Robustness in Pedestrian Detection
Benchmarking HOG, Faster R-CNN, SAM2, & DETR Models on the PnPLO Dataset

Sonal Bhatia & Yash Thakkar
Department of Computer Science, Princeton University, Princeton, NJ 08544

Abstract

This report investigates robustness in pedestrian detection by benchmarking four representative computer vision architectures (HOG, Faster R-CNN, DETR, and YOLO + SAM2) on the PnPLO dataset, which is designed to evaluate performance under person-like visual ambiguity. Rather than focusing solely on accuracy, we emphasize robustness to false positives arising from visually similar nonhuman objects such as statues and mannequins, as these errors impose significant downstream computational costs in robotic and surveillance systems. Using complementary metrics including F1 score, false positive rate, and mean average precision (mAP), we show that classical HOG-based methods suffer from high false positive rates due to reliance on low-level gradient cues, while faster R-CNN reduce false positives through stronger semantic representations. YOLO and Segmentation-first approaches using SAM2 achieve the lowest false positive rate and highest mAP, indicating strong robustness to person-like distractors, whereas transformer-based DETR models demonstrate effective global reasoning and abstention behavior but lower overall precision under ambiguity. Our study demonstrates that robustness to person-like visual ambiguity varies fundamentally across detection paradigms, showing that architectural choices, not accuracy alone, determine false positive behavior and downstream system efficiency in real-world robotic and surveillance pipelines.

1. Introduction

1.1. Problem Statement & Motivations

Vision-based object detection systems are increasingly deployed in always-on, resource-constrained robotic and surveillance environments like home robots, indoor monitoring systems, assistive devices, and autonomous vehicles[1] [2] [9]. In these settings, robustness is as critical as accuracy in detection, as systems must reliably detect humans to ensure safety and appropriate interaction, while also avoiding unnecessary escalation of computationally expensive behaviors in response to visual ambiguity [11].

A common failure mode in such systems is the misclassification of person-like nonhuman objects like statues, mannequins, and human-shaped objects as real people. While conservative behavior is often appropriate, especially in safety-sensitive contexts like pedestrian detection, repeated false positives impose significant costs and affect the overall reliability of the system[11]. In home robotics and surveillance pipelines, pedestrian detections often trigger downstream processes such as identity tracking, pose estimation, action recognition, or human-robot interaction logic [13]. These errors are costly not only in terms of computation, but also in terms of system reliability and downstream decision quality. In persistent false positive scenarios, higher-level modules may repeatedly reason about nonexistent humans, leading to unstable behavior, unnecessary interventions, and degraded user trust.

Similar issues arise in autonomous driving where pedestrian detections invoke costly planning and prediction modules, but in a safety-critical scenario, the focus remains on ensuring the number of false negatives is zero. Downstream modules in an autonomous stack are often expensive for pedestrians than for static or inanimate obstacles. For example, tracking a person's movement as they cross the street compared to tracking a static statue on the sidewalk will require different modules with different compute requirements [17]. As a result, poor robustness at the perception level can propagate inefficiencies throughout the system. Nevertheless, robustness in pedestrian detection is part of a larger challenge of object detection, where an autonomous driving system would first need to ensure safety and prevent any object collisions before evaluating inefficiencies in the system.

This project treats pedestrian detection not purely as a classification problem, but as a robustness problem: how well do vision systems maintain calibrated, reliable behavior when faced with visually similar but semantically different objects? We argue that robustness to person-like distractors is essential not because such objects should be ignored, but because real-world environments routinely contain such ambiguities, and failure to handle them gracefully reveals limitations in how detection systems generalize beyond curated benchmarks.

1.2. Goals & Scope

The goal of this project is to evaluate and analyze the robustness of pedestrian detection systems in the presence of person-like visual distractors, with particular emphasis on their downstream impact in robotic and surveillance systems. In this work, robustness is defined as a system's ability to maintain a low false positive rate when exposed to visually similar but semantically different person-like objects, without sacrificing recall on true pedestrians.

Accuracy alone is a poor measure of robustness in this context, as high accuracy can be achieved while still producing frequent, high-confidence false positives on person-like objects that trigger unnecessary downstream computation. We study how four different classification and detection model architectures (Histogram of Gradients, R-CNN, Vision Transformer, and YOLO + Segmentation) behave under ambiguity and how confidently they make predictions. By examining models of varying complexity, we investigate whether more modern, advanced architectures meaningfully improve robustness, or merely amplifies confidence in ambiguous predictions. This

framing allows us to evaluate results in the context of both system efficiency and model design tradeoffs, rather than treating false positives as interchangeable errors.

1.3. PnPLO Dataset

To support robustness-focused evaluation, we use the PnPLO (Person and Person-Like Objects) dataset, which is designed specifically to assess pedestrian detectors under person-like visual ambiguity [8]. In addition to real pedestrians, the dataset includes statues, mannequins, and other visually similar nonhuman figures that commonly trigger false positives in detection systems.

The PnPLO dataset consists of two classes: `person` and `person-like`. The person-like class includes objects such as statues, mannequins, and human-shaped figures. The dataset contains 944 training images, 160 validation images, and 235 test images, with approximately balanced class distributions across splits to mitigate class imbalance effects.

