CentraleSupélec | université PARIS-SACLAY

Vision-based teleoperation of robotic hands through end-to-end deep neural network

Supervisor: **Guillaume Thomas**

Student: Jintao Ma

Content

- Introduction
- Reproduction of Anyteleop
- Hand Detection
  - Data-preprocessing
  - Resnet+Faster-RCNN
  - YOLO
- Results & Conclusion

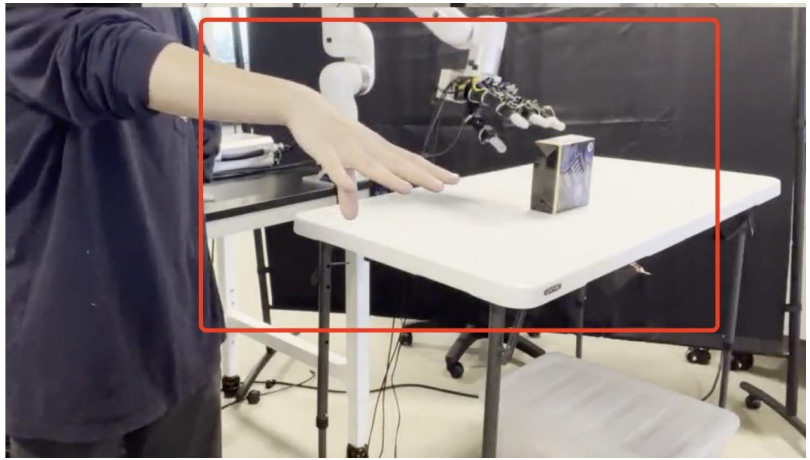
Motivation

Suppose you are working in a dangerous underground coal mine:

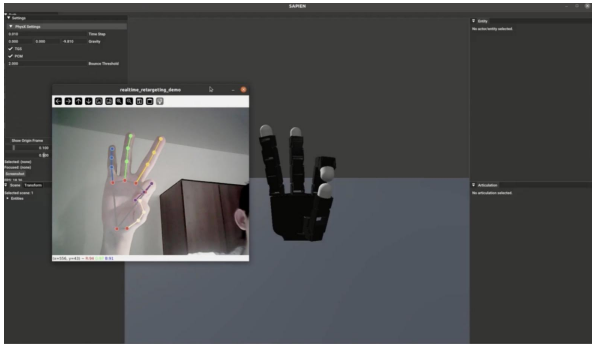
1.Why not do mining remotely without going underground?

2.Can you use robots for repetitive tasks like moving stones without doing it yourself?

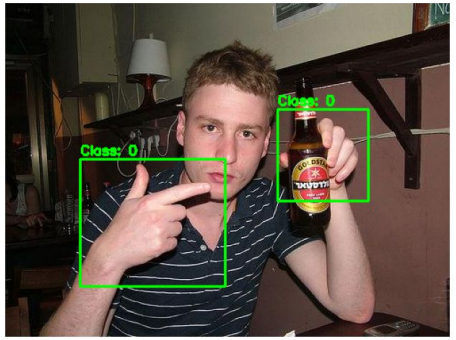
3.How do you detect your (Robots) hands in different environments?



Overview



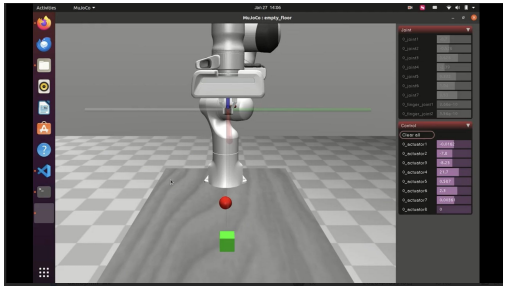
Reproduction of AnyTeleop(Sapien)



Hand Bounding Box Detection

October → November → December → January → February

Usage of Simulation tools (Mujoco)



Faster R-CNN

ultralytics YOLOv8

Some Problems I met

- **Environment Configuration**
- **Lack of GPU(Nvidia)**
- **Hardware Restriction(Camera)**
- **Data Tranformation for Rotated bounding box label**
- **Data Preprocessing for different input format**
- **Data Augmentation for Geometric Transformations**
- **Excessive training time**
- **…**

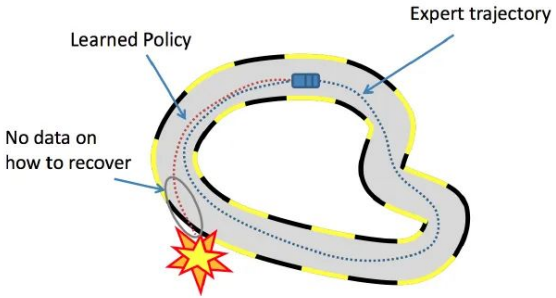
Background

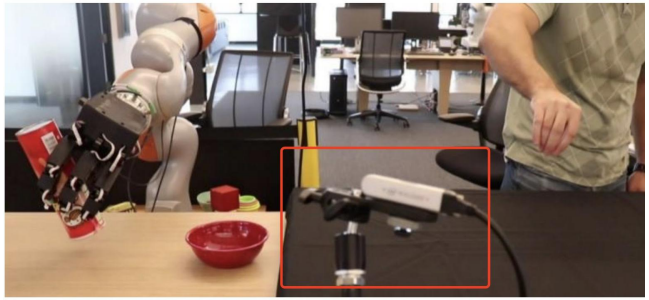
Imitation Learning: Robots hands/arms can **learn tasks** by imitating human behavior**.**

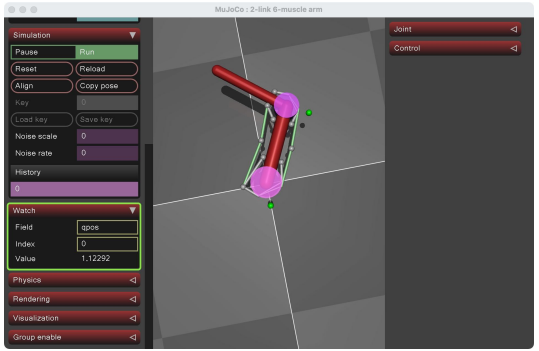
Version-based Teleoperation: Using a camera to **remotely** control robotic arm/hands

Simulation Tools: Simulating the real world and tasks in your computer**(Mujoco,Sapien)**

Dexterous Hand: a highly flexible robotic hand designed for **precise manipulation**, imitating human hand movements(**Allegro,Shadow**)



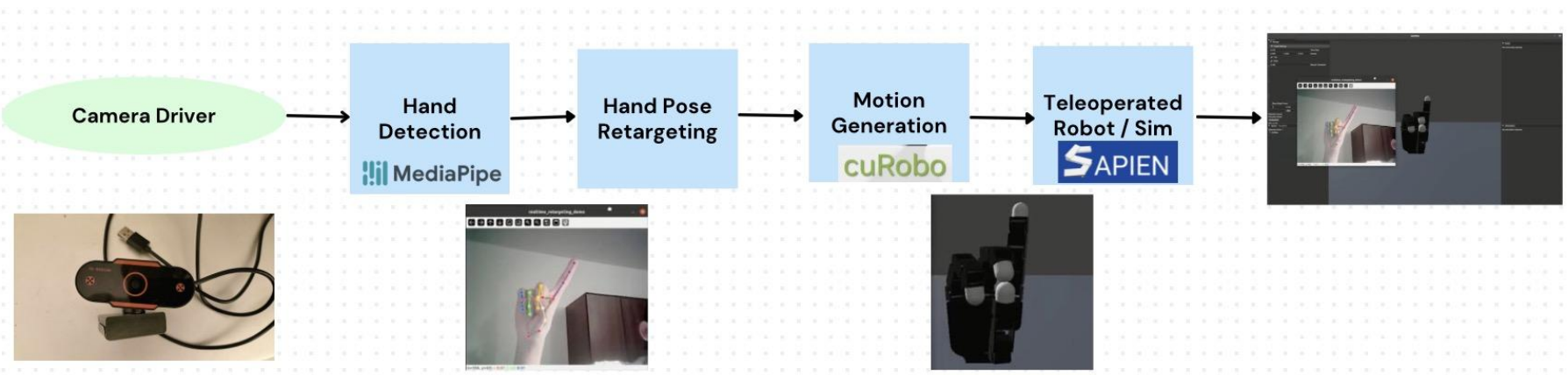




AnyTeleop

A generalized vision-based teleoperation system.

- Support multiple robotic hands and camera configurations.
- Offer flexibility across different simulators and real hardware.
- Maintain high performance in both real-world and simulation tasks.

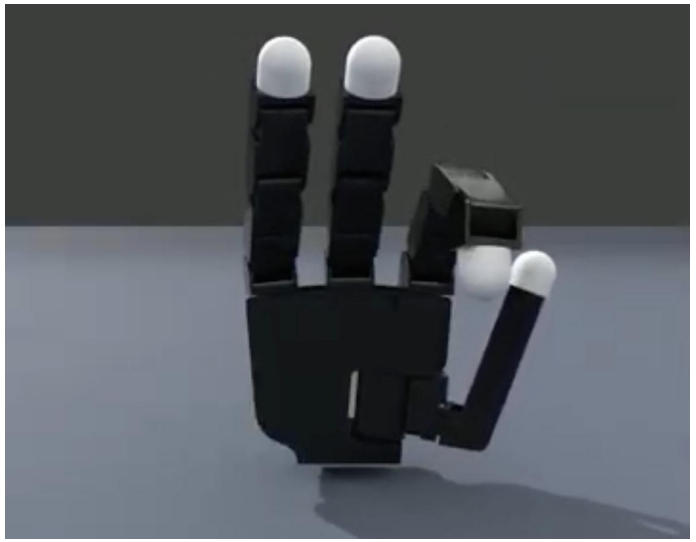


Key Module: Teleoperation Server Framework

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, Dieter Fox. "AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System." Robotics: Science and Systems (RSS), 2023.

