VISION SYSTEMS DESIGN

EXECUTIVE SUMMARY

Machine Vision Reimagined: Step into the Future with Synthetic Data

Brian Geisel, CEO, Geisel Software and Symage, Inc.

OCTOBER 9, 2024

KEY TAKEAWAYS

- Achieving accurate machine vision depends on high-quality images and video.
- Challenges due to data availability and preparation requirements can slow model training.
- Symage by Geisel Software rapidly generates high-quality, bias-free, pre-labeled image data.

in partnership with

SYMAGE By Geisel Software

Machine Vision Reimagined: Step into the Future with Synthetic Data

OVERVIEW

Synthetic data is revolutionizing model training. Real-world data is often scarce, expensive, and fraught with privacy concerns, creating a significant bottleneck in AI development. Forward-thinking companies are leveraging synthetic data to overcome these challenges. Creating and utilizing synthetic images can enhance model accuracy and performance for superior machine vision applications.

To empower AI developers with the data they need to innovate faster, more cost effectively, and with greater confidence in the accuracy and reliability of their models, Geisel Software developed Symage, a state-of-the-art custom synthetic image generator designed to meet the needs of modern AI systems. Symage delivers photorealistic, high-fidelity synthetic data, tailored to simulate a wide range of scenarios. This innovative technology creates custom datasets that fill the gaps in real-world data, enhancing model performance and ensuring that AI systems are prepared for any challenge.

CONTEXT

Brian Geisel discussed the value of synthetic data in model training and explained how Geisel Software's Symage generates synthetic images.

KEY TAKEAWAYS

Achieving accurate machine vision depends on high-quality images and video.

Traditional computing uses a *deterministic* approach, mapping the flow of a process based on pre-defined conditions. This logical approach maintains a finite boundary around machine capabilities; however, that limitation is both useful, such as when debugging code, and restrictive, in that code could not enable open-ended tasks.

The advent of machine learning (ML) has upended the traditional approach.

Although there is still code involved in ML, it is code that allows the system to not only execute tasks, but to make its own decisions about what to do based on context. The ML model can do this because it is not static, but rather it *learns*—by ingesting data, drawing conclusions from that data, and applying probabilities to make decisions on how to move forward.

However, training machine learning models depends on the availability of high-quality, high-quantity data. Training a model for machine vision (enabling a machine to "see" its environment) requires additional labeling of the data, as an image is only as useful to a model as the label that identifies what the image represents.

> "For any chef, [your dishes] can only be as good as your ingredients."
>
> *Brian Geisel, CEO, Geisel Software and Symage, Inc.*

Challenges due to data availability and preparation requirements can slow model training.

When training for machine vision, it takes the same amount of time to label data as it does to obtain the data. However, because training a model to recognize a single object requires an initial ingestion of millions of data points and corresponding labels, the time investment—and associated slow speed—can present a significant hurdle to progress.

Although there are some AI tools that will label data, labeling errors can occur using this approach. Poorly

Machine Vision Reimagined: Step into the Future with Synthetic Data

labeled data is worse than unlabeled data, as it trains a model on incorrect information, leading the model to make wrong decisions.

Figure 1: Poorly labeled data is detrimental in model training

As machine learning applications extend into new areas and are leveraged for more complex tasks, having readily available, high-quality data in large quantities is becoming increasingly critical. However, data availability can be limited due to:

- **Lack of access to the environment** (e.g., photos of the surface of Venus).
- **Data scarcity.** Even when accessibility is not an issue, sometimes sufficient data simply does not exist.
- **Lack of diverse conditions,** such as when training a model to understand weather.
- **A need for highly accurate 3D information** to understand positioning, depth, distance, etc.

Synthetic data provides a solution to these challenges.

Although time-series synthetic data, which enables applications such as predictive maintenance, has existed for at least a decade, it is only in recent years that image and video synthetic data have become more popular. This is due in large part to the advent of more powerful NVIDIA chips, which provide the compute necessary for transformer models—capable of ingesting complex data, such as image and video—to be effective.

These technological advancements, including improved computer vision capabilities, are now driving demand for synthetic data to aid in machine learning and extend intelligence to new applications.

> "Who knew that fooling machines could be the key to unlocking their full potential?"
>
> *Brian Geisel, CEO, Geisel Software and Symage, Inc.*

Symage by Geisel Software rapidly generates high-quality, bias-free, pre-labeled image data.

Symage is the breakthrough synthetic image data generator backed by Geisel Software's years of expertise in robotics, AI/ML, and computer vision. Symage's advanced rendering techniques generate high-fidelity, bias-free image data and enable pixel-precise accuracy in data labeling.

Machine Vision Reimagined: Step into the Future with Synthetic Data

There are a number of way to generate synthetic image data, including:

- **Generative adversarial network (GAN)**, which uses two models. One is the Generator, which generates data (a set of random pixels). The other is the Discriminator model, which is well trained to identify whether the image generated is an accurate representation of what has been requested. The Discriminator's responses help the Generator create more accurate images over time.
- **Diffusion,** which gradually adds noise to the data, after which the model learns to reverse the process. This method produces high-quality, diverse images, but can have unwanted artifacts in the images, such as incorrect text or illegible fonts.
- **Variational autoencoder (VAE)** and other data augmentation techniques, which can generate synthetic data by varying an image's properties.
- **Procedural generation,** which relies on programmed instructions for how to create a given image (e.g., mirroring, contrast ratio).

