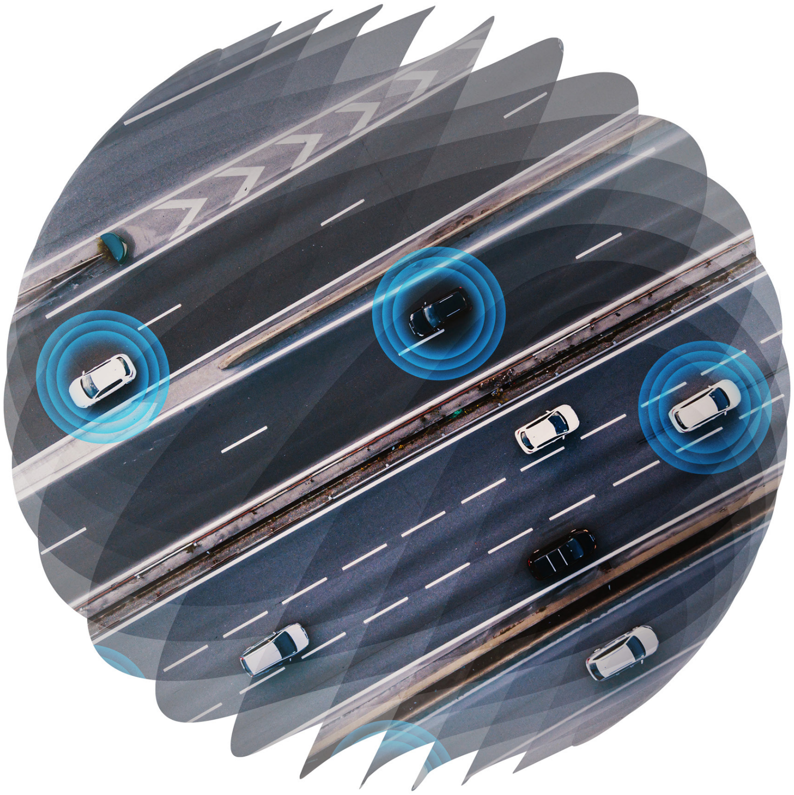
Deloitte.



Using synthetic data to enhance the development of autonomous driving

Despite the tremendous progress being made in autonomous driving technology, high-profile incidents have highlighted some of the significant challenges that remain. From vehicles becoming immobilized to tragic pedestrian accidents,[1] these events underscore the importance of rigorously testing and comparing self-driving systems across a vast array of scenarios.

Some of these incidents can be attributed to rare edge cases that are exceptionally difficult to encounter and analyze through real-world testing alone. Having said that, autonomous vehicles should be prepared to handle a wide variety of complex multi-agent interactions, adverse weather conditions, construction zones, and other potentially hazardous situations.

Collecting sufficient real-world training data to cover this spectrum can be an immense challenge. Road testing and manual data annotation are likely time-consuming, costly, and may be impractical for capturing infrequent events. This is where synthetic data generation is emerging as a powerful tool to augment and diversify the training datasets for autonomous driving perception systems.

By leveraging advanced simulation platforms, it is possible to create virtually limitless permutations of environments, weather conditions, sensor configurations, and edge cases, all with precise, pixel-level annotations. This capability can allow developers to rapidly iterate, test, and validate their autonomous driving solutions, ultimately paving the way for safer deployments on public roads.

In fact, synthetic data is expected to play an increasingly important role in model training for autonomous driving, fundamentally reshaping the automotive industry's approach to solving data-centric challenges.