Reproduction(Hand Pose Retargeting)

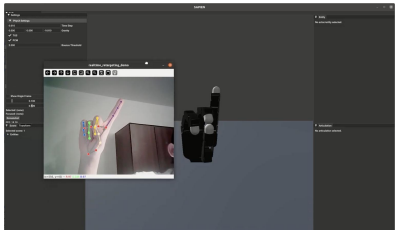
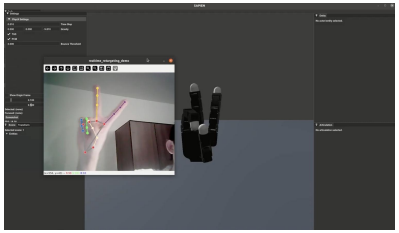
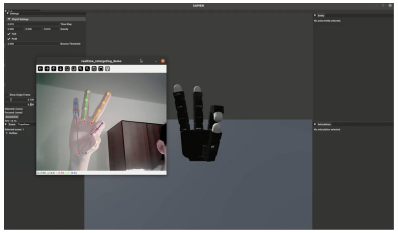
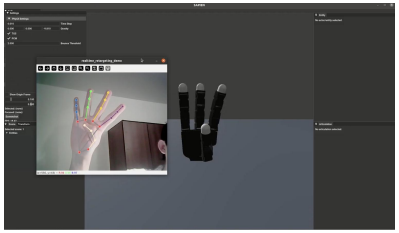
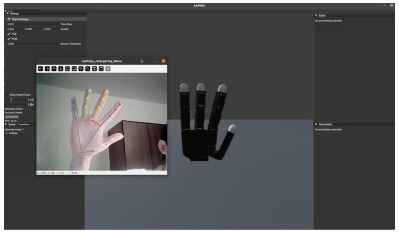
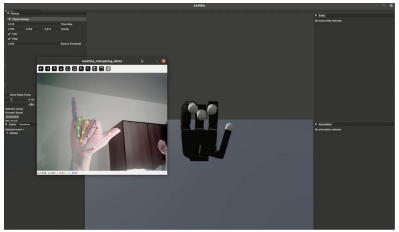
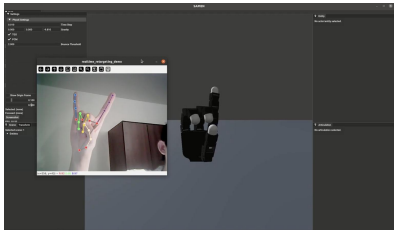
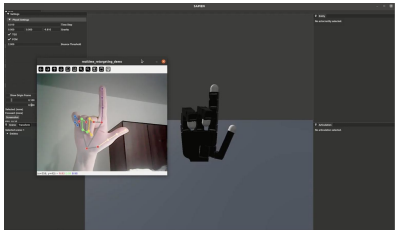
Pre-recorded Video Retargeting(Real hand VS hand on Sapien)



Link:
https://drive.google.com/file/d/1C0_AoNF6GTkLimsCIIYhgZHGCIV3q7Km/view?usp=sharing
https://drive.google.com/file/d/1DE6nNW_nbn92NFw1TIzFMhz5JzwOdwah/view?usp=sharing

Reproduction(Hand Pose Retargeting)

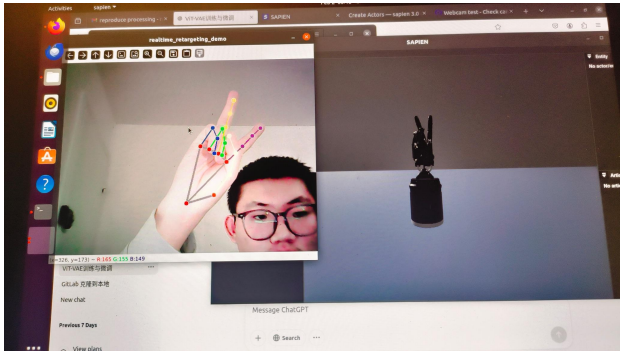
Real-time Mode Retargeting(Real hand VS hand on Sapien)-different postures with Allegro robot hand

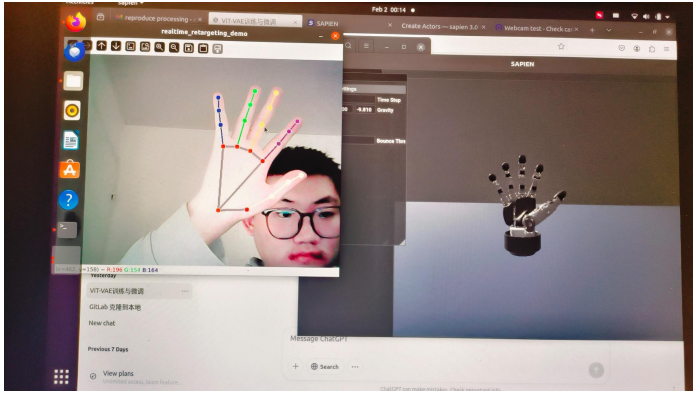









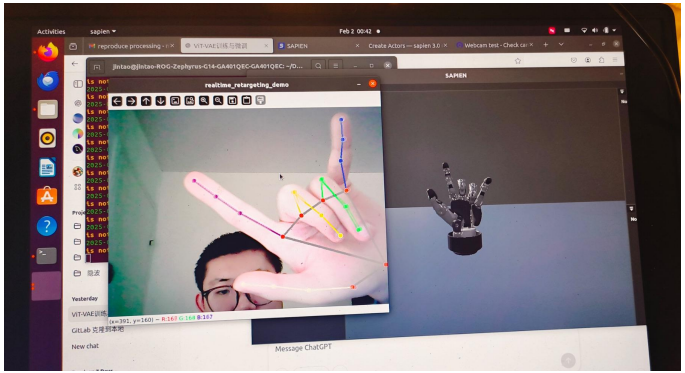
Link: https://drive.google.com/file/d/1tnQRBRV70f7sH1B5BF6acbjSkgz5Q01Z/view?usp=sharing

Reproduction(Hand Pose Retargeting)

Real-time Mode Retargeting(Real hand VS hand on Sapien)-different postures with shadow and SVH robot hands

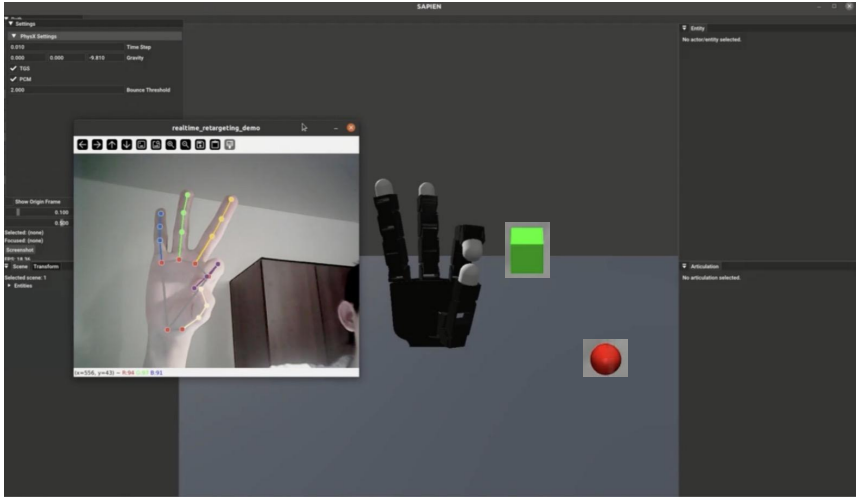




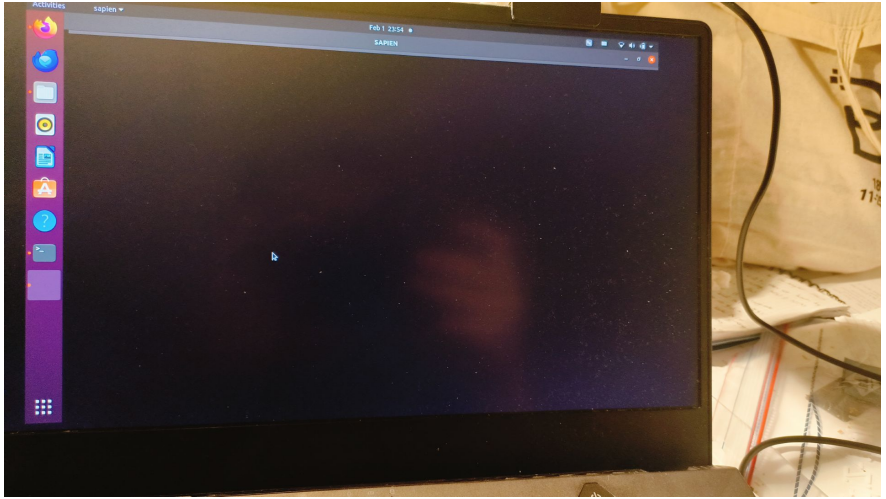


Issue for data collection

- The main issue is **system stability**. After multiple retargeting attempts or prolonged use (2-3 mins), the system fails to capture images, likely due to using an **external USB camera** on a computer without a **built-in camera**, posing a **Objective** hardware limitation.

