SAIGE-
RESEARCH

SAIGE-
VISION

White Paper

Ver 1.2
October 2021

SAIGE-VISION

Introducing the

SaigeVision®

AI Vision Inspection Platform

Automate your vision inspection tasks with the powerful, reliable, and easy-to-use SaigeVision® vision inspection platform. Employing the state-of-the-art in machine learning, *SaigeVision*®offers unmatched accuracy and speed for the most technically challenging inspection problems, even when only limited training datasets are available. *SaigeVision*® is designed for scalability and extensibility, with powerful customized modules can be easily integrated and rapidly deployed for your specific needs.

The Promise of AI

With recent advances in machine learning and AI, machine vision inspection is on the cusp of a revolution. In conventional “rule-based” vision inspection, an expert decides which features—edges, corners, contours, patches, etc.—are relevant for inspection, and then creates a rule-based set of inspection criteria based on these handcrafted features.

The shortcomings of the rule-based approach are well-documented: hours of tedious manual tuning of thresholds and parameters are required; designing features that capture qualitative or subtle criteria is difficult; programming and modifying any of the rules can be notoriously complex. Despite these inherent difficulties, rule-based methods are reliable for basic industrial inspection tasks, and mostly for a lack of better options, they have endured.

With advances in machine learning, especially deep learning networks, the current vision inspection landscape is about to undergo a transformational paradigm shift. Instead of relying on a human expert to design features and create rules, the user instead collects a variety of image samples, and trains a deep learning network. The network can then be deployed for immediate use; little to no intervention by experts is required.

Of course, the training data must be sufficiently clean, varied, and clearly labelled; collection and training must also be repeated for new defect types and products. Yet with the wide availability of deep learning software tools, and the relative ease with which deep learning networks can be trained and deployed by even non-experts, many companies are in a race to build and deploy AI-based vision inspection systems. Some have taken advantage of the relatively low entry barrier offered by publicly available deep learning tools and software, and released AI inspection systems that purport to solve any number of inspection problems in a wide range of application domains.

The Reality

As with most claims and promises about the immediate transformational impact of AI, the reality in vision inspection is far more complex.
Real industrial inspection applications pose many difficult challenges, even for AI:

Accuracy

Defects must typically be correctly detected and identified with greater than 99% accuracy—specifically, the rates for both *false positives* (mistakenly labelling an acceptable part as defective) and *false negatives* (mistakenly labelling a defective part as acceptable) must be kept to within 1%.

Speed

In typical production settings, high-resolution images must be inspected within 10~20 *msec*, often without relying on advanced (and expensive) computational engines.

Flexibility

Inspection systems must be deployed rapidly to the production floor, with on-the-fly training for new defect types and products that may be suddenly introduced.

Scalability

Integration of the inspection system into the existing infrastructure should be seamless. Extension and expansion across different production facilities should also be straightforward, with built-in modularity and scalability.

Given such stringent real-world requirements, making AI vision inspection work in real settings is far from trivial.Below we detail the technical challenges underlying the above performance requirements, and how *SaigeVision*® has managed to overcome these challenges.

I. Achieving Accuracy

The Challenges

The Obstacles to Improving Accuracy Numerous obstacles, some unique to the general industrial inspection sector, stand in the way of achieving the needed accuracy for real-world deployment of AI vision inspection systems:

False Negatives *vs* False Positives: Both Need to be Minimized

The consequences of failing to spot defects, called *false negatives,* can be catastrophic—imagine, for example, the potential damage that can be caused by a faulty battery. Some inspection systems simply impose more stringent thresholds in the hope that it will reduce false negatives, but doing so just ends up leading to more *false positives*—mistakenly labelling an acceptable part or product as defective. Accuracy in inspection means minimizing both false negatives and false positives, not trading off one for the other.

A Deep Learning NetworkPerforms Only as Well as Its Training Data

Deep learning networks, while powerful, are not particularly sophisticated. To work well, they require plenty of training data; this means sufficient quantities of both regular and defective data, cleaned-up, correctly labelled, and in balanced amounts. A network trained with insufficient, imbalanced, or poor-quality training data will most likely perform poorly, with high rates of both false negatives and false positives. A deep learning network will also behave unpredictably to data that it has not encountered before. While the choice of network architecture and training algorithm can influence performance, without doubt the most important factor affecting a deep learning network's accuracy is the quantity and quality of the training data.