Nonhuman annotations were manually labeled using the `labelImg` tool and stored in PASCAL VOC format. Objects marked as difficult were excluded, as these cases are visually ambiguous even to human annotators and would confound robustness analysis. Person images were sourced in part from the PASCAL VOC 2007 and 2012 datasets, ensuring diversity while maintaining balanced class distributions. This binary labeling setup enables targeted analysis of false positives arising specifically from person-like nonhuman objects.

1.4. Background & Related Work

Pedestrian detection has evolved through several major methodological paradigms, each introducing tradeoffs between robustness, computational cost, and representations. In this section, we briefly review the four representative approaches evaluated in this work: classical feature-based detectors, convolutional neural network (CNN) detectors, segmentation-first models, and transformer-based architectures.

Histogram of Gradients (HOG). The Histogram of Gradients (HOG) descriptor encodes local edge orientation statistics within sliding windows and, when paired with a linear Support Vector Machine (SVM), became a foundational baseline for pedestrian detection [6]. HOG captures upright human silhouettes effectively but relies exclusively on local gradient structure, lacking semantic understanding or global context. As a result, HOG-based detectors are particularly vulnerable to person-like nonhuman objects that exhibit similar contours or poses. These robustness failures are largely systematic, making false positives a well-known limitation of gradient-based approaches.

Faster R-CNN. CNN-based detectors such as Faster R-CNN introduced region proposal networks and end-to-end feature learning, substantially improving detection performance in complex scenes [15]. Learned hierarchical features provide stronger semantic discrimination than hand-crafted descriptors, often reducing false positives from background clutter. However, robustness gains remain heavily dependent on training data diversity and confidence calibration. CNN-based detectors can still assign high confidence to visually familiar but semantically incorrect objects, indicating that robustness is empirical rather than architecturally guaranteed.

Segmentation-First Models (SAM2). Segmentation-first approaches decouple object localization from classification by producing class-agnostic masks prior to semantic labeling. The Segment Anything Model (SAM) introduced a prompt-able segmentation paradigm trained on large-scale annotated data [10], while SAM2 extends this framework to video and temporal consistency [14]. This design contrasts with end-to-end detection models such as YOLO, which jointly optimize localization and classification in a single forward pass and therefore entangle geometric precision with semantic confidence. While precise object boundaries from segmentation pro-

vide strong geometric cues, segmentation alone does not resolve semantic ambiguity. Robustness to person-like distractors therefore depends entirely on downstream classification, raising questions about whether improved geometry alone suffices to reduce false positives.

Vision Transformers (DETR). Vision transformer-based detectors reformulate object detection as a set prediction problem using global self-attention [5]. By modeling long-range dependencies and avoiding hand-crafted heuristics such as anchor boxes and non-maximum suppression, DETR incorporates global scene context that may improve robustness. However, recent analyses show that transformer-based detectors can exhibit calibration issues and overconfident predictions in visually ambiguous or out-of-distribution settings [12], making their robustness behavior an open empirical question.

Taken together, prior work suggests that robustness to person-like objects is not explicitly addressed by any single detection paradigm. Classical methods fail due to limited representational power, while modern deep models risk overconfidence despite stronger semantics. Segmentation-first approaches improve geometry but defer semantic robustness, and transformer-based detectors trade local cues for global context without clear guarantees on ambiguity resolution. Motivated by these gaps, our work provides a controlled comparative study of robustness across these paradigms using the PnPLO dataset, focusing on false positive behavior under visual ambiguity.

2. Methodology & Implementation

2.1. Data Extraction & Preparation

All four models evaluated in this study use the PnPLO dataset and therefore share a unified data ingestion and annotation parsing pipeline to ensure that observed performance differences reflect modeling choices rather than preprocessing artifacts. The dataset is organized into training, validation, and test splits, each containing a list of image identifiers, a `JPEGImages/` directory with RGB images, and an `Annotations/` directory with PASCAL VOC-style XML files.

For each split, annotation files are parsed to extract object class labels and bounding box coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$. Labels are normalized into a binary classification setting, where `person` instances are treated as positive examples and `person-like` nonhuman objects are treated as negative examples. No annotations outside these two categories are present in dataset.

Data are organized at the image level rather than treating individual bounding boxes as independent samples. Each image is represented by its file path and a set of associated annotations, allowing downstream models to operate on full images or localized regions while preserving contextual information. Images without valid annotations after filtering are discarded. To monitor dataset composition, the number of positive and negative instances is tracked independently for each split. All extracted features, trained models, and evaluation artifacts are stored under a `results_model/` directory to enable efficient reuse and ablation without reprocessing the dataset.

2.2. Evaluation Metrics

We evaluate all four models using a common set of quantitative metrics designed to capture both standard detection quality and robustness to person-like false positives. In addition to reporting aggregate performance, we emphasize metrics that reflect system-level costs in robotics and surveillance settings, where false positives can repeatedly trigger expensive downstream modules.