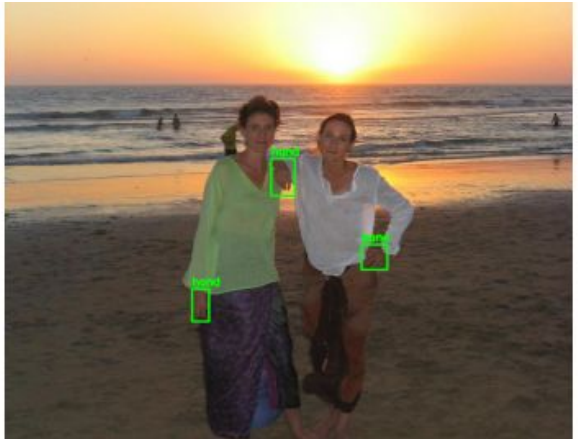


Original Plan



Collapsed systems

Hand Bounding Box Detection:



Background:

- **Supervised learning**: the model is trained on labeled data
- **one-stage detection**: directly predicts bounding boxes and class labels in one step
- **two-stage detection**: first generates region proposals and then refines them for classification
- **bounding box**: A rectangular box used in object detection to define the location and size of an object within an image.

Confidence Score

Confidence threshold is the **minimum** confidence score required for a detected object to be considered valid and not discarded.

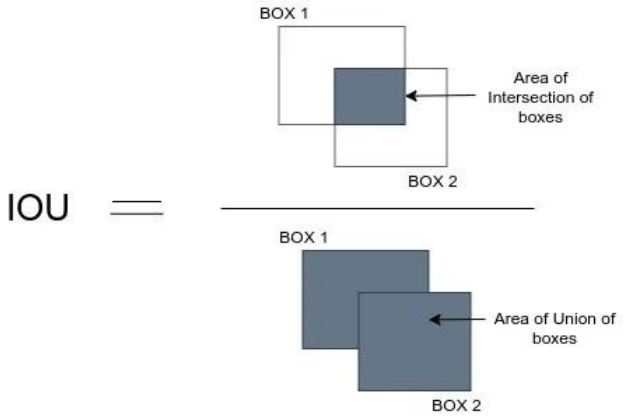
Confidence Threshold= 0.3

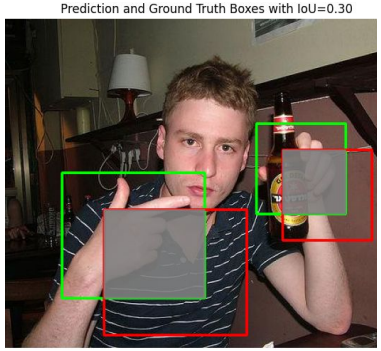
Increasing it leads to **missed detections**.

Decreasing it leads to **false detections**.

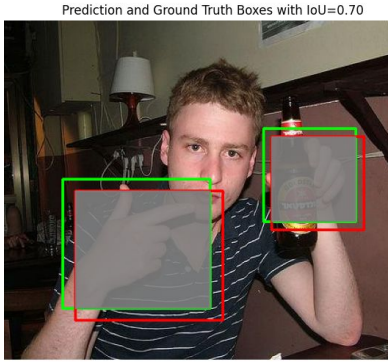
IoU

IoU (Intersection over Union) measures the overlap between the predicted bounding box and the ground truth bounding box, calculated as the ratio of their intersection area to their union area.

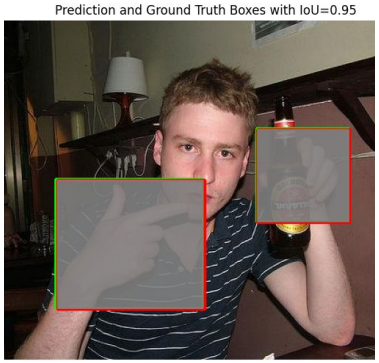




Poor(IoU=0.30)



Good(IoU=0.70)



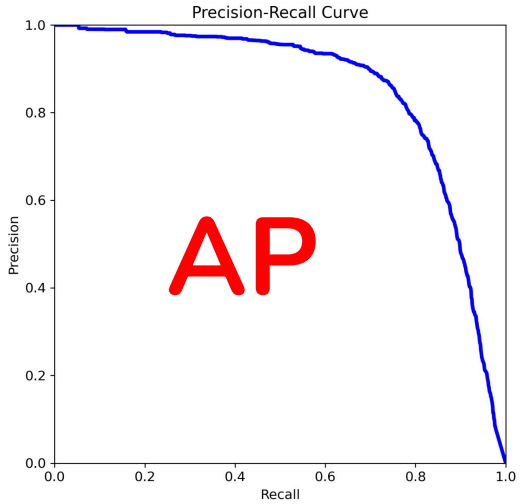
Excelent (IoU=0.95)

Confusion Matrix

	(Confidence Score ≥ Threshold)	(Confidence Score < Threshold)
(IoU ≥ Threshold)	**True Positive (TP)** ✅ Correct detection	**False Negative (FN)** ❌ Missed detection
(IoU < Threshold)	**False Positive (FP)** ❌ False detection (wrong object)	**True Negative (TN)** ✅ Correct rejection (no object detected)

Metric	Formula	Purpose
Precision	$\frac{TP}{TP+FP}$	Reduces false positives (FP)
Recall	$\frac{TP}{TP+FN}$	Reduces false negatives (FN)
F1-Score	$2 \times \frac{P \times R}{P+R}$	Balances Precision and Recall

mAP



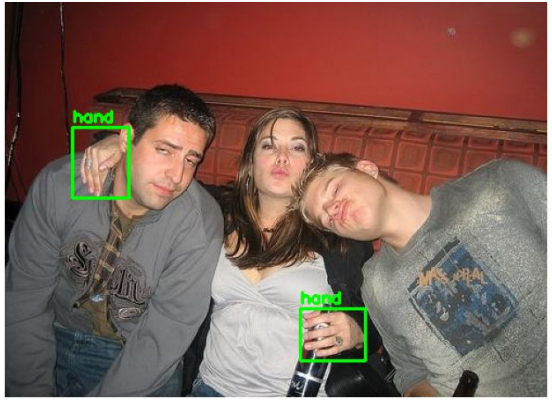
AP(Average Precision) measured the model's detection performance across different recall levels.

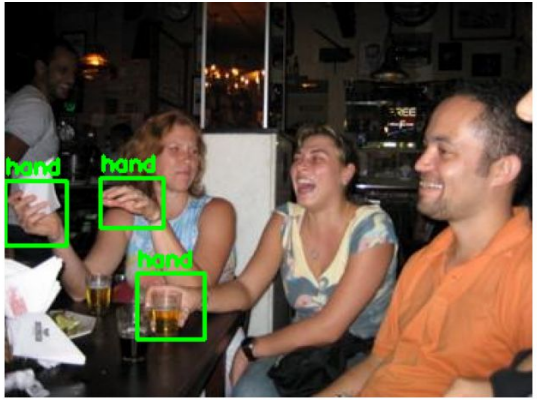
$$AP = \int_0^1 P(R)dR$$

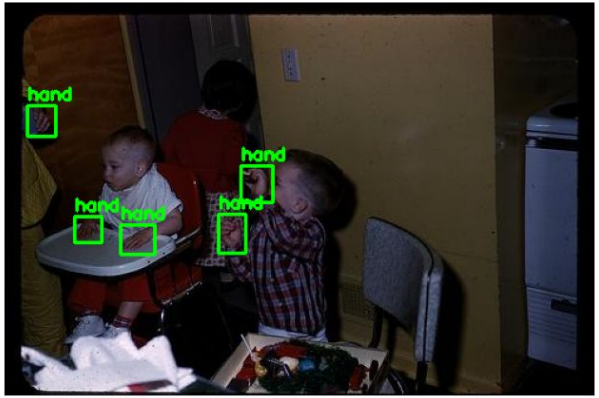
mAP(mean Average Precision) is the mean of AP values across all classes.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

Dataset



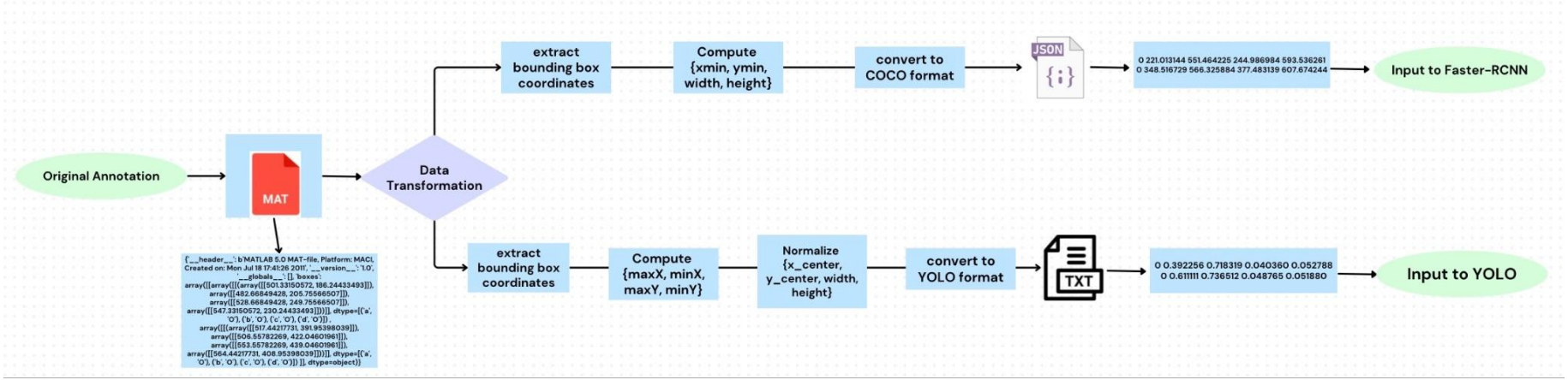




Dataset Characteristics:

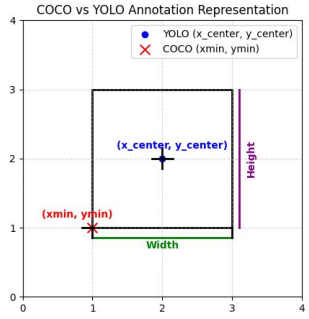
- **Train Dataset:** 4089 images and 9163 hand instances,
- **Vaidation Dataset:** 738 images and 1856 hand instances,
- **Test Dataset:** 821 images and 2031 hand instances
- from oxford hand dataset: https://www.robots.ox.ac.uk/~vgg/data/hands/

Annotation Transformation(Label Data)



Key Points:

- The original `.mat` file contains bounding box annotations, where each box is defined by four corner points (`a, b, c, d`), each represented as a array with `(x, y)` coordinates.
- COCO format stores bounding boxes as absolute coordinates `(xmin, ymin, width, height)` in a structured JSON file.
- YOLO format uses normalized `(x_center, y_center, width, height)` coordinates in a simple TXT file.



Data Augmentation Strategies

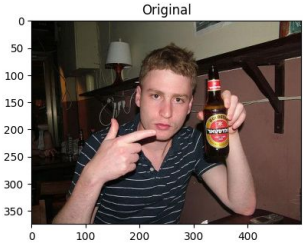
Why:

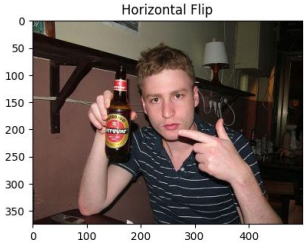
- **Training data** is limited.
- Need to **improve model generalization** and **avoid overfitting.**

How:

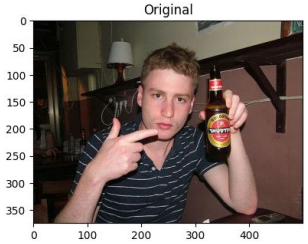
- **Training Set**:
  - Geometric transformations
  - Color & contrast adjustments
  - Blurring & distortion
  - Automated strategies
- **Validation Set**: Only Normalization & Resizing

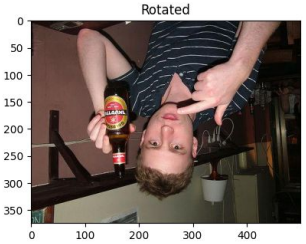
Geometric Transformations



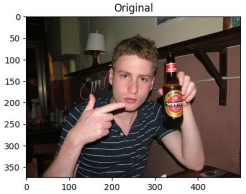


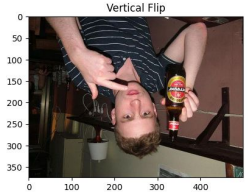
Horizontal Flip



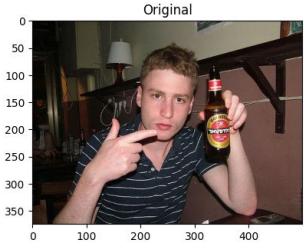


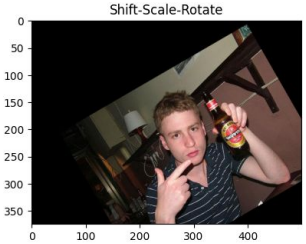
Rotate



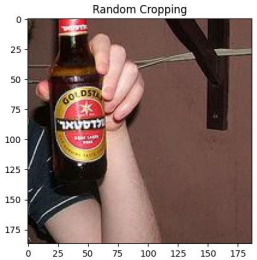


Vertical Flip

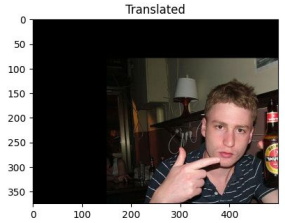




Shift-Scale-Rotate

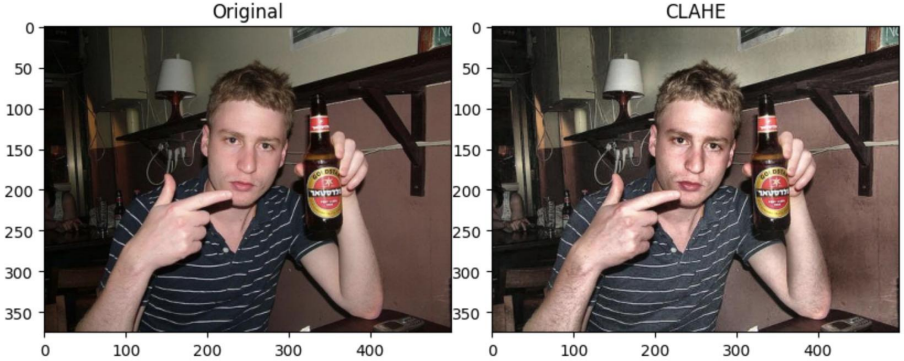


Cropping

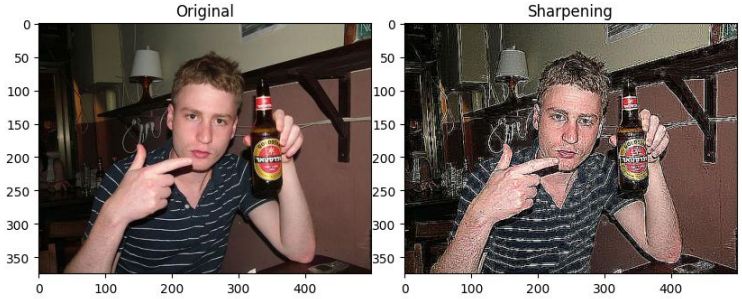


Translation

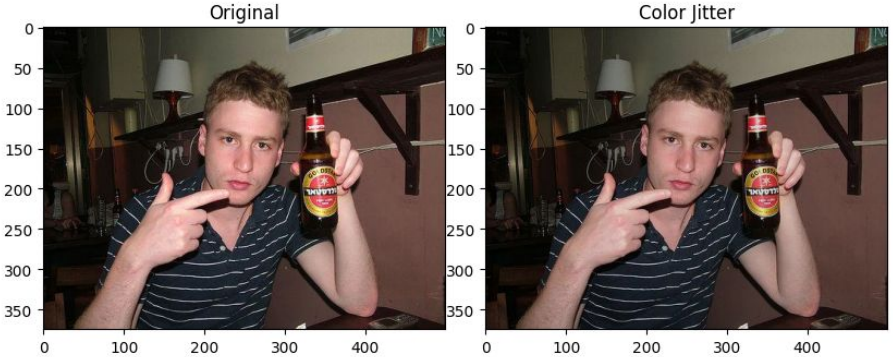
Color & Contrast Adjustments



CLAHE(Contrast Limited Adaptive Histogram Equalization)