Where real-world data falls short

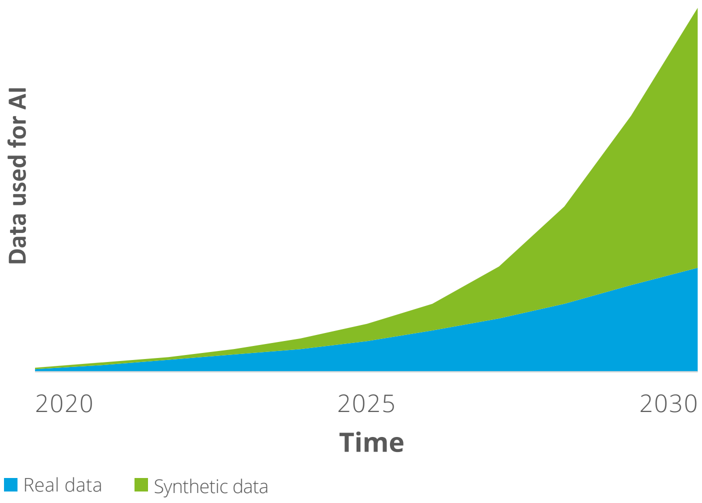
When it comes to model training for autonomous driving, the need for high-quality, annotated data is critical. Traditional data collection methods involve operating fleets of sensor-equipped vehicles that gather data from real-world environments. While effective, this approach is plagued by several important issues. First, collecting and annotating real-world data is an arduous task that requires a significant investment of time and resources. Autonomous vehicles can generate up to 8 terabytes (TB) of data per day,[3] and the manual labor required to label this data is likely unsustainable. Second, the range and diversity of scenarios that real-world data is able to capture can be limited. Rare edge cases, though infrequent, hold disproportionate importance as litmus tests for the robustness and safety of autonomous vehicles.

In addition, there are ethical and legal concerns to consider. Data privacy regulations are in place for certain types of data, thereby limiting the scope of real-world datasets that can be collected and used. Biases inherent in real-world data also pose significant challenges, particularly when algorithms trained on such data are expected to operate in diverse and unpredictable environments. For instance, if autonomous driving algorithms are primarily trained on data collected from urban environments, they may struggle to navigate rural roads or handle scenarios unique to countryside settings.

Why synthetic data should be the way forward

By definition, synthetic data is information generated via computer algorithms or simulations, designed to mimic the properties of real-world data. Unlike traditional data, synthetic data can be generated quickly and in large quantities, drastically cutting down on time (up to 80% time savings in data generation and validation) and resource investments (up to 95% cost savings).[4] Notably, synthetic data can be obtained directly from simulation platforms complete with ground truth labels. This includes labels such as semantic segmentation, in which each pixel in an image is assigned to a specific category (e.g., “car” or “road”); bounding boxes, which are rectangular coordinates that identify the location and size of an object in an image; and point clouds, which are a set of data points in space often used in 3D modeling or computer vision to represent the external surface of an object or environment. The inclusion of ground truth labels within the simulation is important for streamlining the annotation process and accelerating algorithm training in various industries, including autonomous driving. Moreover, synthetic data offers a controlled environment to run a wide range of scenarios, helping capture real-world situations such as chairs flying across a highway or a jaywalking pedestrian. Finally, the risk of data privacy issues is eliminated because synthetic data is generated and not collected.

Figure 1: Growth of synthetic data use for AI models (2020 – 2030F)



Source: Gartner[2]

Creating synthetic data with an open-source autonomous driving simulator

The integration of synthetic data into the development process of autonomous driving systems underscores a broader industry trend toward leveraging digital simulations to overcome the constraints of real-world data collection. Open-source simulators for autonomous driving research are emerging as essential tools in this arena. Open-source simulators not only provide a foundational suite of assets and vehicles, but also invite expansion to encapsulate a broader spectrum of real-world scenarios. By integrating and enhancing open-source simulator capabilities with additional assets, diverse environments and conditions can be simulated with a high degree of realism. This augmentation process broadens the simulated scenarios available, which is imperative for training robust and reliable autonomous systems.

Phase 1: Object development

Static objects are designed using an open-source 3D computer graphics software tool, helping ensure a high level of detail in textures and aesthetics. Each object is then assigned unique tags, making them identifiable for tasks like semantic segmentation. Within a 3D computer graphics game engine, Deloitte researchers established a hierarchical folder structure that houses these created assets along with their materials and textures. As these objects were integrated into an open-source simulator, the material and texture settings were replicated to maintain a coherent visual experience.