Precision and Recall. Precision measures the fraction of predicted pedestrians that are correct, while recall measures the fraction of true pedestrians that are successfully detected. Formally,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Precision reflects robustness to false positives (especially on `person-like` objects), while recall reflects sensitivity to real pedestrians.

F1 Score. The F1 score summarizes the balance between precision and recall as their harmonic mean:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

We report class-specific F1 scores at fixed confidence thresholds to reflect realistic operating points for deployment.

False Positive Rate (Robustness Metric). Because our primary robustness concern is nonhuman objects being incorrectly classified as people, we explicitly report the false positive rate (FPR) with respect to the `person-like` class:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{4}$$

where TN denotes true negatives. In our setting, this corresponds to the fraction of `person-like` objects incorrectly predicted as `person`. This metric directly captures robustness failures that can trigger unnecessary downstream computation.

Precision-Recall Curves and Average Precision (AP). Precision-Recall (PR) curves visualize the tradeoff between precision and recall as the confidence threshold varies. From a PR curve, average precision (AP) summarizes the area under the curve and provides a threshold-independent measure of ranking quality. We report AP for each class when applicable.

ROC AUC. Receiver operating characteristic (ROC) curves plot true positive rate versus false positive rate as the decision threshold varies. The area under this curve (ROC AUC) provides a threshold-independent measure of separability in score space. ROC AUC is reported for classifier-style models and is useful for diagnosing overlap between `person` and `person-like` predictions.

Intersection-over-Union (IoU). For detection models, we determine whether a predicted bounding box matches a ground-truth box using intersection-over-union:

$$\text{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}, \tag{5}$$

where B_p is the predicted box and B_g is the ground-truth box. A prediction is considered a correct localization if its IoU exceeds a chosen threshold (e.g., $\text{IoU} = 0.5$).

Mean Average Precision (mAP). For object detectors, we report mean average precision (mAP), which averages AP across classes and, when applicable, across multiple IoU thresholds. We report mAP at $\text{IoU} = 0.50$

(mAP@0.5) as well as mAP averaged across IoU thresholds from 0.50 to 0.95 (mAP@0.50:0.95) to evaluate both detection quality and localization tightness.

Confusion Matrices (Including Background). We use confusion matrices to provide a discrete summary of performance at fixed operating points. For classifier-style models, the matrix captures `person` vs. `person-like` misclassifications. For detection models, we additionally include a background category to capture missed detections (unmatched ground-truth boxes) and, when applicable, spurious detections in empty regions. Although PnPLO contains only two semantic classes, background appears in evaluation because detection frameworks allow for predictions (or misses) that do not correspond to any labeled object.

3. Implementation

3.1. Histogram of Gradients (HOG)

3.1.1. System Overview. As a classical baseline, we implement a Histogram of Gradients (HOG) feature extractor paired with a linear classifier to evaluate how far hand-crafted gradient features can go in distinguishing real pedestrians from visually similar nonhuman objects in the PnPLO dataset. The primary purpose of this baseline is not to maximize accuracy, but to provide a transparent and interpretable reference point for robustness analysis, with particular emphasis on false positives arising from person-like objects.

Our implementation follows the general architecture introduced by Dalal and Triggs [6], with modern preprocessing, explicit hyperparameter optimization, and evaluation tailored to robustness rather than aggregate accuracy. Data ingestion and annotation parsing are shared with the other models in this study to ensure that performance differences reflect modeling choices rather than preprocessing artifacts.

3.1.2. Feature Extraction. HOG features are extracted at the object level using annotated bounding boxes. For each labeled instance, we perform the following steps:

- Crop the image region corresponding to the ground-truth bounding box.
- Convert the cropped region to grayscale, as HOG operates on intensity gradients.
- Resize the crop to a fixed window size to satisfy the descriptor's geometric requirements.
- Compute the HOG feature vector and flatten it into a one-dimensional representation.

3.1.3. Hyperparameter Optimization. Rather than treating HOG as a fixed descriptor, we explicitly optimize its configuration to improve discriminative performance on the PnPLO dataset. This step is critical, as prior work demonstrates that HOG performance can vary significantly with descriptor parameters [3]. We perform a grid search over a predefined set of HOG hyperparameters, where for each candidate configuration, we first enforce geometric consistency constraints required by the HOG formulation, such as divisibility of block dimensions by cell dimensions and proper alignment of the sliding window with the block stride. Invalid parameter combinations are discarded to prevent malformed descriptors. For each valid configuration, HOG features are extracted independently for the training and validation splits using a shared extraction function. Hyperparameter selection is driven exclusively by validation set performance, preventing information leakage from the test set.

3.1.4. LinearSVC Training and Validation. For each HOG configuration, we train a linear Support Vector Classifier (LinearSVC) using the extracted training features. We employ an L2-regularized linear classifier with a fixed regularization strength ($C = 0.01$) to compensate for minor class imbalance, and an increased maximum iteration count to ensure convergence. Linear classifiers are well suited to HOG's high-dimensional feature space and enable efficient optimization via the LibLinear backend. After training, each model is evaluated on the validation set using accuracy, per-class F1 scores, ROC AUC, and Precision-Recall AUC.