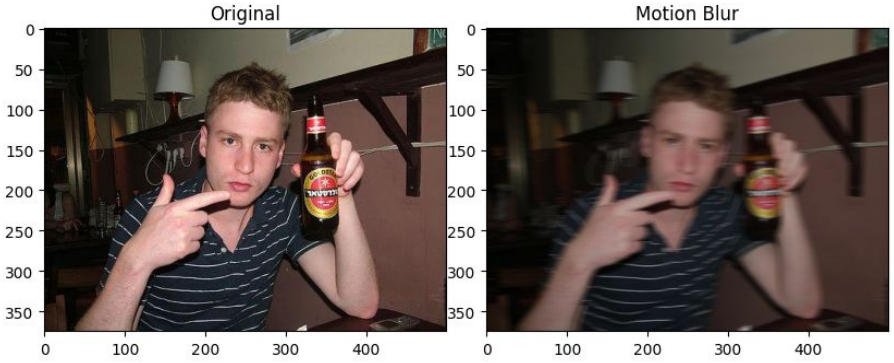
Sharpening



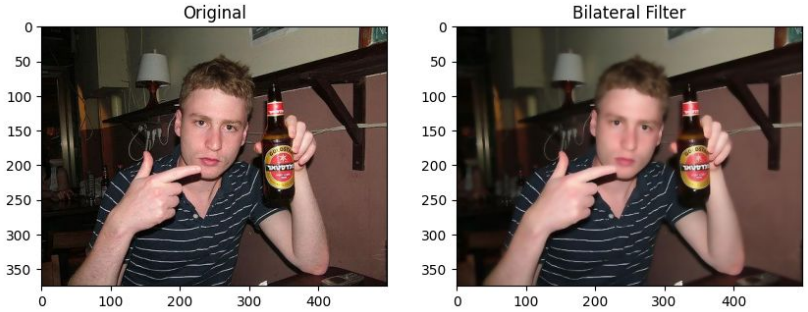
Color Jitter

Hue-Saturation Adjustments

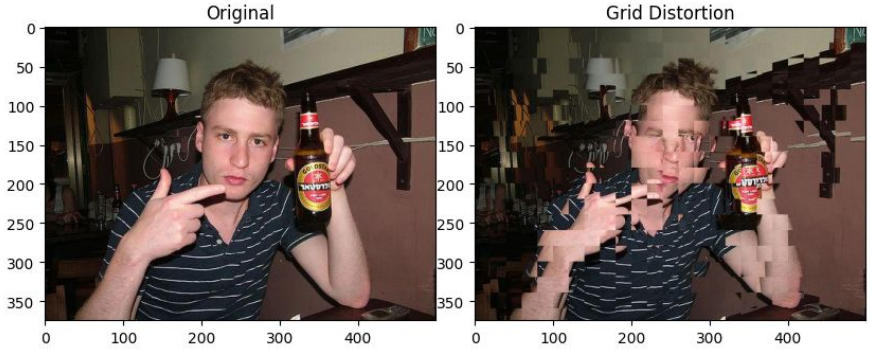
Blurring & Distortion



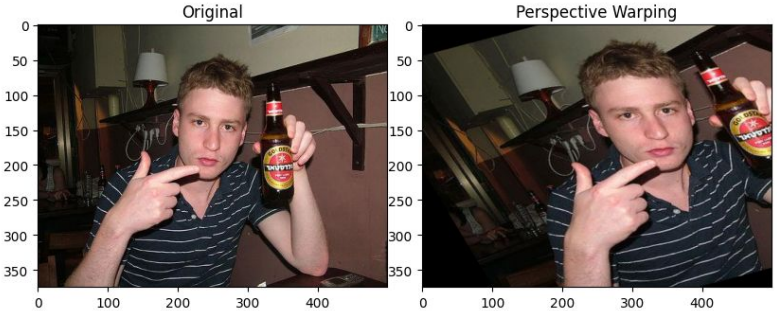
Motion Blur



Bilateral Filter

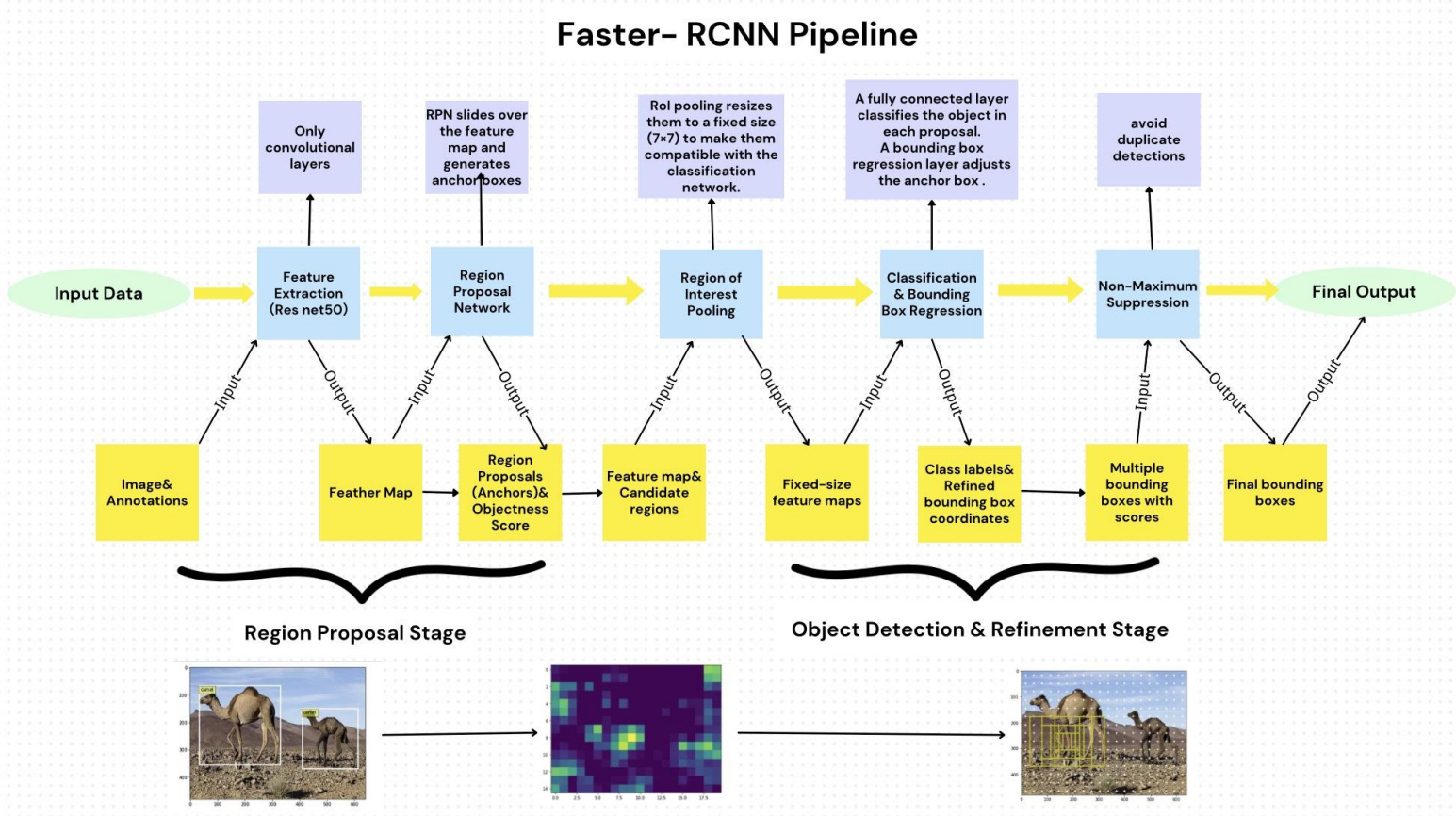


Grid Distortion

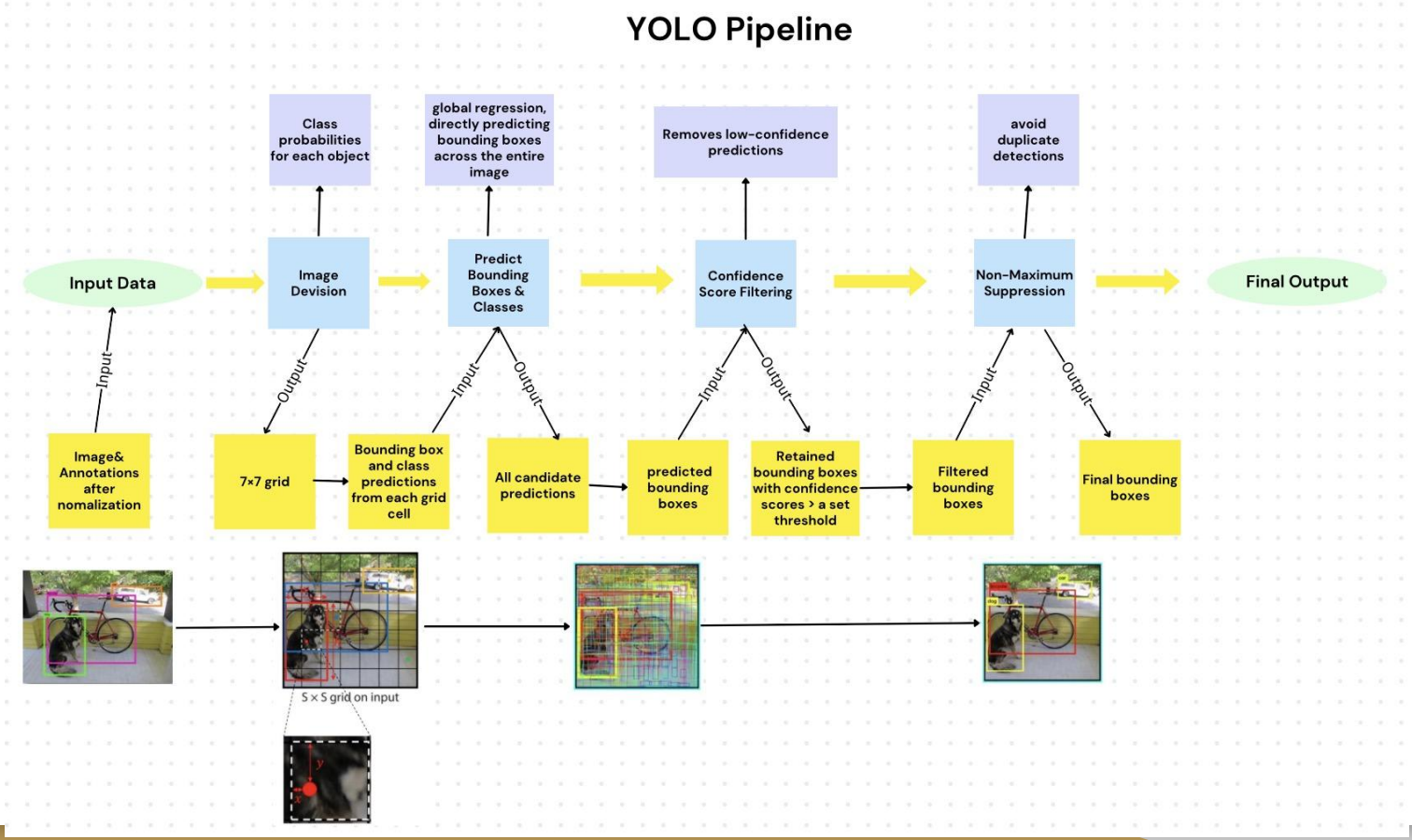


Perspective Warping

