Safeguards-Informed Hybrid Imagery Dataset

Project Number SL21-Hybrid Image Datasets-PD3SZ
Authors: Joshua Rutkowski, Zoe Gastelum, Tim Shead, Ahmad Rushdi, Jason Bolles and Arielle Mattes

National Nuclear Security Administration

The Challenge

Deep Learning computer vision models require many thousands of properly labelled images for training, which is especially challenging for safeguards and nonproliferation, given that …

… safeguards-relevant images are typically rare due to the sensitivity and limited availability of the technologies.

… creating relevant images through real-world staging is costly and limiting in scope.

… expert-labeling is expensive, time consuming, and error prone.

Goal

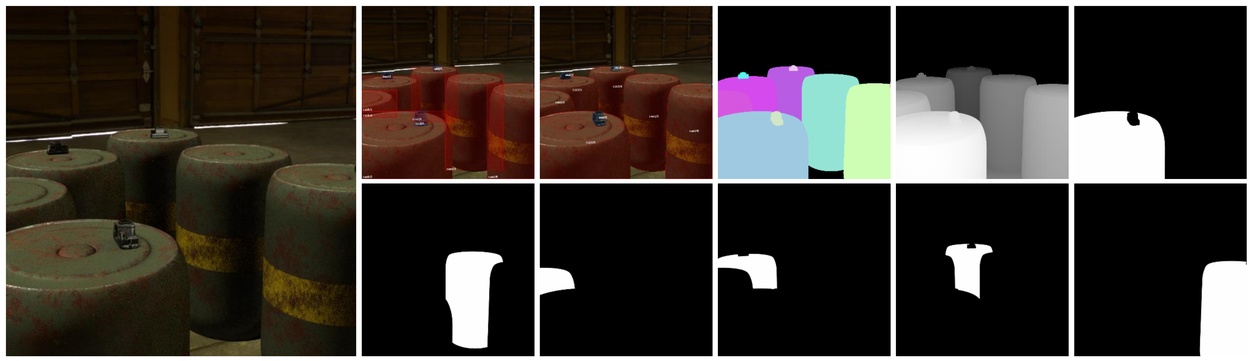
We aim to develop a data set of both real-world and synthetic images that are relevant to the nuclear safeguards domain that can be used to support multiple data science research questions. In the process of developing this data, we aim to develop a novel workflow to validate synthetic images using machine learning explainability methods, testing among multiple computer vision algorithms, and iterative synthetic data rendering.

We will deliver one million images – both real-world and synthetically rendered – of two types uranium storage and transportation containers with labelled ground truth and associated adversarial examples.

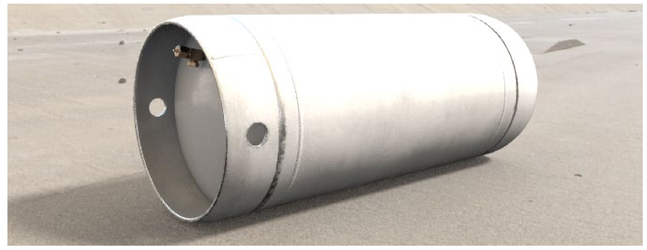
Method: Synthetic Images

Synthetic images can be created using a combination of real-world reference images, parameterized 3D CAD models, and random sampling, rendering the results as 2D images for training.

Because the process is generative, perfect ground truth labels can be created as an integral part of the rendering process. In the following sample synthetic scene, the automatically generated ground truth includes bounding boxes, contours, per-class-instance bit masks and more.



Synthetic data can contain as much variance as time, creativity, and physical constraints permit. **Below, parameterized aging effects have been applied to a synthetic object:**



The synthetic 30B UF_6 cylinder in the above image can be displayed with a variety of paint jobs, aging, environments, lighting conditions, and camera angles to create a wide variety of training exemplars:

Deep Learning for Object Detection

We use a standard Deep Learning workflow for object detection based on TensorFlow and Keras. Our baseline models aim to localize objects in an image. We experimented with Coco and VOC as datasets and Yolo/Darknet/Detectron as detectors. We have had initial success using a Keras-YOLOv3 workflow.

Validation

We use quantifiable quality metrics when training a deep learning model on **synthetic** images and testing on **real** ones. For example, for a synthetic image generator G, the Train-Synthetic-Test-Real (TSTR) factor is:

$$TSTR(G) = \frac{1}{k}\sum_{i=1}^{k} m\left(\mathcal{M}_i(\mathcal{D}_{tr}^{G}), \mathcal{D}_{ts}\right)$$

averaged over k models, where $m(.,.)$ is a distance meteric between the i^{th} model trained on synthetic data $\mathcal{M}_i(\mathcal{D}_{tr}^{G})$ and the real test data $\mathcal{D}_{ts}$.

Findings So Far

Open sources for actual images of Type 30 and 48 containers for this work is limited to public domain and commons-licensed venues – which leads to a significant number of duplicates but an almost insignificant number of images relative to the target quantity for our data set.

Commercial and editorial sources show promise for a marginally larger subset, but legal concerns preclude their use.

Relevance

Our data will be available for machine learning researchers creating computer vision models. Access to a large collection of safeguards-relevant data could spur new innovations and allow apples-to-apples comparisons among domain relevant models.

ENTER THE ERCA NUMBER HERE

Sandia National Laboratories

CONTACT
Zoe Gastelum
505-845-1002
zgastel@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

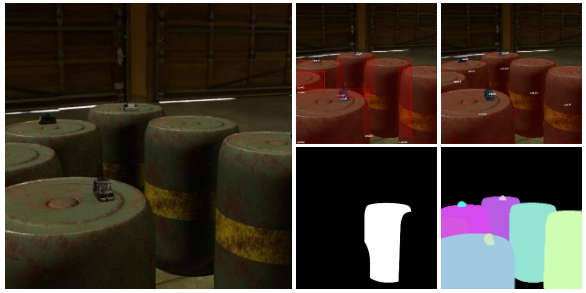
U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration
Defense Nuclear Nonproliferation

Machine Learning-Validated Hybrid Image Dataset for Object Detection and Classification

- Deliver **one million images** of uranium transportation and storage containers, which are relevant to international safeguards
- Synthetic images created using a combo of real-world reference images, parameterized 3D CAD models, and random sampling, rendering the results as 2D images for training
- **Our data will be available for machine learning researchers creating computer vision models**

Parameterized aging effects applied to a synthetic object.



Perfect ground truth labels can be created as an integral part of the rendering process

Poster #8A
Joshua Rutkowski
Sandia National Laboratories