3.1.5. Final Training and Test Evaluation Once the optimal HOG parameters are identified, features are re-extracted for the training and test sets using the selected configuration. A final LinearSVC model is trained on the full training set using the same classifier settings as during tuning. To support a deeper analysis, we also extracted raw decision function scores from the trained classifier. These scores are used to generate probability histograms, Precision-Recall curves, and threshold-based analyses presented later in the report.

3.1.6. LinearSVC vs. Kernel SVM Instead of a traditional kernel-based Support Vector Machine, we use `LinearSVC` from scikit-learn. While kernel SVMs can model nonlinear decision boundaries, they require solving a quadratic programming problem and scale poorly with feature dimensionality. In contrast, LinearSVC optimizes a linear decision boundary directly using LibLinear, offering substantially faster training and better scalability [16]. For HOG features, linear separation is often sufficient and empirically effective [4]. From a robustness perspective, linear classifiers also provide more interpretable decision behavior. The resulting decision function enables direct analysis of confidence distributions and false positive behavior without additional kernel-induced complexity. This aligns with our goal of studying robustness and calibration rather than maximizing raw classification capacity.

3.2. Faster R-CNN

3.2.1. System Overview. Faster R-CNN with a ResNet-50 Feature Pyramid Network (FPN) backbone is a widely used two-stage detection framework that decouples region proposal from classification. In the first stage, a Region Proposal Network (RPN) generates candidate object regions directly from convolutional feature maps [15]. In the second stage, these region proposals are passed to the R-CNN head, which classifies each proposed region and refines its bounding box coordinates. The model is pre-trained on the ImageNet dataset, enabling it to learn generalizable visual features. The goal is to evaluate the model's performance on its ability to distinguish a real human with person-like objects. We hypothesize that a region-based detector will be able to create a rich representation of semantic features to correctly detect and classify between the given classes.

3.2.2. Dataset Handling. To ensure compatibility with torchvision detection models, we implement a custom `torch.utils.data.Dataset` that returns each sample as an `(image, target)` pair. Bounding boxes are stored in `(xmin, ymin, xmax, ymax)` format using absolute pixel coordinates, and labels are stored as integer class indices. A key requirement for Faster R-CNN is variable-length annotation support, as such, each image may contain a different number of objects. For this reason, we use a custom collate function that assembles a batch as a *list of images* and a *list of target dictionaries*, rather than stacking targets into a fixed-size tensor. This mirrors common practice in detection pipelines and prevents padding/truncation artifacts.

Background class handling. In torchvision's Faster R-CNN implementation, class index 0 is reserved for `__background__`. Importantly, the dataset does *not* include explicit background annotations; instead, background is learned implicitly from region proposals (RPN) or RoIs that do not match any ground-truth box. However, the model head must still be parameterized with `num_classes = 1 + #foreground classes` so that it can assign low-confidence proposals to background during classification. In our case, we train with three classes total: `__background__` (0), `person-like` (1), and `person` (2). Accordingly, we map raw dataset labels into the 1-indexed foreground space (`person-like` $\rightarrow$ 1, `person` $\rightarrow$ 2) while leaving background implicit. Finally, we apply lightweight augmentations and normalization via torchvision transforms. We resize images to a consistent resolution for stable training dynamics and apply random horizontal flips to reduce overfitting and improve robustness to viewpoint variation.

3.2.3. Training Procedure. We fine-tune a pre-trained `fasterrcnn_resnet50_fpn` model and replace its RoI classification head to match the PnPLO label space. This isolates adaptation to the final prediction layers while

retaining transferable backbone and FPN features learned from large-scale pretraining. Training is performed for seven epochs using SGD with momentum and weight decay. Faster R-CNN optimizes a *multi-term loss* that includes RPN objectness classification, RPN box regression, RoI classification loss, and RoI box regression loss. This structure is important for robustness because even if localization is correct, misclassification at the RoI head still yields a false positive (e.g., predicting `person` on a mannequin). We use a step learning-rate scheduler to reduce the learning rate mid-training, stabilizing fine-tuning as the model transitions from learning broad detector behavior to refining decision boundaries between visually similar classes.

3.3. DETR

3.3.1. System Overview. To study whether global context and attention-based reasoning improve robustness to person-like nonhuman objects, we implement a transformer-based object detector using the DEtection TRansformer (DETR). DETR formulates object detection as a direct set prediction problem, in which the model outputs a fixed-size set of object predictions per image rather than relying on anchor boxes, region proposals, or non-maximum suppression. This design enables joint reasoning over object presence, class identity, and spatial localization through global self-attention [5].

We fine-tune a pretrained DETR model with a ResNet-50 backbone on the PnPLO dataset, adapting it to a binary detection task with two foreground classes: `person` and `person-like`. Pretrained backbone and transformer weights are obtained from the Hugging Face implementation, while dataset integration, training orchestration, and robustness-oriented evaluation are implemented by us. Our goal is not to modify the DETR architecture itself, but to analyze how its global attention and structured prediction behave under visual ambiguity.