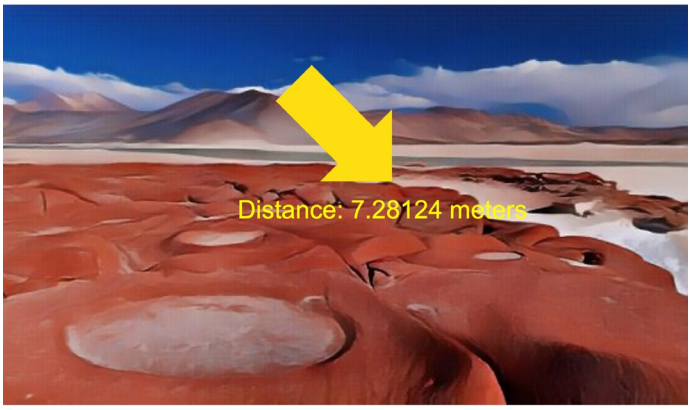
When Symage generates synthetic images, the data, labels, depth map, semantic segmentation, and other attributes that can help the model are all generated simultaneously. Symage's automated labeling process means zero time is spent on manual efforts, allowing efforts to remained focused on training and refining AI models.

Symage's physics-based simulator creates photorealistic scenes that perfectly mirror real-world conditions, enabling adjustments to environmental variables like lighting, camera angles, and object placements to simulate complex scenarios and edge cases with precision.

Figure 2: A GAN uses adversarial training to refine and generate more accurate images

Machine Vision Reimagined: Step into the Future with Synthetic Data

Figure 3: Symage includes highly accurate 3D information with generated data



Scene-based synthetic data can also simulate real-world situations, which can be especially useful for testing dangerous, impractical, or even impossible conditions, and/or deriving long-term assessments from a rapid, condensed simulation.

Looking to the future, multimodal model training can provide greater context to enable even more complex applications, such as synthesizing visual, light, and audio, or advanced detection uses such as chemical and radiological threat sensing.

SIMULATING THE MARTIAN ENVIRONMENT

How Geisel Software helped NASA train a model to recognize a landscape 140 million miles away.

Training models that enable rovers to drive on Mars is an exceptional challenge, as the available real visual data about the Red Planet is extremely limited. To provide its rovers with the intelligence to navigate Mars, NASA needed to close the enormous simulation-to-reality gap with synthetic data.

NASA approached Geisel Software with the request to develop a Martian simulation that could create data for model training. Using available footage from Mars, Geisel Software compared elements from the Martian landscape, such as rocks, to similar elements that existed on Earth, and for which quality 3D models already existed.

With the combined existing Mars images and similar Earth objects, Geisel created a simulation of the Martian landscape. However, the model also had to understand the varying and extreme conditions on Mars, including different atmospheric effects. By adding noise, adjusting the atmosphere, and shifting color within the simulated landscape, a wide range of data that reflected environmental variation was generated for model training—and all data was labeled with detailed information about each object's properties.

Machine Vision Reimagined: Step into the Future with Synthetic Data

The combined synthetic images and pixel-level labels allows the model to fulfill specific requests. For example, semantic segmentation made possible by the detailed information generated by the simulated environment enables the model to search for and identify hematite rocks, which indicate a high likelihood of a past presence of water and a potential location where traces of life might be found.

> "The semantic segmentation of this exact same scene is pixel-for-pixel identical to the other. But here . . . blue is stuff you want to be able to identify, green is not. This is one of the things that we're able to do in this scenario, which [enables] something as awesome as searching for particular rock formations via AI on Mars."
>
> *Brian Geisel, CEO, Geisel Software and Symage, Inc.*

Figure 4: Simulated environments based on real footage were used to create synthetic training data

Footage from Mars | Scanned Earth Objects | Create Simulation

Figure 5: Semantic segmentation allowed the model to identify the potential presence of hematite

Original Object | Instance Labeling | Segmentation of Objects of Interest

Machine Vision Reimagined: Step into the Future with Synthetic Data

Best practice recommendations for using synthetic data

Synthetic data does not apply to every training set. If non-synthetic data already exists in some capacity, it is often best to begin training the model with data before supplementing with synthetic data.

To determine where synthetic data fits best, it is critical to have a benchmarking system in place. Identifying where a model is falling short depends on understanding how the model is performing, to start with. The quantity and type of synthetic data will depend on which areas of the model need to be improved.

There is no special requirement for testing the efficacy of synthetic data. With a benchmarking system already in place, the model's measurements will reflect whether and how much the synthetic data helped.

ADDITIONAL INFORMATION

To learn more visit:

- Geisel Software
- Symage
- Symage Mars Case Study

BIOGRAPHY

Brian Geisel

CEO
Geisel Software and Symage, Inc.

Brian Geisel is the CEO of Geisel Software and Symage, Inc., where he leads pioneering advancements in artificial intelligence, machine learning, and robotics. With a focus on AI-driven applications, autonomous systems, and synthetic data, Brian has dedicated over two decades to tackling some of the most challenging problems in technology. His expertise extends across embedded systems, AI/ML, deep learning, and automation.

Under Brian's leadership, Geisel Software has earned a reputation for creating innovative, resilient solutions for high-stakes fields. The company's partnerships with NASA and major robotics firms exemplify Brian's commitment to precision, reliability, and pushing the boundaries of what's possible in AI and synthetic data integration. Through his visionary approach, Brian and his team deliver trusted, state-of-the-art software solutions that shape the future of intelligent systems and autonomous technology.