The Big Data Myth: Using Diffusion Models for Dataset Generation to Train Deep Models

Roy Voetman, Master Computer Vision & Data Science
Supervisors: Maya Aghaei Gavari, Klaas Dijkstra

Summer 2023

Figure 1. By using 20 images of a real-world scenario (left) and leveraging recent advancements in image synthesis, we demonstrate the capability to generate images (right) used to train deep detection models.

Introduction

- Over the past decade, deep learning has revolutionised the field of computer vision.
- Acquiring representative datasets for training deep models is challenging in data scarce scenarios.
- This research:
  - *Genfusion*, a framework that combines recent advances in image synthesis to generate images to train an object detector.
  - Feasibility is demonstrated within the context of apple detection because well-established benchmark datasets are present.
  - Concretely, using 20 real-world images, a pretrained text-to-image diffusion model is fine-tuned to produce images that are highly similar to the real-world scenario.

Abstract

This study explores the challenge of acquiring large amounts of training data for deep object detection models. We propose a framework for generating synthetic datasets by fine-tuning pretrained stable diffusion models. These datasets are manually annotated and used to train object detection models, which are then compared to a baseline model trained on real-world images. Results show that models trained on synthetic data perform similarly to the baseline, highlighting the potential of synthetic data generation as an alternative to extensive data collection for deep model training.

Experiments and Results

- Train YOLOv8 object detectors to detect apples in apple trees.
- Compare training on generated with training on real-world data.

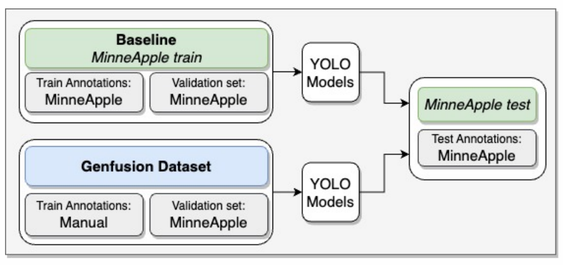

Figure 4. Experiment Setup

Table 1. AP evaluation metrics on the MinneApple test set between different YOLOv8 models trained on the generated train set and the model trained on the MinneApple train set.

Dataset	AP@0.5:0.05:0.95	AP@0.50	AP@0.75
Baseline yolov8x	0.45 ±0.005	0.79 ±0.005	0.47 ±0.011
Genfusion yolov8x	0.33 ±0.050	0.66 ±0.060	0.29 ±0.070
	0.12 –	**0.13** –	**0.18** –
Baseline yolov8l	0.45 ±0.008	0.78 ±0.009	0.47 ±0.008
Genfusion yolov8l	0.35 ±0.019	0.70 ±0.011	0.32 ±0.031
	0.10 –	**0.08** –	**0.15** –

(a) MinneApple Baseline (YOLOv8l)

(b) Genfusion (YOLOv8l)

Figure 5. The predicted (white) and ground truth (green) bounding boxes of the YOLOv8l Baseline and Genfusion model from three of the images from the MinneApple test set. The blue arrows indicate instances where the Genfusion trained model predicts two bounding boxes for a single apple.

Materials and Methods

- Benchmark dataset: MinneApple
  - Dataset for apple detection in apple orchards. 536 image train set with different kinds of apples at a variety of growth stages.
- Generated dataset: Genfusion
  - Goal: preserve all the characteristics of the benchmark's training set, including:
    - The number of images.
    - The distribution of apple colours.
    - Only annotating images in foreground.
      - Annotation process was manual
- Generative model: Stable Diffusion
  - Open Source pretrained Text-To-Image model based on diffusion models.
- Fine-tuning: DreamBooth
  - This enables the generation of images that closely resemble the real world.
  - DreamBooth is a technique for fine-tuning diffusion models addressing language drift and overfitting by regularising the model on its own generated images before fine-tuning.