Insufficient and Inaccurate Data: Inspection isa Small-Data Problem

Given that most industries are obsessed with eliminating production defects, it should come as no surprise that data for defective parts and products is seldom plentiful. Any data used to train a deep learning network must also be "cleaned" and labelled by human experts. Not only is it costly to assign humans to perform such mundane tasks, but assessments of the same sample can vary considerably even among human experts.

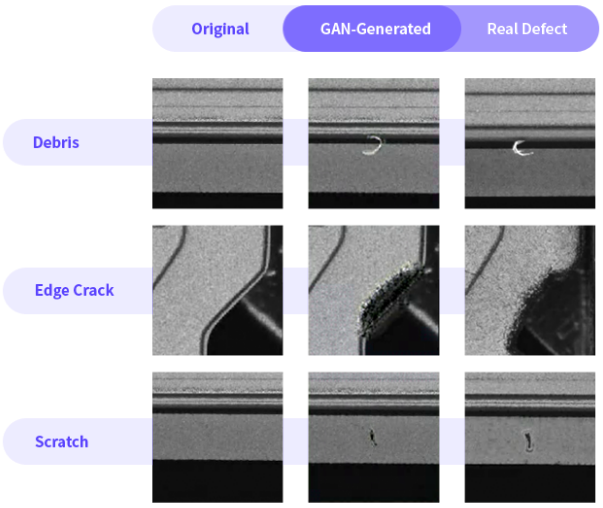
Our Solution

SOLUTION I

Synthetic Data Generation Technology

The synthetic data generation technology of*SaigeVision*®draws upon the latest advances in machine learning to allow users to generate realistic and meaningful synthetic data for training purposes. Generating useful synthetic data is trickier than it seems: simply blurring, scaling, rotating, distorting or augmenting existing image data in some form, or overlaying defects over existing image data, will not measurably improve detection accuracy, especially for more challenging inspection tasks.

SaigeVision® has developed proprietary synthetic data generation technology that overcomes the limitations of existing methods, and produces realistic and robust synthetic data that significantly improves accuracy. Our custom proprietary algorithms merge advanced *generative adversarial network* (GAN) and autoencoder techniques with domain-specific pre- and post-processing methods.



GAN methods, while powerful, are notoriously difficult to use in practical applications: improperly designed GANs fail to generate new meaningful data, and training a GAN can be time-consuming, and highly sensitive to mode collapse and various other instabilities. *SaigeVision*® uses in-house algorithms optimized for industrial inspection that overcomes these and other challenges, delivering state-of-the-art performance for real application scenarios.

The system allows the user to manually create highly customized defect samples, or to automatically create an array of defect types, generating randomized defect samples for training.

> "Using the synthetic data generation capabilities of SaigeVision® in our electric vehicle (EV) battery production lines, we were able to achieve a 77% reduction in false negatives. Just as importantly, the synthetic data generation allowed us to perform inspection immediately on our new product lines, for which defect data was not available. We are extremely satisfied with the overall performance of SaigeVision® and its powerful and robust capabilities, and are continuing to deploy SaigeVision® throughout our wider production facilities."

Testimonial from Samsung SDI EV Production Inspection Manager (2020)

Solution II

Auto-Labelling Technology

As the name implies, auto-labelling refers to a collection of machine learning techniques for automatically labelling data. The most basic collection of methods are *one-class learning* (or *manifold learning*) algorithms, which attempt to find a lower-dimensional representation of the data. As a useful visual analogy, imagine a point cloud in some higher-dimensional space, with each point corresponding to observed data: the underlying assumption—the *manifold hypothesis*—is that most of the data belong to some lower-dimensional surface (the dominant class, or *manifold*) embedded in this higher-dimensional space. By characterizing this surface, one can more meaningfully measure how distant some point is from the surface.