Faster-RCNN Methodology Framework



YOLO Methodology Framework

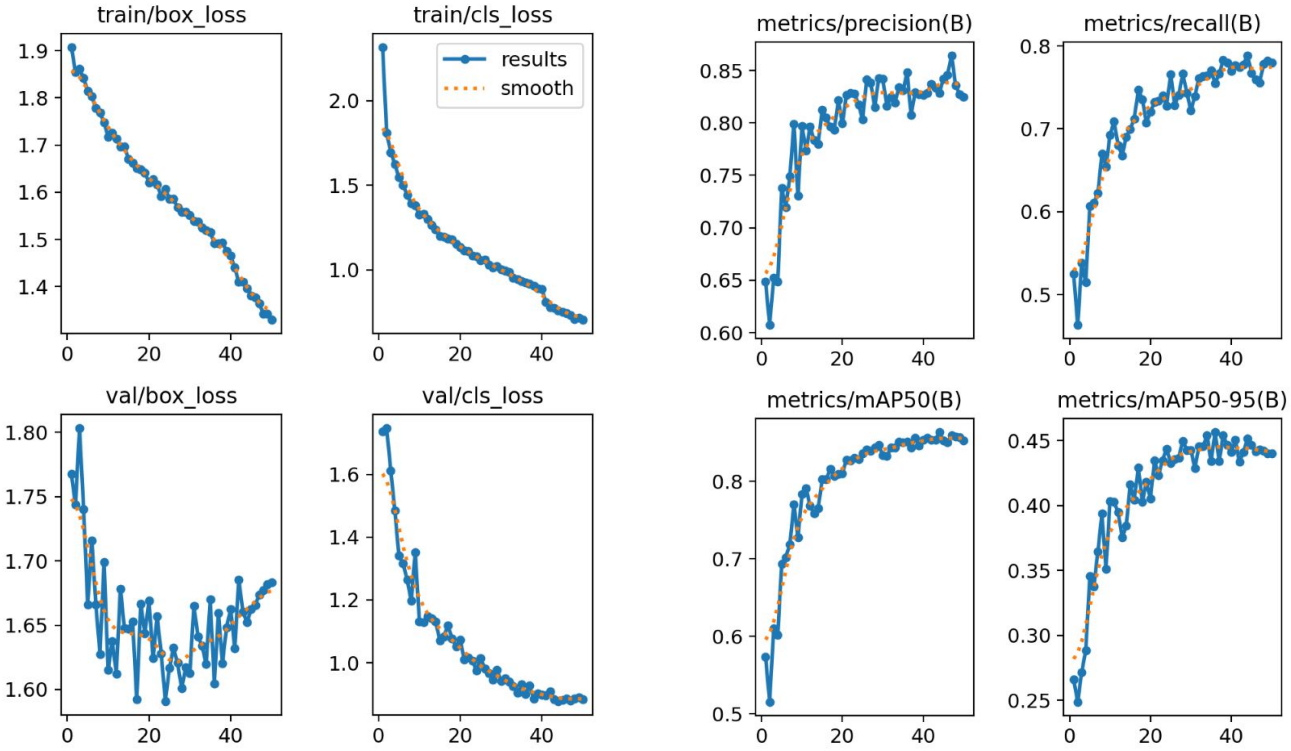


Model Performace

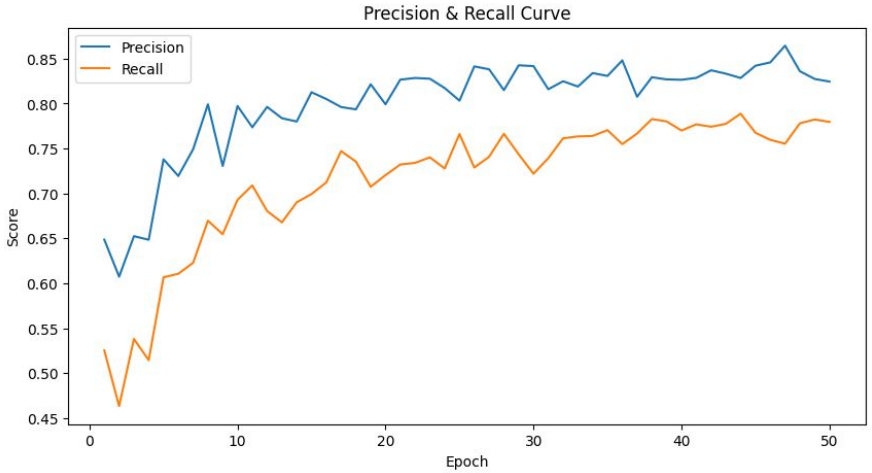
Model	**mAP@50(%)**	**mAP@50-95(%)**
Resnet+Faster-RCNN	63.2	31.7
YOLO V8	86.9	48.2

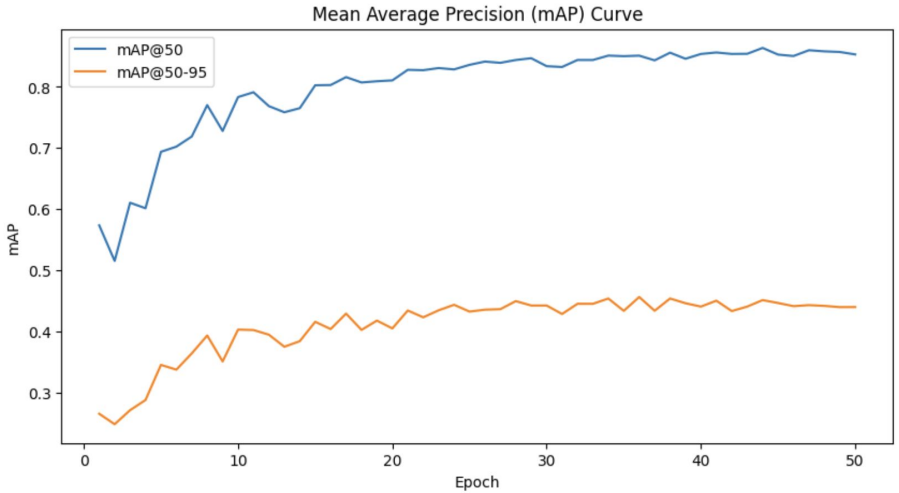
mAP@50 measures AP at IoU ≥ 0.50, while mAP@50-95 averages AP over IoU thresholds from 0.50 to 0.95.