3.3.2. Annotation Processing. DETR expects annotations in COCO-style format, all VOC-style bounding boxes are converted into the required representation and associated with image-level metadata. Images are loaded and normalized using the DETR image processor, which performs resizing and pixel normalization consistent with the pretrained model. During training, images are standardized to a fixed resolution to ensure stable optimization and predictable memory usage. This ensures DETR receives consistent supervision to the inherent ambiguity of the PnPLO dataset.

3.3.3. Set-Based Detection and the Hungarian Algorithm. A defining characteristic of DETR is its use of set-based detection via bipartite matching. For each image, the model produces a fixed number of object queries, each predicting a class label, confidence score, and bounding box. Because the number of predicted queries exceeds the number of ground-truth objects, DETR must determine which predictions correspond to real objects and which should be treated as background.

This assignment is solved using the Hungarian algorithm. During training, DETR computes a matching cost between every predicted query and every ground-truth object. The algorithm then finds the optimal one-to-one assignment that minimizes the total cost. Each ground-truth object is matched to at most one predicted query, and all unmatched queries are explicitly assigned to the background class. Only matched predictions contribute to object-specific losses, while unmatched predictions are trained to predict background. This structured supervision encourages DETR to produce a small, non-redundant set of detections and directly influences how missed detections and background assignments manifest during evaluation.

3.3.4. Training Procedure. We fine-tune DETR end-to-end using a standard detection training loop. Each image is processed by the convolutional backbone and transformer encoder–decoder, producing object query predictions that are matched to ground-truth objects using the Hungarian algorithm. Training is performed for twenty five epochs using the AdamW optimizer with a small learning rate, following best practices for

transformer-based models. We monitor total loss as well as individual loss components throughout training to verify stable convergence.

3.3.5. Inference. At inference time, DETR processes each image independently and produces a fixed set of predicted bounding boxes, class labels, and confidence scores. Many of these predictions correspond to background or low-confidence detections. We apply confidence-based filtering to retain predictions relevant for evaluation.

3.4. SAM2

3.4.1. System Overview. SAM2 is a segmentation model rather than a standalone detector-classifier. We implement a two-stage *Det–SAM2* pipeline that combines a lightweight object detector, YOLO [7], to propose bounding boxes and class labels, and SAM2 [14] to generate high-quality segmentation masks for those regions. The primary objective of this system is to study robustness in the presence of visually ambiguous `person-like` objects. In particular, we aim to understand how detection errors propagate into downstream segmentation, and whether high-quality masks can mitigate or expose false positives produced by the detector. Our central hypothesis is that decoupling semantic recognition (detection) from precise boundary estimation (segmentation) yields interpretable failure modes and improved qualitative outputs, even when classification errors persist.

3.4.2. Data Preparation for Detection. The PnPLO dataset is originally annotated in VOC-style XML with absolute pixel coordinates. To train a modern one-stage detector, we convert each annotation into YOLO format $(c, x_{\text{center}}, y_{\text{center}}, w, h)$, where coordinates are normalized to $[0,1]$ relative to image width and height. This conversion ensures compatibility with YOLO's training pipeline while preserving geometric consistency across models.

3.4.3. Object Detection with YOLO. We fine-tune a pre-trained YOLOv11-nano detector using transfer learning. YOLO is selected for its strong speed–accuracy trade-off and its ability to produce dense, single-stage predictions in cluttered scenes. The model is trained for twenty epochs with an input resolution of 640×640 and a batch size of 16. Pre-training on large-scale datasets provides strong low-level visual features, while fine-tuning adapts the detector to the specific appearance statistics of the PnPLO dataset.

During inference, YOLO outputs bounding boxes, confidence scores, and class predictions for each detected instance. To reduce the impact of spurious detections, we apply confidence thresholding before passing boxes to SAM2. This step is critical for robustness: false positives at the detection stage directly translate into unnecessary segmentation prompts, increasing the likelihood of hallucinated masks on background regions.

3.4.4. Segmentation with SAM2. For each input image, detector-predicted bounding boxes are used as prompts to SAM2. The image is embedded once using SAM2's image encoder, after which all bounding boxes are processed in a single forward pass. We disable multi-mask output to obtain a single best segmentation per detection, simplifying downstream analysis and visualization. SAM2 excels at refining spatial boundaries when provided with a reasonable prompt; however, it does not perform semantic verification. As a result, segmentation quality is tightly coupled to detection quality This behavior allows SAM2 to act as a diagnostic tool, exposing upstream detector errors rather than masking them. In contrast to end-to-end instance segmentation models, this decoupled design enables clearer attribution of failure modes. This separation is central to our analysis of false positives and robustness in the Det–SAM2 pipeline.

4. Results & Evaluation

4.1. HOG Model

4.1.1. Evaluation Overview. We evaluate the HOG + LinearSVC system using two complementary perspectives. First, we report threshold-independent metrics such as Precision-Recall curves, average precision, and ROC AUC. These metrics characterize the intrinsic discriminative power of the HOG feature representation and linear classifier without assuming any particular operating point. Second, we analyze confusion matrices, false positive rates, and qualitative examples at fixed decision thresholds, including both the default threshold and an optimized threshold chosen to emphasize robustness. These results describe how the system would behave in practice when deployed in a robotics or surveillance pipeline. In such settings, false positives are costly because they trigger unnecessary downstream processing and repeated alerts. Reporting both views allows us to distinguish limitations of threshold choice from limitations of the representation itself. All evaluations are performed on the held-out test set unless otherwise noted.