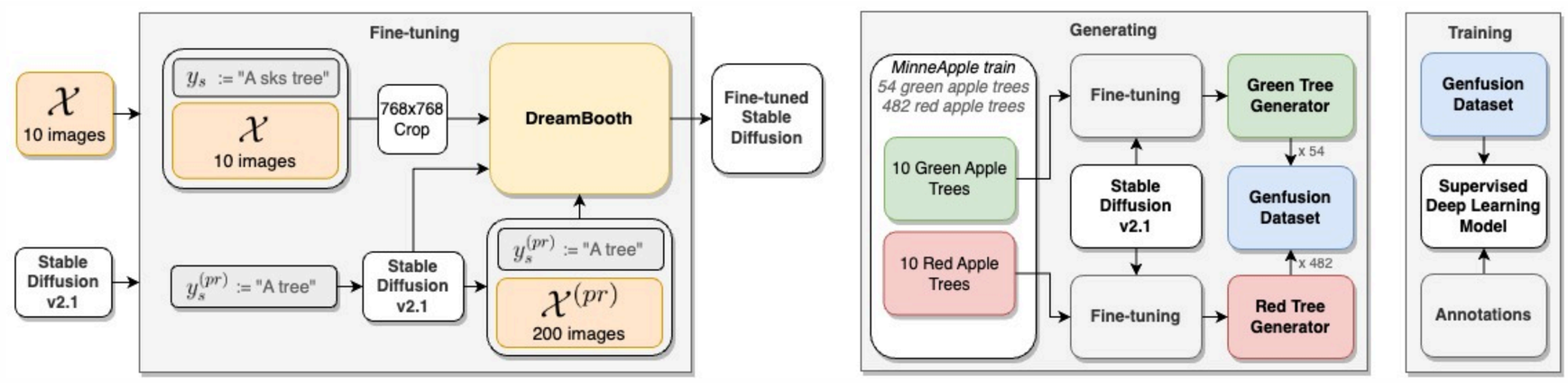

Figure 2. The Genfusion Pipeline *(1) Fine-tuning:* Starting with 10 real-world images and a pretrained Stable Diffusion model. DreamBooth fine-tuning is then performed on the model using 200 prior preservation images. *(2) Generating:* Fine-tuning is applied separately for green apple trees and red apple trees to control the colour distribution in the generated dataset. *(3) Training:* We use the generated data and annotate it with bounding box information to train a deep model.

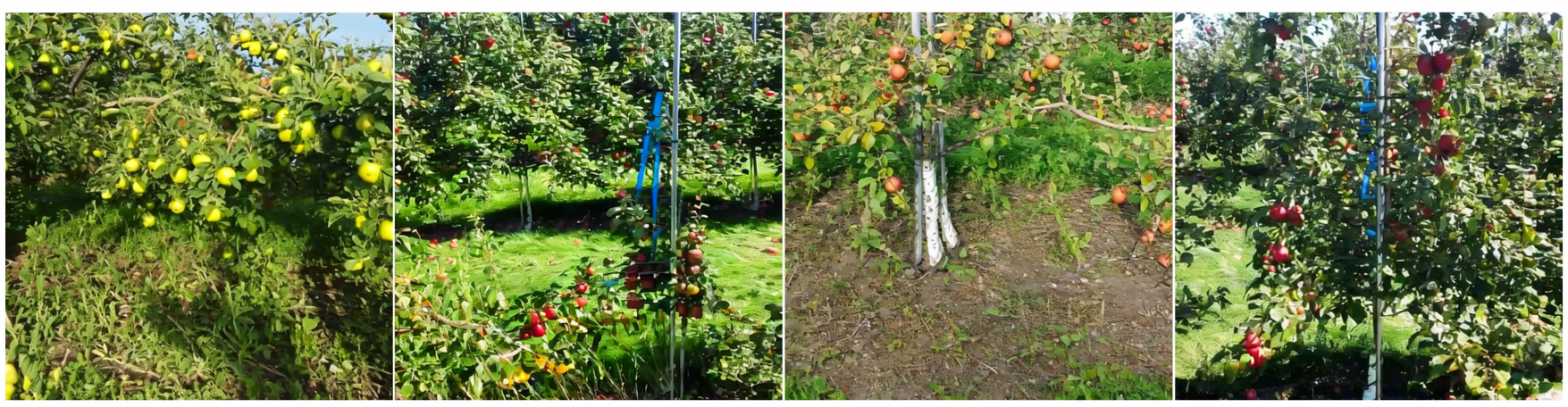

Figure 3. A subset of the images in the Genfusion dataset which is composed of instances generated by the green apple tree and the red apple tree generators.

NHL STENDEN
computer vision & data science

Acknowledgements

- This project is financially supported by Regieorgaan SIA (part of NWO) and performed within the RAAK PRO project Mars4Earth.
- We would like to thank our collaborators at Saxion University of Applied Sciences for insightful discussions.

Regieorgaan SIA
SAXION UNIVERSITY OF APPLIED SCIENCES

Conclusions

- The proposed framework shows promising results with close to real-world performance when compared to a well-established benchmark dataset.
- Feasibility of the approach demonstrated for apple detection, but potential for extension to other (data-scares) domains.
- Overall, our study demonstrates the versatility and potential of generative approaches in object detection.