YOLO mAP & Loss Trends

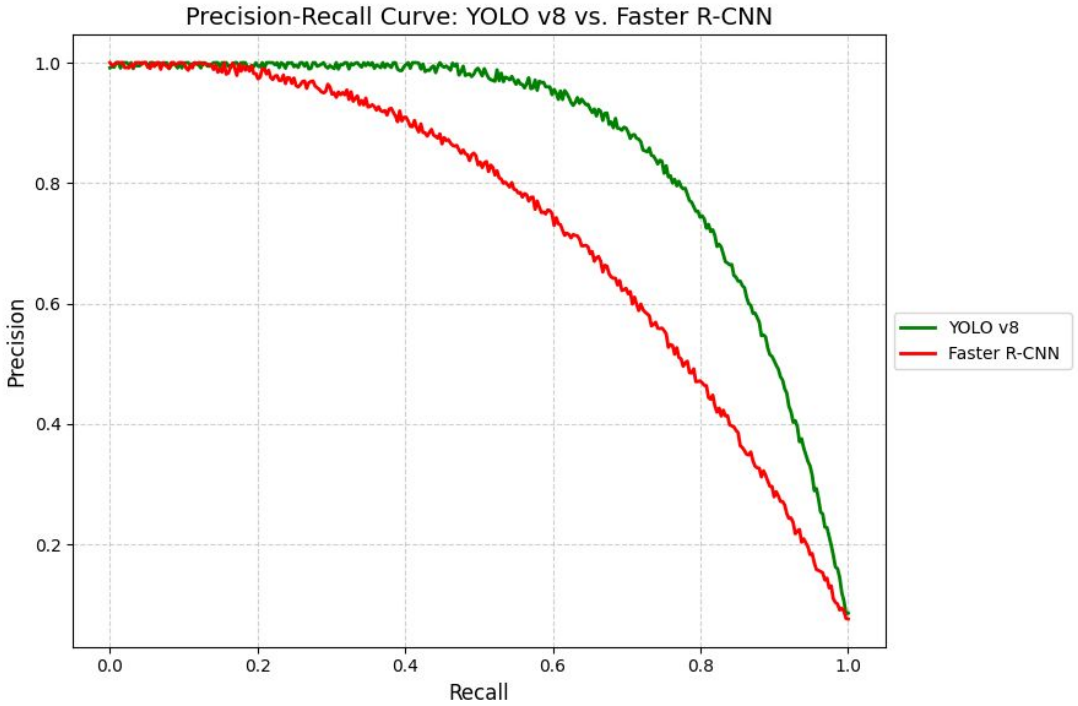


YOLO P&R & maP Trends

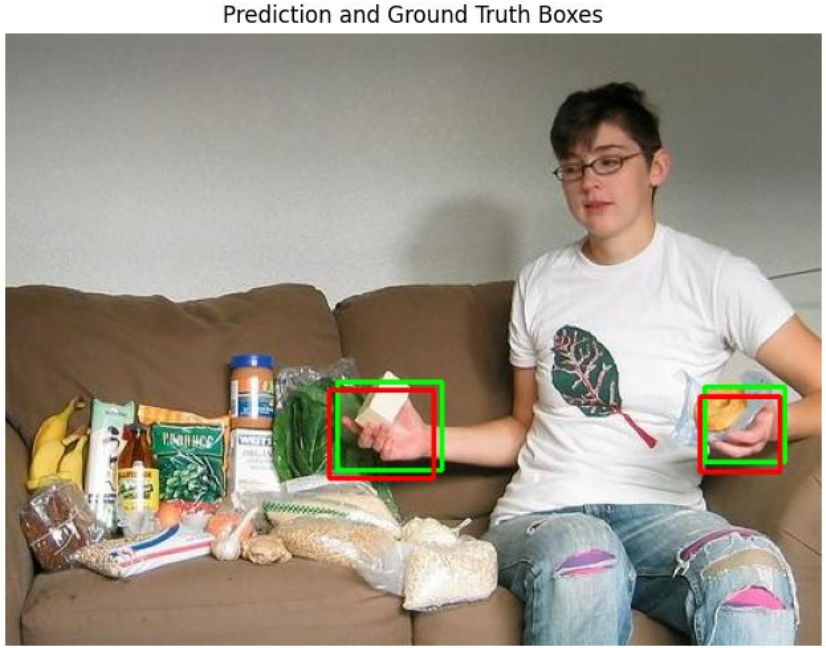




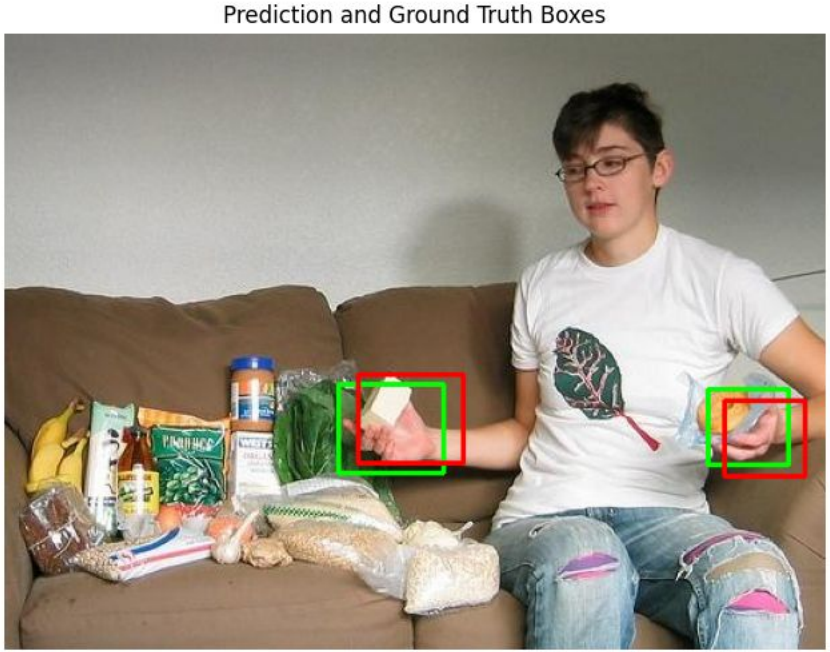
P-R Curve Comparison



YOLO Triumphs Where Faster-Rcnn Stumble(**True Positive**)

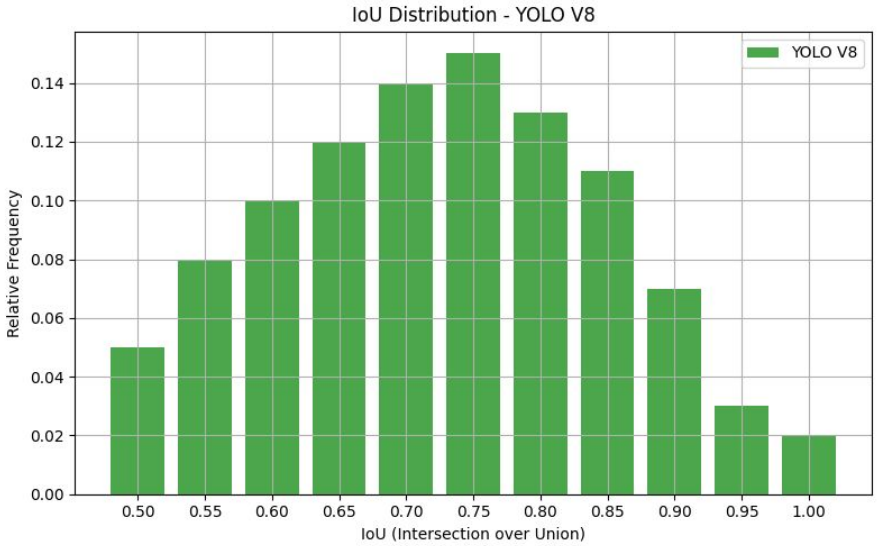


YOLO

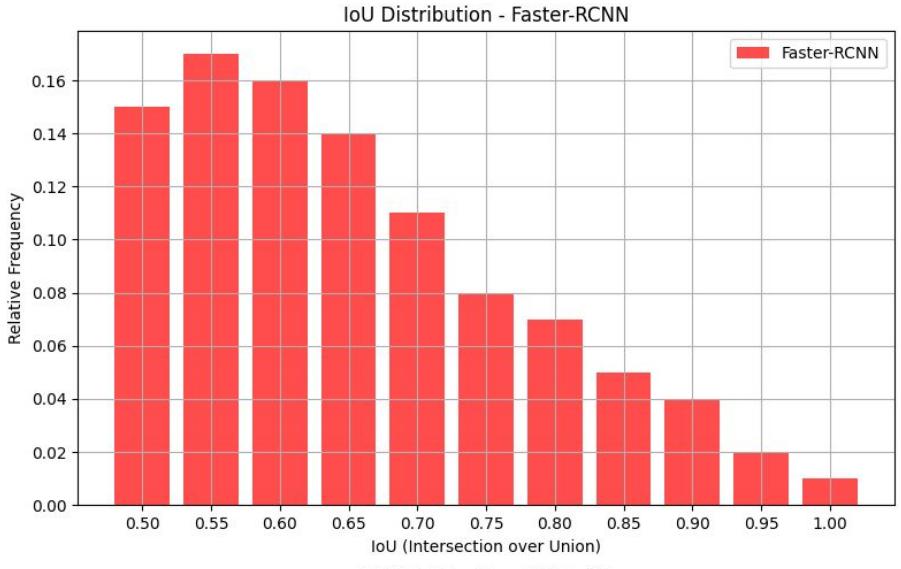


Faster-RCNN

YOLO Triumphs Where Faster-Rcnn Stumble(**IoU Distribution**)

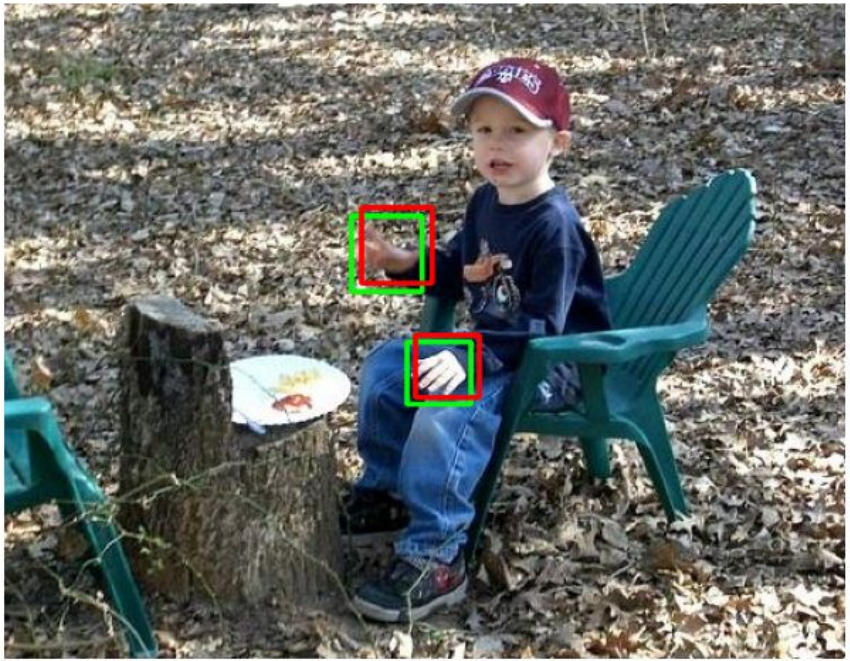


YOLO

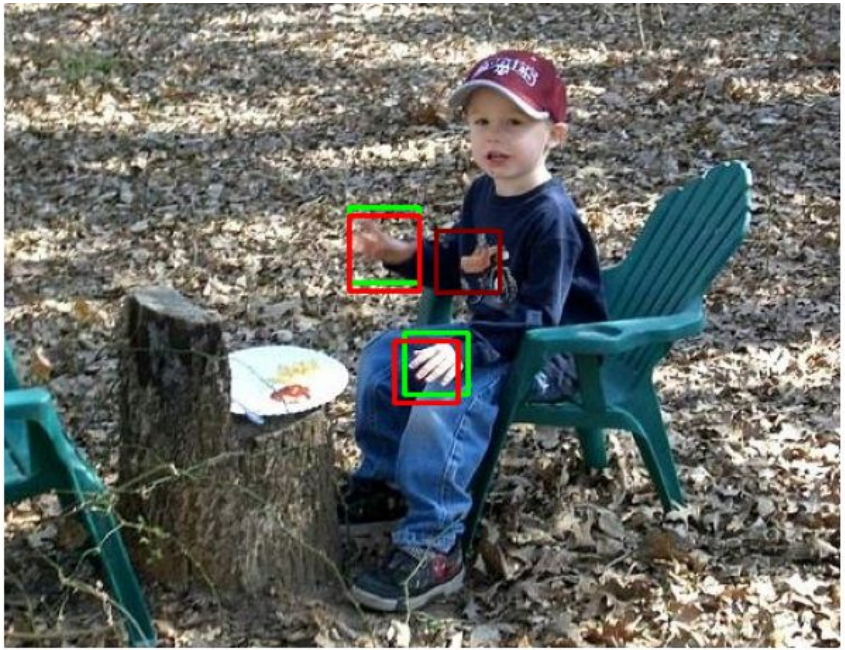


Faster-RCNN

YOLO Triumphs Where Faster-Rcnn Stumble(**False Positive**)