Since most of the data collected in typical industrial inspection settings will be of non-defective parts and products, one-class learning can identify those images that are non-defective (*i.e.*, they lie on the surface), separating these from the much smaller number of images that are likely to contain defects (*i.e.*, those points whose distance from the surface exceeds some threshold). Human experts then only need to examine and label this much smaller set of possibly defective image data.

There are a wide range of one-class learning algorithms, with new promising ones being proposed continuously, and their performance highly dependent on the nature of the task and data. Some of the latest one-class learning methods for image data are able to accurately detect and identify the location of defects from a large collection of unlabeled data.

The caveat is that these methods, at least in their original form, cannot meet the accuracy requirements of industrial inspection. At Saige Research we have developed proprietary one-class methods that meet these stringent requirements: our algorithms are optimized for the domain of industrial image inspection, and have been tested and validated to be robust and reliable for a wide range of industrial image inspection tasks.

[**Note:** Auto-labelling is currently available only for customized applications, and will be included as a regular feature in the upcoming 2021 third-quarter release of *SaigeVision v2.1*—Other more advanced auto-labelling features will be released in the near future.]

II. Achieving Speed

The Challenges

The Obstacles to Achieving Speed

Deep learning networks require lots of computation, both during inspection as well as in network training; this is a primary reason why GPUs are essential to implementing practical deep learning systems. Applications like semiconductor wafer inspection require fast inspection times on the order of 5-10 *msec* per image. Training also needs to be done rapidly, and not take hours or days.

With such demanding speed requirements, it's not hard to see why general-purpose deep learning packages don't do the job. What are the underlying sources of the intensive computation, and what can be done to improve speeds?

First, the networks produced by many general-purpose tools are often unnecessarily large and redundant. The greater the number of nodes (where computations are performed) and connections (where data is passed), the slower the network and greater the memory requirements. Choosing a needlessly large network, or training it poorly, often results in a network with many idle nodes and connections. Constructing the smallest deep learning network that meets accuracy and other performance requirements is essential to maximizing a network's speed.

Training deep learning networks is especially computation-intensive; it is not unheard for training to take days for large training data sets. Naturally smaller networks, although generally less expressive than larger networks, can be trained more rapidly. Pre-processing and curation of the training data, for example by removing samples that are redundant or devoid of useful information content, can further reduce network training times.

The training algorithms and code behind general-purpose deep learning networks are not particularly efficient, because they are not optimized for the task at hand. Code optimization and better memory management alone can significantly improve network inference and training speeds. With a better understanding of the underlying problem structure and by leveraging efficient state-of-the-art numerical algorithms, speed can be further improved in myriad ways.

Our Solution

SOLUTION I

Optimal Transfer Learning and Network Compression

The idea behind *transfer learning*—taking a pre-existing network that has already been trained with a general data set like *ImageNet*, and resetting and retraining some of the node weights with more

task-specific data—is well-known. Transfer learning is a popular and widely used technique in many real-life applications. Choosing the network architecture and training data set, and deciding which of the nodes to reset and retrain, often involves many *ad hoc* decisions involving considerable trial-and-error and manual fine-tuning. Done incorrectly, the final network will yield only minimal improvements in training efficiency, or worse, lead to degraded performance.

Going hand-in-hand with transfer learning is *pruning,* which refers to a collection of methods for reducing the size of a network. Methods for pruning involve, variously, techniques for ranking the importance of nodes and subnetworks and systematically pruning the least important nodes, and further fine-tuning of the reduced network. Like transfer learning, methods for pruning involve many *ad hoc* choices and manual tuning and, done wrong, can actually worsen a network's performance.

SaigeVision® relies on transfer learning algorithms that have been optimized for the inspection domain, leading to highly sparse networks that are amenable to pruning. The result is a lightweight network with considerably reduced training and inference times, all while meeting the stringent accuracy and performance requirements of the inspection task at hand.

SOLUTION II

Focused Attention Learning

The interesting and most information-rich parts of an image usually occupy only a small part. Humans have the intuitive ability to focus their attention on these information-rich parts, but typical machine learning training algorithms do not; they process the training image dataset in their raw form, further slowing down training.

