Bridging the Synthetic to Real Data Gap

YRIKKA

BREAK THROUGH TECH

August 2025

We’re excited to be your Challenge Advisors!

Kia Khezeli
YRIKKA
Co-Founder & CEO
kia@yrikka.com

John Kalantari
YRIKKA
Co-Founder & CTO
john@yrikka.com

Maxim Clouser
YRIKKA
Founding ML Engineer
max@yrikka.com

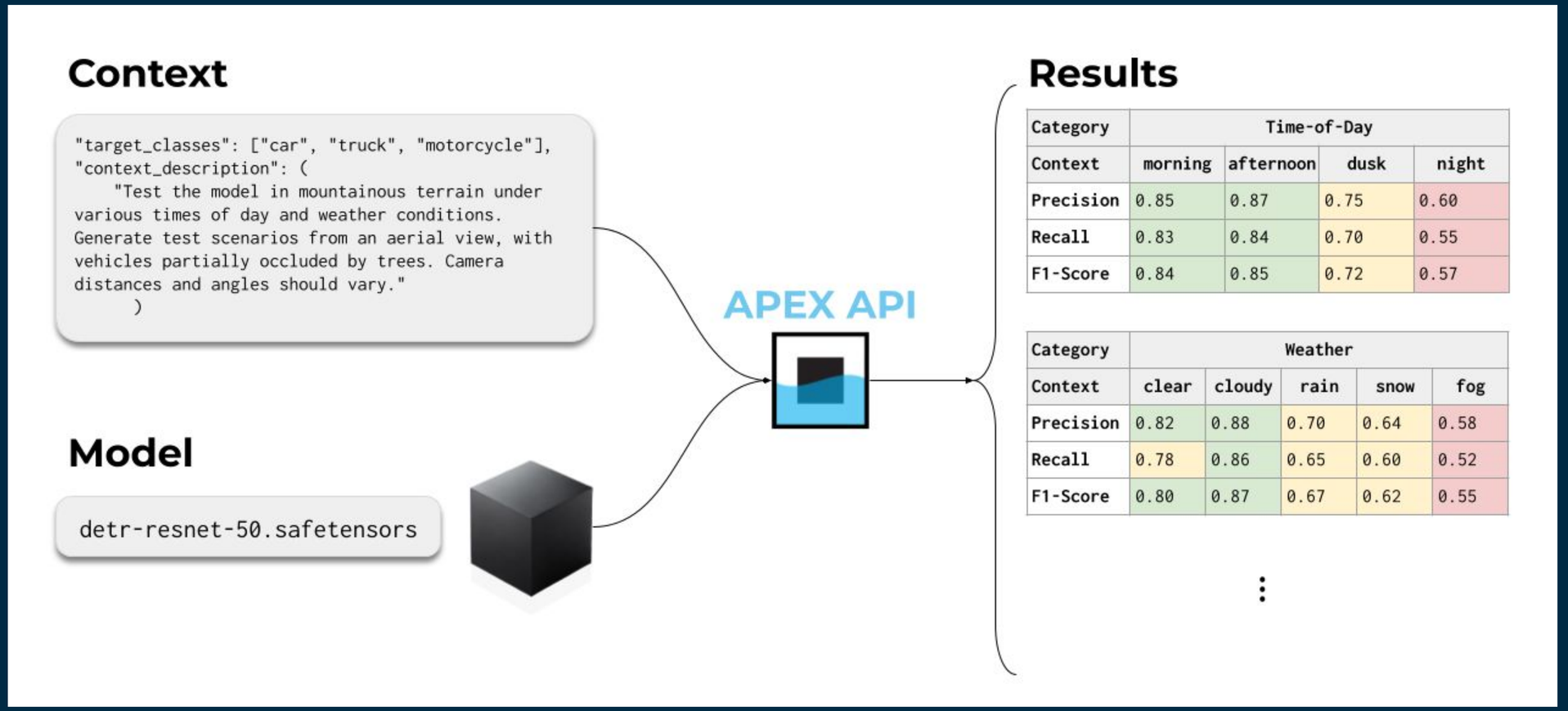
Company overview

Perception is the most failure-prone layer in autonomous systems, especially in dynamic, open, and unpredictable environments. Our mission is to build the trust layer for physical AI by hardening the perception stack and making it adaptive to real-world complexity.