For dynamic assets, like vehicles, the process begins by retrieving vehicle skeletons from an open-source simulator's repository. These skeletons act as the structural basis upon which 3D models are constructed. One of the key steps involves the careful process of binding and alignment in the vehicle model. This procedure helps ensure the different parts of the vehicle fit together accurately and move in relation to each other, producing a realistic depiction of physics and movement. This process is essential for the model to behave and interact with its environment in a manner that mimics a real-world vehicle.

Phase 2: Integration into open-source simulator's inventory

Once the modeling is complete, the assets are exported as tailored FBX files. This is followed by the detailed configuration of various parameters such as physics, animations, and skeleton, preparing the vehicle for its final integration into an open-source simulator's native vehicle factory.

Phase 3: Simulation and data collection

The simulator offers diverse virtual environments, known as "towns," each mimicking real-world scenarios from busy urban areas to rural settings. These towns can be populated with dynamic entities like vehicles and pedestrians, providing a robust dataset for training machine learning models on various traffic conditions. Additionally, the simulator incorporates a range of weather conditions, allowing for comprehensive testing of autonomous systems under different atmospheric conditions.

Within these environments, the "ego" vehicle can be autonomously navigated through complex scenarios that contain many randomizations, collecting ground truth data through integrated sensors. These sensors are designed to collect diverse data types. Forward-facing cameras provide visual context, while depth sensors assist in measuring object proximity. The system also employs both semantic and instance segmentation to isolate and differentiate scene components and individual objects. Additionally, a simulated light detection and ranging (LIDAR) sensor generates a 3D point cloud.

The synthetic data generated covers a broad range of scenarios, such as vehicle variations tailored to mimic emergency vehicles and varying traffic conditions. It also accounts for occlusions and distant objects to capture the intricacies of object visibility. Camera angle randomizations offer multiple viewpoints. In addition, the dataset is enriched by randomizations in weather and road conditions. The robustness and variability of the synthetic dataset was enhanced in a postprocessing step by randomly choosing images to undergo geometric and color space transformations, as well as noise injection.

Evaluating the impact of synthetic data on model training

The methodology employed by Deloitte researchers involved using a mix of synthetic and real-world data for model training. Initially, the model was trained solely on synthetic data, and later it was fine-tuned using available real-world data.

To comprehensively evaluate the efficacy of synthetic data, Deloitte researchers conducted a study focusing on firetruck detection using the fine-tuned YOLOv8 model. The test set included only real-world images. Ensuing experiments ranged from training models exclusively on synthetic data to gradually incorporating real-world data. This step-by-step approach led to several key insights:

1. **Mean average precision (mAP), precision, recall, F1 score:** These are all measures of the model's accuracy. *mAP* is the average precision at different recall levels, *precision* measures what percentage of detections were actually correct, *recall* measures what percentage of actual positives were identified correctly, and *F1 score* is the harmonic mean of precision and recall. Fine-tuning with even a small set of real-world data improved these measures.

2. **Overfitting mitigation:** Overfitting occurs when a model learns the training data too well, to the point where it performs poorly on new, unseen data. Synthetic data helped create a more balanced performance, unlike real-world data, which showed tendencies to overfit when the dataset size was small.