The focused attention learning algorithms provided by *SaigeVision*®take advantage of common features in industrial inspection. By pre-processing the images to focus on the most relevant parts, training can be achieved much more effectively with a considerably reduced number of images

SOLUTION III

Code and Algorithm Optimization

Most general-purpose machine learning tools and packages employ algorithms that do not exploit any special features of the problem. With its focus on industrial image inspection, *SaigeVision*® employs the most efficient numerical algorithms that are optimized for the task domain. Our code implementation has also been optimized to squeeze maximum performance out of the available computing and memory resources.

III. Achieving Flexibility and Scalability

The Challenges

Obstacles to Achieving Flexible and Scalable Inspection Systems

Finding the sweet spot that balances flexibility with performance is always an ongoing challenge. Inspection criteria are adjusted, products are regularly upgraded or re-designed, new defect types appear, and the inspection environment—cameras, lighting, hardware—is continuously subject to changes and disturbances large and small. Moreover, as seasoned engineers know, the transition from proof-of-concept to production line deployment is significant, as is the transition from a single production line to across multiple production lines and facilities—the required additional work can be 3-5 times that of the proof-of-concept stage.

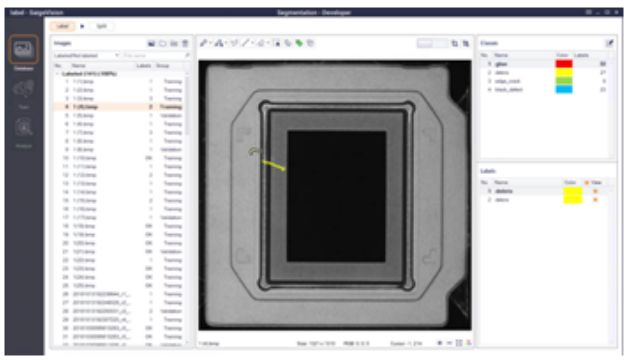
Our Solution

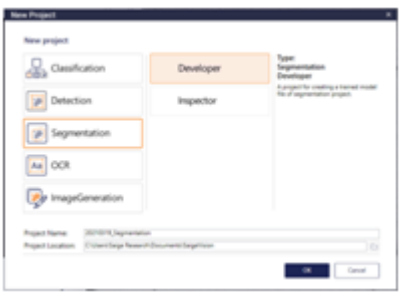
SaigeVision® has been designed to exploit as much as possible the common features of industrial image inspection to maximize performance, while keeping the overall platform flexible and scalable. The interface is intuitive and easy to learn, allowing users to rapidly upload and process data, to efficiently train a network for inspection, and to deploy the network to production lines. Its modular software architecture and components enable easy customization for specific customer needs and seamless integration with existing inspection systems and infrastructure.

SOLUTION I

A Light, Intuitive, and Responsive Interface

The *SaigeVision*®interface is designed to be light, intuitive, and easy to learn and use. Unlike other vision inspection systems that require multiple parameters to be initialized and manually adjusted by the users, and a myriad of seldom-used buttons and menus, the*SaigeVision*®interface has been carefully designed to be intuitive and clutter-free. Uploading and editing data, network training, and inspection is straightforward and fast—no long delays waiting for the system to respond.





SOLUTION II

Easy Integration with Existing Systems and Infrastructure

Inspection often involves multiple systems—including legacy systems—serving different functions that must all be seamlessly integrated together. Rule-based inspection systems, for example, may already be in place at a production line, and migration to a purely AI-based system may take a long time. With its modular, light, and flexible architecture, *SaigeVision*®can be easily integrated with existing systems: data transfer and communications for, e.g., cloud applications, monitoring, and reporting can be done with minimal programming and overhead.

SOLUTION III

Easily Deployable and Scalable

With its lightweight and modular design, transitioning from proof-of-concept to production line deployment is made easier with *SaigeVision*®. A trained network deployed at one production line can be easily modified and ported to another production line, and information collected and shared across different networks.

www.saige.ai

+82-2-877-0566
contact@saigeresearch.ai

Saige Research 5F, 49, Seocho-daero
40-gil, Seocho-gu, Seoul, 06656,
Republic of Korea

SAIGE-
RESEARCH