- **We’re an AI start-up founded in 2022 based in NYC**
- **Our platform generates hyper-realistic, multi-sensor synthetic data with pixel-aligned EO/IR pairs and full-fidelity annotations.**
- **Our platform detects the vulnerabilities of computer vision models on edge cases and help address them by fine-tuning on synthetic data.**

Website, LinkedIn, Documentation

More about YRIKKA APEX API



Category	Time-of-Day			
Context	morning	afternoon	dusk	night
Precision	0.85	0.87	0.75	0.60
Recall	0.83	0.84	0.70	0.55
F1-Score	0.84	0.85	0.72	0.57

Category	Weather				
Context	clear	cloudy	rain	snow	fog
Precision	0.82	0.88	0.70	0.64	0.58
Recall	0.78	0.86	0.65	0.60	0.52
F1-Score	0.80	0.87	0.67	0.62	0.55

⋮

AI Studio Challenge Project Overview

CHALLENGE SUMMARY

Training downstream computer vision models on synthetic data may result in poor performance in the real world, a phenomenon known as the synthetic-to-real (or sim2real) data gap.
Project goal: Evaluate whether augmenting a YOLO model with synthetic data can improve object-detection performance.

YOUR TEAM'S OBJECTIVE

The overall objective of the project is to improve the performance of a YOLO model in detecting five everyday objects [*"potted plant"*, *"chair"*, *"cup"*, *"vase"*, *"book"*].

1. Evaluate the performance of the YOLO model on synthetic data with imperfect annotations. 10% of the annotations will be either missing or incorrect. Use the YOLO model predictions to improve the annotations using CVAT.
2. Assemble 200 challenging real-world test images (e.g., with low lighting or partial occlusion) and annotate them in CVAT. To accomplish this, each team member should aim to collect 10-15 test images from the internet or with their camera.
3. Generate synthetic images via YRIKKA's data engine, and then fine-tune the YOLO model. The goal is to improve the MAP@50 metric by 0.10 compared to the baseline model.

DESIRED OUTCOMES

We'd like to see students present the improvement from fine-tuning a model on synthetic data to demonstrate its utility.

Business context

Perception is the most failure-prone layer in autonomous systems, especially in dynamic, open, and unpredictable environments.

- We’re enhancing the perception robustness of autonomous systems (e.g., drones)
- Collecting real data for all possible scenarios and edge cases is difficult. Synthetic data has the potential to address this problem. The goal of the project is to demonstrate it on a general purpose use-case.
- The planning and reasoning capabilities of state-of-the-art multi-modal LLMs has created an opportunity for autonomous systems including robots, drones, and vehicles to operate in the real world. Successful completion of this project, will help students learn the necessary skills to develop cutting edge computer vision models that can be deployed in the real world.

Suggested ML approach

- This project involves a combination of supervised learning and active learning approaches.
- The YOLO family of models are beginner-friendly yet adaptable to real-world deployment.
- Mean Average Precision at 50 (mAP@50) measures the accuracy of object detection models by evaluating how well predicted bounding boxes match ground truth boxes, using an Intersection over Union (IoU) threshold of 0.5 to determine a correct detection. It is a good metric because it balances precision and recall while requiring only a moderate IoU, making it useful for assessing whether models can generally localize objects without requiring perfect alignment.