4.1.2. Qualitative Results. Visual inspection is particularly important for robustness analysis because it reveals why false positives and false negatives occur, rather than merely how often they occur.

Figure 1: ***True positives in the HOG+LinearSVC pipeline.*** **Canonical upright pedestrian poses with strong silhouette structure are correctly classified.**

The true positive examples demonstrate that HOG performs well on canonical pedestrian imagery of upright human poses, clear limb articulation, and strong contrast at body boundaries. These cases align well with HOG's design assumptions, as the descriptor effectively captures vertical edges and consistent gradient orientations associated with human silhouettes. Even when appearance varies in clothing or background, the underlying pose geometry is sufficient for correct classification.

Figure 2: ***False negatives in the HOG+LinearSVC pipeline.*** **Real pedestrians with non-canonical poses or occlusions are misclassified as nonhuman.**

The false negative examples show real people incorrectly classified as nonhuman. These cases often involve unusual poses, occlusions, or non-canonical appearances such as sitting, leaning, or partially visible bodies. Since HOG primarily encodes local gradient orientations, it struggles when the global upright silhouette assumption is violated, highlighting a limitation of the architecture.

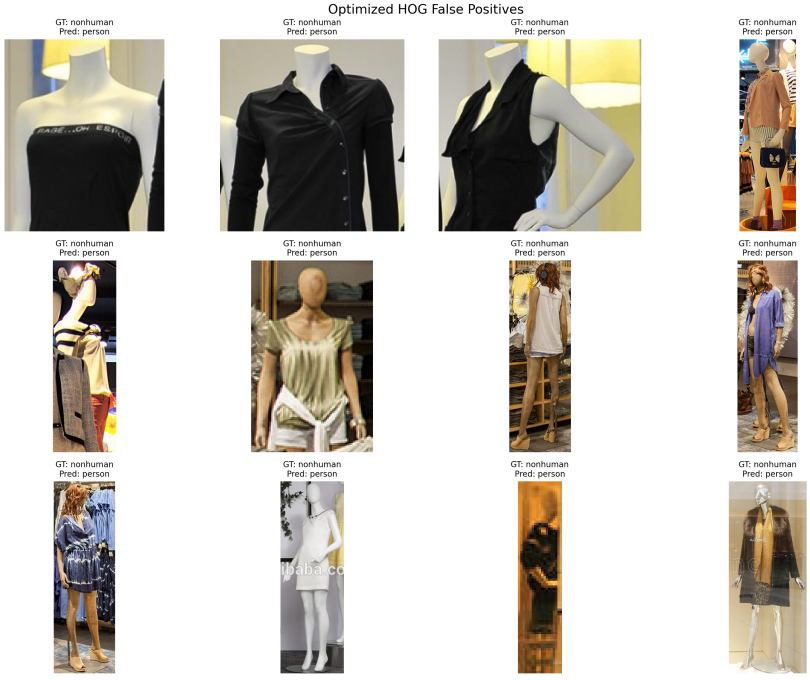


Figure 3: ***False positives in the HOG+LinearSVC pipeline.*** **Mannequins with strong human-like silhouettes are incorrectly classified as people.**

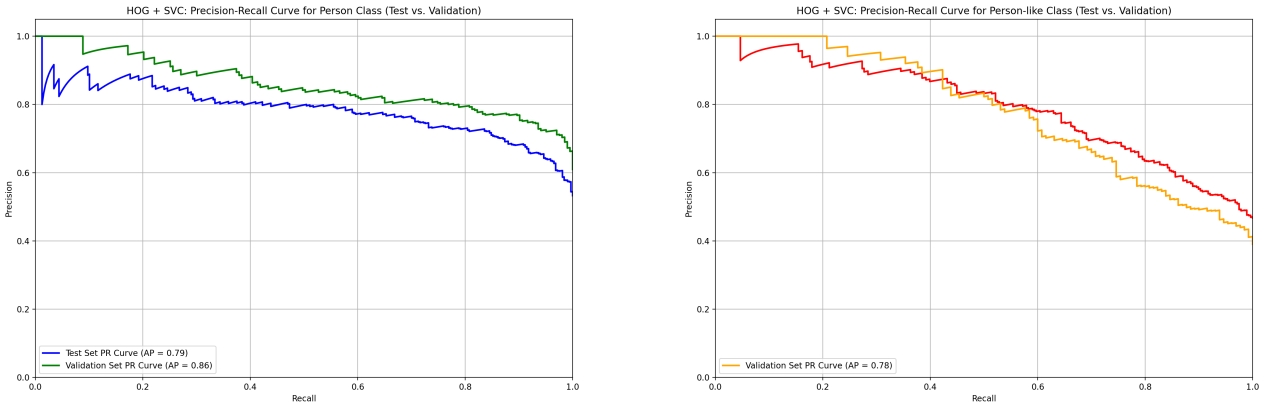
Conversely, false positives predominantly consist of mannequins with strong vertical edges and limb-like structures. These examples underscore HOG's core weakness in this task that it captures shape similarity but

lacks semantic understanding. From a robustness perspective, these false positives are especially costly, as they would repeatedly trigger human-specific downstream processing in long-running robotic systems despite the objects being static and nonhuman.

