COS 429: Computer Vision
Final Project Milestone Report

Sonal Bhatia & Yash Thakkar
November 6, 2025

Pedestrian Detection: Benchmarking HOG, CNN, ViT, and SAM2 Approaches on the PnPLO Pedestrian Dataset

Q2. Description of the Problem / Paper to Replicate

The problem we aim to study is pedestrian detection: specifically, how computer vision models distinguish real people from person-like nonhuman objects such as mannequins, statues, or robots. While deep learning-based detectors have achieved high accuracy on general datasets (referring to our discussions of COCO dataset in lecture), they often generate false positives when encountering human-shaped nonhumans in real environments. This leads to safety and reliability concerns in applications like autonomous driving and assistive robotics.

We will use the dataset from the paper "A Novel Dataset for Pedestrian Detection: PnPLO (Persons and Person-Like Objects)" (Karthika et al., 2022). The paper introduces a dataset of 944 annotated images containing both human and person-like nonhuman classes and evaluates the detection performance by training the SSD object detection model on their new dataset and measuring detection accuracy and false positives using metrics such as Average Precision (AP) and mean Average Precision (mAP) to see how well these models differentiate real people from person-like nonhuman objects.

Our goal is to test the generalizability of modern object detection and classification techniques (some of which we learned in this course) on this dataset, comparing:

1. A traditional HOG pipeline (Dalal & Triggs, 2005) then fed into a classifier.
2. A CNN-based detector (Faster R-CNN) (which could confirm/complicate the paper's claim that SSD can run in real time with an accuracy comparable to that of faster RCNN, pg. 1085).
3. A Transformer-based detector (ViT),
4. A segmentation-first approach using SAM2, a foundation segmentation model by Meta.

By reproducing the Karthika et al. paper's core task and extending it to vision transformers and SAM2 segmentation model, we aim to understand whether recent advances meaningfully reduce misclassification of nonhuman objects in the context of pedestrian detection.

Paper Citation: Karthika, N.J., Chandran, S. (2020). Addressing the False Positives in Pedestrian Detection. In: Mallick, P.K., Meher, P., Majumder, A., Das, S.K. (eds) Electronic Systems and Intelligent Computing. Lecture Notes in Electrical Engineering, vol 686. Springer, Singapore. https://doi.org/10.1007/978-981-15-7031-5_103

Kaggle Link:
https://www.kaggle.com/datasets/karthika95/pedestrian-detection?resource=download

Q3. Pointers to Related Course Topics

This project is directly related to course topics such as Histogram of Oriented Gradients (HOG), Segmentation, R-CNN, and Vision Transformer model architectures discussed in Lecture as we will be building and implementing the models for the project. Additionally, the project takes inspiration from the pedestrian detection task discussed in lecture 7 by extending it to new ML models. We aim to use evaluation metrics discussed in the course such as Precision, Recall, F1, Accuracy, and Intersection of Union.

Q4. Plans for Acquiring the Necessary Data and Computational Resources

- Dataset: We will use the open-access PnPLO Pedestrian Detection Dataset from Kaggle (https://www.kaggle.com/datasets/karthika95/pedestrian-detection), which contains ~900 images with XML annotations for "person" and "nonhuman person-like object." No additional data collection should be needed.
- Data Preparation: Annotations likely will need to be parsed from the provided XML format to one that is compatible with the PyTorch implementations of the models we are studying.
- Compute: All experiments will be run on Google Colab Pro (provided Student subscription), using the T4 or A100 GPU compute resources. Given the small dataset size and moderate model complexity, Colab's resources are sufficient.
- Libraries: PyTorch, TorchVision, OpenCV, NumPy, Matplotlib, and Meta's SAM2 library for segmentation.

Q5. Plans for Quantitative and Qualitative Evaluation

Quantitative Metrics:

- IoU ≥ 0.5: Detection considered correct if overlap exceeds 50%.
- AP and mAP: Average Precision per class and mean AP across classes.
- False Positive Rate: Specifically tracking misdetections of nonhuman objects as "person."
- F1 Score: Balance Precision and Recall metrics.

Qualitative Evaluation:

- Visualize predicted bounding boxes and segmentation masks across models.
- Compare representative success and failure cases, especially where mannequins or statues are mistaken for humans.
- Generate Precision-Recall curves to illustrate trade-offs.

Together, these evaluations will measure both accuracy and robustness to ambiguity, thus revealing whether modern architectures like ViT improve interpretability and reliability in pedestrian-like scenarios.

Q6. Target Outcome

By the end of the project, we will aim to deliver:

1. A Colab notebook pipeline implementing all four detection methods (HOG+Classifier, Faster R-CNN, ViT, SAM2).
2. A quantitative comparison (AP, mean AP, F1-Score) of detection performance across the two classes (person and non-person).
3. Visual examples of successful and failed detections.
4. A technical report summarizing experimental findings and discussing whether newer transformer and segmentation-based methods reduce false positives relative to classical and CNN baselines.

Q7. Fallback Plan

Potential challenges include:

- Compute limitations on Colab (session timeouts, GPU limits).
- SAM2 model integration issues due to large memory usage and compute restrictions.
- Limited dataset size leading to overfitting.

If these arise, the minimum deliverable will include:

- Full evaluation of HOG+SVM, Faster R-CNN, and ViT models (both feasible on CPU/GPU), removing SAM2 from our study.
- A complete quantitative comparison between the various architectures we described with clear documentation and visual examples.

Q8. Optional Questions / Concerns

1. Is comparing false positives on person-like nonhumans sufficient as a primary research metric, or should we emphasize mean average precision more heavily in evaluation?
2. Is it feasible to create four separate evaluations of different architectures (HOG, R-CNN, ViT, SAM) to compare their performance? Or should the focus be only on two?