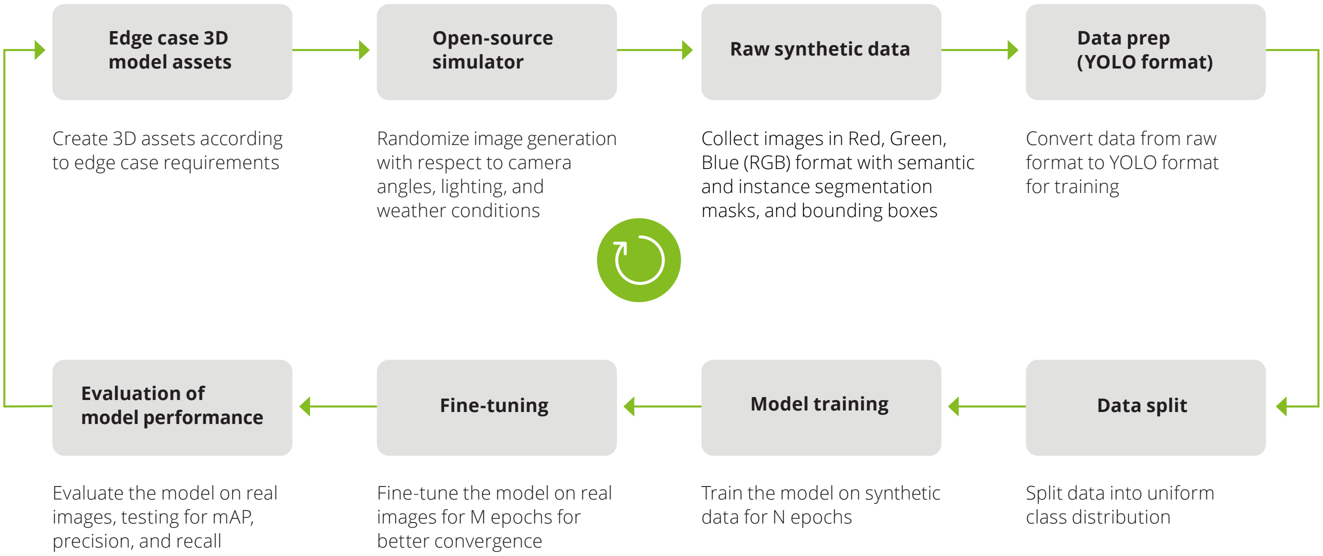
3. **Synergy and balance:** A careful combination of synthetic and real-world data resulted in models that were not only accurate but also comprehensive in their detection capabilities.

4. **Importance of synthetic dataset size:** Increasing the synthetic dataset size with various randomizations increased performance on all metrics, approaching the performance of the models trained only on real-world data, even without adding real-world images.

The experiments showed that when real-world data is limited, supplementing with synthetic data improved all measures of accuracy. Furthermore, in the absence of real-world data, generating a diverse and randomized synthetic dataset of sufficient size can bring the model's performance remarkably close to those trained on real-world data. This highlights the potential and value of synthetic data in enhancing model training and performance.

As the field of autonomous driving continues to progress, synthetic data will likely become an invaluable tool for training artificial intelligence (AI) models. Extensive experiments using an open-source simulator have shown that synthetic data can provide a solid foundation for model training. It helps to avoid the trap of overfitting, allows for specific customization, and can reduce data-related costs, particularly for rare or unusual scenarios. Furthermore, blending synthetic and real-world data can lead to AI models that are not only accurate but also resilient and capable of handling a variety of situations. Experiments also underscored the fact that when the availability of real-world data is limited, the accuracy of AI models can be enhanced by supplementing synthetic data. In situations where real-world data is completely absent, creating a large, varied, and randomized synthetic dataset can bring the performance of the AI models remarkably close to those trained exclusively on real-world data. This underscores the significant potential value of synthetic data in improving autonomous model training and performance.

Figure 2: Flow diagram—synthetic data evaluation via model training



Edge case 3D model assets

Create 3D assets according to edge case requirements

Open-source simulator

Randomize image generation with respect to camera angles, lighting, and weather conditions

Raw synthetic data

Collect images in Red, Green, Blue (RGB) format with semantic and instance segmentation masks, and bounding boxes

Data prep (YOLO format)

Convert data from raw format to YOLO format for training

Data split

Split data into uniform class distribution

Model training

Train the model on synthetic data for N epochs

Fine-tuning

Fine-tune the model on real images for M epochs for better convergence

Evaluation of model performance

Evaluate the model on real images, testing for mAP, precision, and recall

Contact

Dr. Teymur Sadikhov Kagan
Specialist Leader
Deloitte Consulting LLP
tkagan@deloitte.com

Endnotes

[1] Yiwen Lu and Cade Metz, "Cruise's driverless taxi service in San Francisco is suspended," *New York Times*, October 24, 2023.

[2] Leinar Ramos, *Maverick Research: Forget about your real data — synthetic data is the future of AI*, 2021.

[3] Adam Grzywaczewski, "Training AI for self-driving vehicles: The challenge of scale," NVIDIA Developer Blog, October 9, 2017.

[4] Deloitte research.

This publication contains general information and predictions only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional adviser. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

About Deloitte
Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States, and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.

8928158