Figure 4: ***True negatives in the HOG+LinearSVC pipeline.*** **Mannequins with broken silhouettes or weak gradient structure are correctly rejected.**

Of note, all the false positives and true negatives examples shown above are mannequins, and this is a direct consequence of how the PnPLO dataset is constructed and what HOG actually encodes. In PnPLO, the nonhuman class is dominated by mannequins and statues that are intentionally chosen to be visually similar to humans. When we restrict our attention to false positives (nonhuman predicted as person) and true negatives (nonhuman predicted as nonhuman), we are effectively examining two subsets of the same object category: mannequins that HOG finds sufficiently human-like versus mannequins that deviate from its internal silhouette template.

4.1.3. Precision-Recall Analysis. For our robustness-focused evaluation, the PR curve reveals how precision degrades as recall increases, particularly due to false positives on nonhuman objects.



(a) *Precision-Recall curve for the* `person` *class.* (b) *Precision-Recall curve for the* `nonhuman` *class.*

Figure 5: ***Precision-Recall curves for the HOG+LinearSVC model on the PnPLO test set.***

For the person class, the PR curve shows relatively strong performance at low to moderate recall values. The behavior reflects HOG's effectiveness at identifying canonical pedestrian appearances that strongly match its learned gradient template. Average precision on the test set is approximately 0.79, with slightly higher performance on the validation set (0.86), suggesting limited generalization to new pose and appearance variations. The nonhuman class PR curve presents a more critical view of robustness. Precision drops rapidly as recall increases, indicating that many nonhuman objects receive high person scores and are difficult to reject confidently. Only the most visually distinct mannequins are consistently classified as nonhuman at high confidence.

4.1.4. False Positive Rate and Confusion Matrix. At the default decision threshold, accuracy is approximately 71%, with precision and recall around 0.73 for the person class and 0.69 for the nonhuman class. The false positive rate is approximately 30%, indicating that nearly one third of nonhuman objects are misclassified as people.

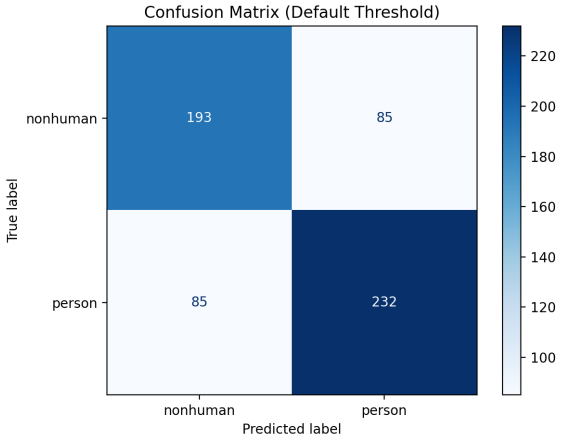


Figure 6: ***Confusion matrix for HOG+LinearSVC at the default threshold.***

At the optimized threshold, recall for the person class improves, but the false positive rate increases substantially to approximately 48%, demonstrating that threshold tuning shifts errors rather than resolving ambiguity.

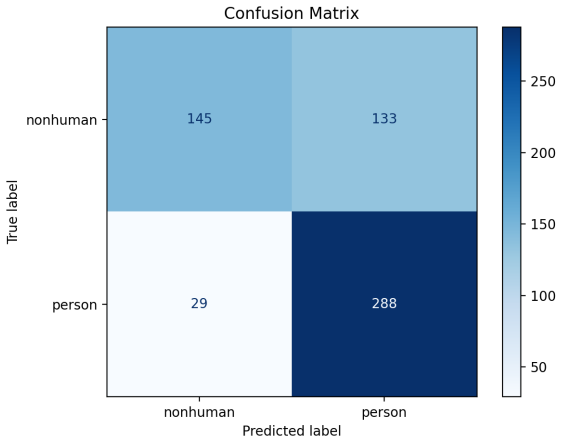


Figure 7: ***Confusion matrix for HOG+LinearSVC at the optimized robustness threshold.***

4.1.5. Predicted Probability Distributions. Finally, we analyze histograms of predicted person probabilities for true people and true nonhuman objects.