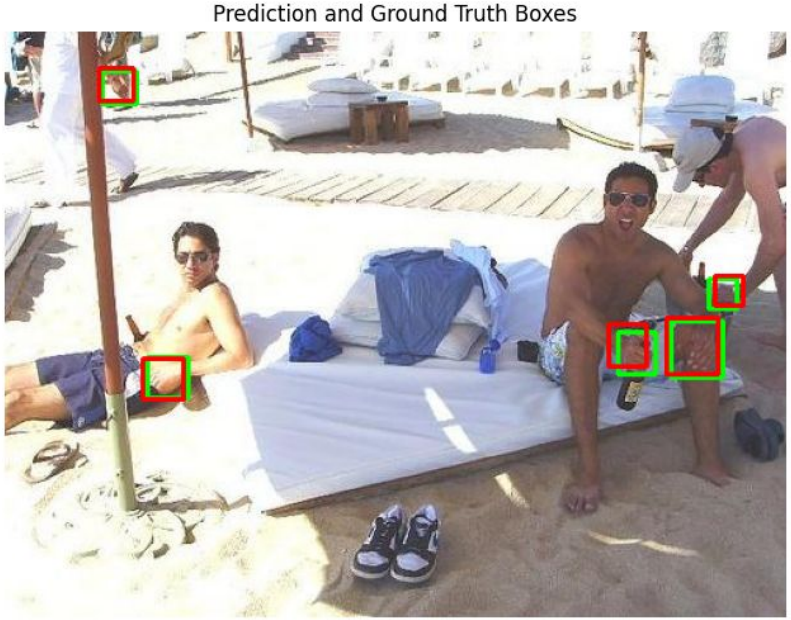


YOLO

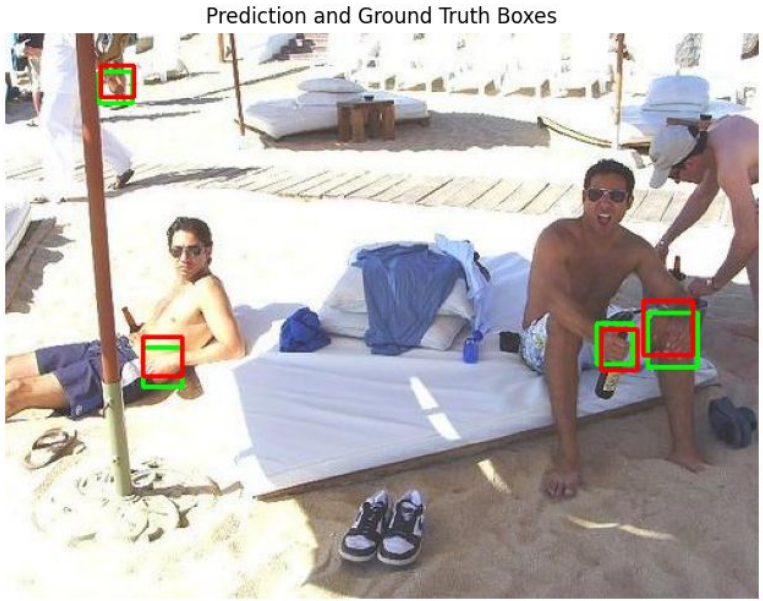


Faster-RCNN

YOLO Triumphs Where Faster-Rcnn Stumble(**False Negative**)

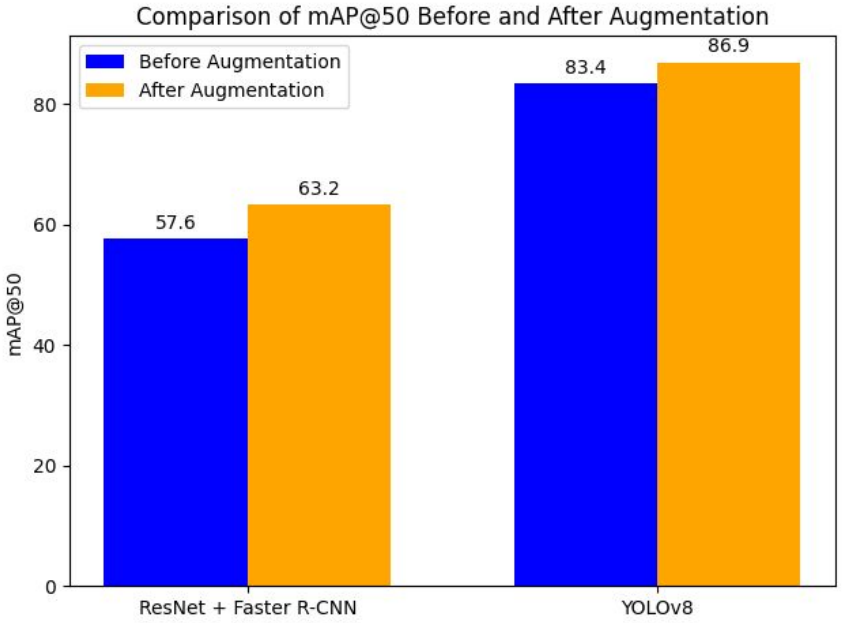


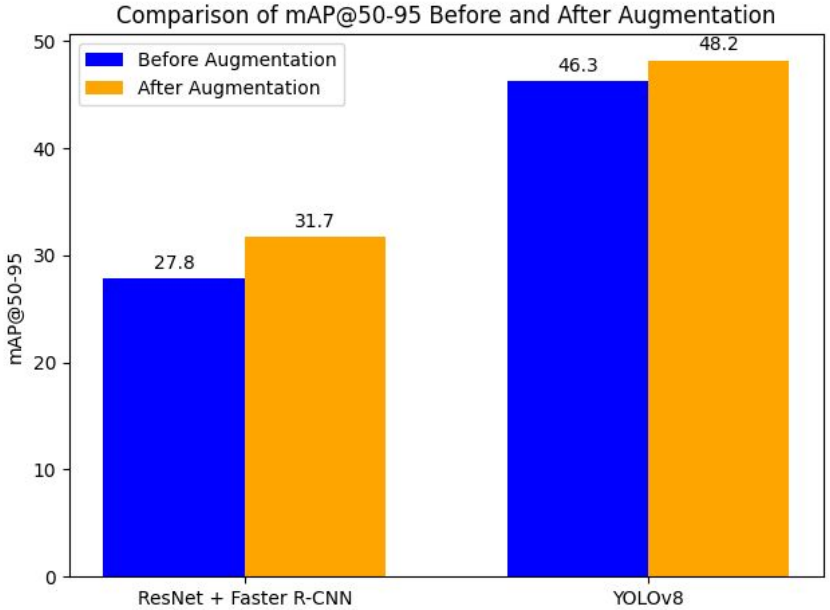
YOLO



Faster-RCNN

How Augmentation Impacts two models





Explanation:

Data augmentation benefits Faster R-CNN more significantly because its two-stage architecture relies on **region proposals (RPN)**, which require diverse training samples to improve robustness. Enhancing data variety helps RPN generate better candidate boxes, leading to **more accurate detections**, especially for small or occluded objects. Additionally, Faster R-CNN is more prone to overfitting on small datasets, making augmentation essential for improving generalization.

YOLOv8, being a one-stage detector, is less dependent on data augmentation as it already incorporates **built-in optimizations** and **Augmentation**. Its design inherently provides strong generalization, making it effective even with minimal augmentation. While data augmentation still improves performance, the gain is smaller compared to Faster R-CNN, as YOLOv8 is already optimized for efficient feature extraction and localization.

Conclusion

Comparison	YOLOv8	Faster R-CNN
Detection Framework	One-stage detection	Two-stage detection
Optimization Strategy	End-to-end optimization	Region proposal-based optimization
Feature Extraction	Efficient and lightweight	Relies on its backbone
Performance on Small Datasets	Highly effective	Struggles due to proposal-based mechanism,needs large amounts of data, otherwise difficult to optimize candidate boxes
Region Proposal Mechanism	No region proposals (direct regression)	Relies on region proposals
Computational Efficiency	Requires fewer computational resources	Higher computational cost

Future Work

I plan to further enhance detection accuracy by exploring **data augmentation techniques and hyperparameter tuning**.

Additionally, incorporating **Transformer-based models or hybrid architectures** could improve feature extraction and localization performance

Project Conclusion

- **Reproduced Anyteleop**
- **Try to perform data collection**
  a. **But teleoperation is a engineer-heavy task**
  b. **It is hardware dependent**
- **Focus on a smaller part: hand detection**
- **Comparison of two object detection models**

References

- Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, Dieter Fox. "AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System." Robotics: Science and Systems (RSS), 2023.
- S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), pages 627–635, 2011.
- J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2016.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.
- S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems (NeurIPS), volume 28, 2015.

CentraleSupélec | université PARIS-SACLAY

Thanks!

Supervisor: **Guillaume Thomas**

Student: Jintao Ma