquire and who have never had to go through the exercise of designing the ideal dataset that they believe might fit their need.

Right Simulation for Right Job

Depending on the sensor domain and the type of analysis or process being investigated, different simulators or generation types will have varying relevance for specific AI training and validation efforts. In cases where precise pixel-level accuracy is required, physics-based simulation techniques are the only path in today's technology landscape. The specific type of analysis may drive the technical sophistication of the simulator applied. Standard ray tracing tools from the entertainment industry may suffice for many visible light simulation scenarios. Highly accurate validated physics models, such as DIRSIG, will be required for more technically sophisticated sensor types and analysis. Generative tools offer new opportunities for augmenting data sets and creating data that are just being explored today.

Bespoke Implementation vs. Open Source vs. COTS

The problem domain and the availability of proven tools to simulate relevant data will drive the pattern of implementation and acquisition of synthetic data technology. Bespoke technology stacks or bespoke components may be required in cases where novel sensors or technical specificity are required, and which have not been previously addressed in COTS tools.

With the need for diverse skills in physics, 3D, ML, and data collection, organizations will be better served if they can acquire off-the-shelf solutions, either open source or commercial, or minimize custom technical implementation to isolate skill gaps to areas that are feasible to be hired or outsourced.

In the area of AI training, which may require collection or generation of ongoing datasets, cloud compute offers the best path to scale data generation and processing, but also presents a technical barrier for some organizations without staff or knowhow for cloud-based software implementation.

THE FUTURE OF SYNTHETIC DATA

As applications for computer vision diversify and specialize, cost and availability of data will be critical constraints on the ability of organizations to deploy AI systems. Evolving operational needs will drive the continuous need for additional data instances and new data sources. Synthetic data is already being adopted for workflows in US intelligence agencies, especially when availability and cost of training data can impact timeliness of intelligence analysis (Andrews 2023).

Gartner and other organizations believe that synthetic data will become the dominant type of data used for training AI and ML systems in the next decade (Woodward 2022). With benefits such as the ability to create data on demand, the reduction in quantities of real data required for training (Chandrasekaran 2022), and opportunities for standardizing data packaging and traceability, synthetic data is already being shown to have tangible positive impacts on cost and performance when training and validating AI algorithms.

Looking forward, with the adoption of AI broadly across organizations such as the US Department of Defense and in civilian agencies, such as the US Department of Energy, synthetic data may offer opportunities for establishing more explainable boundaries of algorithm effectiveness and for detecting injection of adversarial data during training. These areas are in early research, but offer significant opportunities for synthetic data to be part of the defense and explainability of decision support systems that will have critical import for national security, supply chain, and infrastructure.

REFERENCES

Andrews, C., Fleming, S., Kenney, P., Mohammed, S., Widener, D., Lepretre, S. (2023) *The Evolving Role of Synthetic Data in GEOINT Tradecraft.* Herndon, VA: USGIF, from https://trajectorymagazine.com/file/the-evolving-role-of-synthetic-data-in-geoint-tradecraft/

Andrews, C. (2023) Case Study: Synthetic Data Reaching New Heights with Rendered.ai and Planet, Rendered.ai blog. Retrieved on February 2, 2024 from: https://rendered.ai/case-study-remote-sensing-synthetic-data-with-planet/

Bacastow, T.M., Knapp, M., Neary, P., Park, S., Polaski, D., Ronlov, M., & Spugnardi, J. (2021) *The State of Artificial Intelligence and Machine Learning (AI/ML) in the GEOINT Community and Beyond.* Herndon, VA: USGIF, from https://usgif.org/wp-content/uploads/2021/09/USGIF-White-Paper-3-AI-ML.pdf

Chandrasekaran, A., Linden, A., Mullen, A., Choudhary, F. (2022) *Innovation Insight for Synthetic Data*, Gartner: ID G00757632.

Di, L., Meyer, D.J., Yu, E. (2023 pre-publication) *Engineering Report for OGC Testbed 19 Analysis Ready Data*, Open Geospatial Consortium.

Eigen Innovations, Inc. (2023) Scalable Deep Learning Models in Manufacturing - A Surface Inspection Case Study, Technical paper. Available at: https://eigen.io/imagetwinwhitepaper/

Goodenough, A.A., Brown, S.D. (2017) *Dirsig5: Next-generation remote sensing data and image simulation framework*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 11, pp. 4818-4833.

Lavander, S., Tinker, T. (2024 pre-publication) *Testbed-19: Machine Learning Models Engineering Report*, Open Geospatial Consortium.

McKeag, B. (2023) *Why Altering the Resolution in Stable Diffusion Gives Strange Results*, RunPod Blog. Retrieved on February 2, 2024 from: https://blog.runpod.io/why-altering-the-resolution-in-standard-diffusion-gives-strange-results/

Szeliski, R. (2021) Computer Vision: Algorithms and Applications. London: Springer-Verlag.

Toews, R. (2022) *Synthetic Data Is About To Transform Artificial Intelligence*, Forbes.com. Retrieved on February 2, 2024 from:

https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=514f25c77523

Taulli, T. (2023). Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business. United States: Apress.

Woodward, A., Chitkara, V., Jury, B., Hare, J. (2022) *Emerging Technologies: When and How to Use Synthetic Data*, Gartner: ID G00751766.

Zhu, J., Park, T., Isola, P., Efros, A.A. (2020) Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv: 1703.10593.