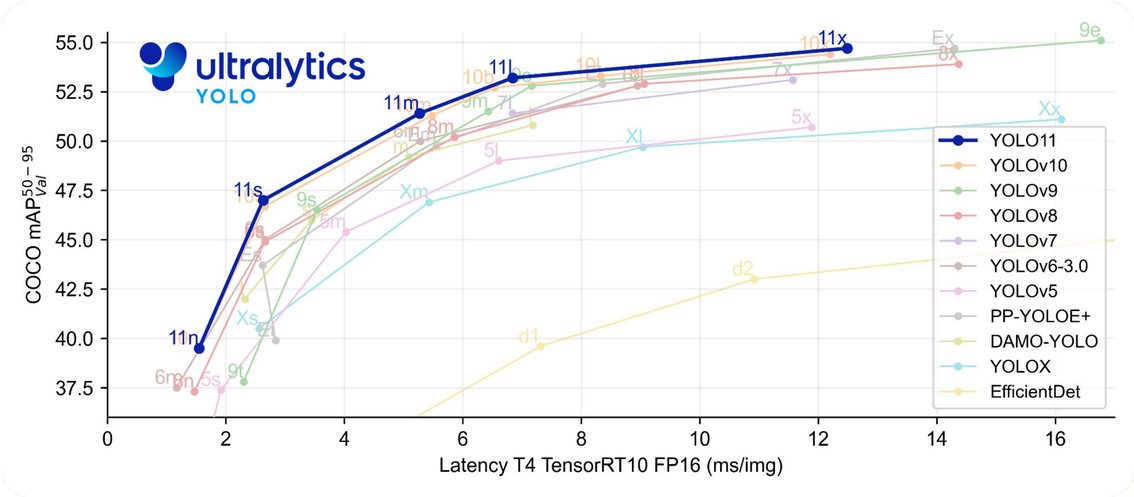
Data overview

- You will be provided a synthetic data set that includes images of everyday objects [“potted plant”, “chair”, “cup”, “vase”, “book”] along with their bounding box annotations. The data set size is 2000 and is provided in the YOLO format.
- The bounding box annotations are imperfect. There may be cases where the annotations are missing or incorrect.
- Students will need to use the YOLO model to detect potential incorrect annotations and correct them using the CVAT platform (an open source platform for data annotations)
- The data set is hosted on Github::

  https://github.com/YRIKKA/yrikka-btt-aistudio-2025

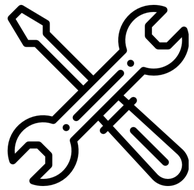
Python libraries

- Main Python libraries that are useful for this project: scikit-learn, PyTorch
- Ultralytics library includes streamline methods for training and testing YOLO models.
- CVAT is an open source platform that facilitates the annotation of images.



Suggest tools and workspaces

- Recommended free tools for:
  - Data Sharing and Collaboration: GitHub, Google Drive
  - Coding & IDE: VS Code, PyCharm
  - Project Management (e.g., GitHub Projects, Notion, Linear)
  - Other (e.g., Google Colab and Kaggle notebooks for free GPU resources)
- Creating a GitHub repository for the team is highly recommended to ensure version control across different team members.



Helpful resources

- Tutorial on mAP as a metric for object detection: https://blog.roboflow.com/mean-average-precision/
- Ultralytics documentation for training, fine-tuning, and testing YOLO models: https://docs.ultralytics.com/guides/#guides

Project milestones and timeline

These are the monthly Milestones for your Challenge Project. You will need to complete several tasks for each of these.

MILESTONE 1

September

Utilize the YOLO model predictions and the CVAT platform to correct any incorrect annotations.

MILESTONE 2

October

Collect challenging real-world test set images to test the baseline YOLO model against.

MILESTONE 3

November

Fine-tune the YOLO model on synthetic data.

MILESTONE #1 TASKS

Data preprocessing

- **Model Guided Data cleaning** Use the YOLO model to identify images with missing annotations and correct them in CVAT.
- **Data Augmentation** Standard image augmentation techniques such as cropping, rotation, and flipping may be helpful.
- Suggested tools and libraries for this stage of the project (e.g., Matplotlib, OpenCV)

MILESTONE #2 TASKS

Real-World Testing

- Fine-tune the YOLO model on the synthetic data from Task 1 using the corrected annotations.
- Collect 200 challenging test images (e.g., low light conditions, partial occlusion, uncommon shape or positioning).
- Compare the performance of the base and fine-tuned YOLO model on your test data to understand the model failures.

MILESTONE #3 TASKS

Generate Additional Synthetic Data to Fine-Tune the YOLO Model

- Based on the results of Milestone 2, generate an additional 500-2000 synthetic data using the YRIKKA data engine.
- Fine-tune the YOLO model from Milestone 2 on the additional synthetic data.
- Compare the performance of models on your test set and evaluate the utility of synthetic data for improving the performance of object detection models.

How we'll work together this semester

Check-in meetings	• Please prepare a short slide deck summarizing your progress and include questions or action items for the YRIKKA team. Please email the slides 48hrs before the meeting time.
Reporting	• Please provide a weekly progress update on Slack and a bi-weekly report summarizing your progress and potential model metrics.
Communication	• Please use work email for urgent requests and the Break Through Tech Slack workspace for general questions (expect a response within 24 hours)
Tools and platforms	• FREE industry-standard tools or platforms to complete your project work: GitHub, VS Code, Kaggle/Google Colab notebooks, CVAT

How to get started

Here’s what I suggest for your immediate next steps. I’ll follow up on your progress and help address any challenges in our next check-in meeting:

Review these slides and note down questions

I’ll email you a copy of this deck. Review it as a team and note down any questions you’d like to discuss in our next meeting.

Complete your “Project Brief and Workplan”

Continue working on your Project Brief and Workplan assignment, which is due next month. We’ll review it again in our next meeting.

Access to YRIKKA Data Engine

In the last stage of the project, we will provide you access to our data engine to generate synthetic data.

Questions?

What questions do you have?

Anything I can help clarify?

What are you most excited about?

Anything you’re unsure about?