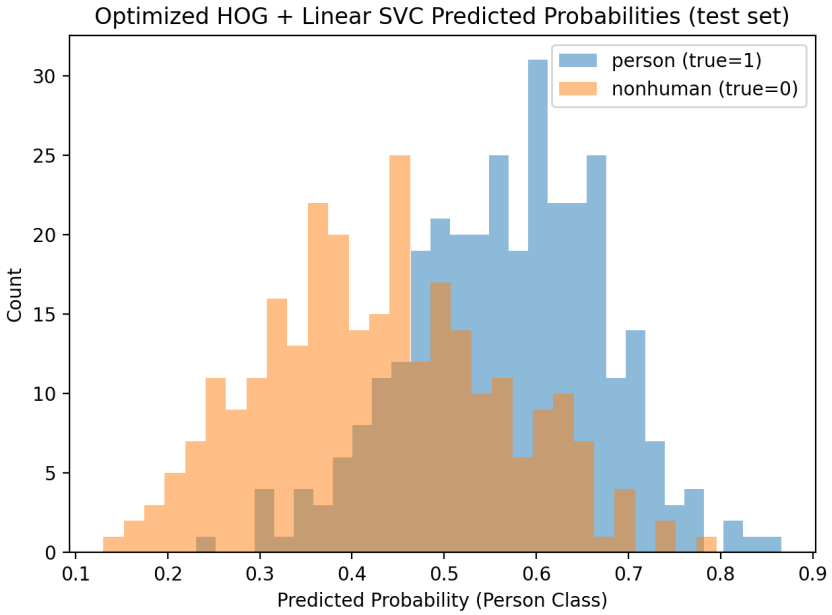


Figure 8: ***Predicted person probability distributions for true persons and nonhuman objects.***

These distributions show significant overlap, particularly in the mid-confidence range between approximately 0.4 and 0.6, explaining the smooth Precision-Recall tradeoffs and lack of a clean decision boundary.

4.1.6. Summary of HOG Performance. Overall, HOG + LinearSVC serves as an interpretable classical baseline but exhibits clear robustness limitations on the PnPLO dataset. Its strengths lie in detecting canonical pedestrian silhouettes, while its weaknesses stem from an inability to distinguish humans from realistic person-like objects. Errors are systematic, predictable, and driven by low-level gradient similarity. These findings motivate the use of more expressive models explored in subsequent sections.

Metric	Nonhuman	Person
Precision	0.69	0.73
Recall	0.69	0.73
F1-score	0.69	0.73
Number of Ground-Truth Samples	278	317
Overall Accuracy	0.71	
False Positive Rate (Nonhuman → Person)	0.3058	
ROC AUC (Test)	0.796	
Average Precision (PR AUC, Test)	0.787	

Table 1: ***Performance of the optimized HOG+LinearSVC model on the PnPLO test set.*** **Class-wise precision, recall, and F1-score are shown along with aggregate accuracy, robustness-focused false positive rate, and threshold-independent ranking metrics.**

4.2. Faster R-CNN

4.2.1. Evaluation Overview. Evaluation is performed on a held-out test set using an IoU threshold of 0.50 to define correct detections. To fairly balance precision and recall, we select the confidence threshold that maximizes the F1-score for the person class on validation data (optimal threshold = 0.81). We report precision, recall, F1-score, and Average Precision (AP@0.50), and analyze precision–recall curves to assess ranking quality across confidence levels.

4.2.2. Qualitative Analysis. Across diverse scenes, Faster R-CNN reliably detects upright, naturally posed humans with clean localization, even under varied lighting and background clutter (e.g., Images 0, 25, and 213). High-confidence true positives tend to correspond to canonical full-body or upper-body configurations, indicating strong reliance on global shape and part-based cues.

Figure 9: ***Qualitative Grid (GT vs. Predictions) Qualitative detection results for Faster R-CNN on the test set. Ground truth boxes are shown in purple (person) and brown (person-like), while predictions above the optimal confidence threshold (0.81) are shown in green (true positives), red (false positives), and orange dashed boxes (false negatives).***

Failure cases usually happen in the presence of cluttered images, crowded backgrounds, or fuzzy patches. In such scenarios, person-like distractors such as mannequins, statues, and human-shaped objects are frequently predicted as real people (Images 46, 63, 65, and 98). These objects share similar silhouettes, clothing textures, and limb arrangements, leading to confident false positives that are spatially well-aligned but semantically incorrect.

Crowding and partial occlusion further degrade performance, leading to missed detections or boxes that narrowly miss the IoU=0.50 threshold (Image 98). In ambiguous person-like scenes, near-correct boxes may fail the IoU=0.50 threshold, explaining precision drops at very high recall. Overall, the qualitative results indicate that Faster R-CNN is dependable for real pedestrians in typical conditions, with errors primarily driven by hard negatives and scene ambiguity rather than systematic detection failure.

4.2.3. Quantitative Results. At the selected operating point (IoU = 0.50, score threshold = 0.81), Faster R-CNN achieves a balanced precision–recall tradeoff, with precision = 0.876 and recall = 0.871. Concretely, this means that approximately 88% of predicted persons correspond to true humans, while 87% of all true humans are successfully detected. The resulting F1-score of 0.873 (TP = 276, FP = 39, FN = 41) confirms that neither false positives nor false negatives dominate error at this threshold.

Metric	**Value**
True Positives (TP)	276
False Positives (FP)	39
FP from person-like distractors	1
FP from other background regions	38
False Negatives (FN)	41
True Negatives (TN, approx.)	277
Precision	0.8762
Recall	0.8707
F1-score	0.8734

Table 2: ***Final evaluation metrics for the `person` class using Faster R-CNN at IoU = 0.50 and confidence threshold = 0.81 on the PnPLO test set.***

From a robustness perspective, the false positive count (FP = 39) is particularly informative. These errors represent cases where the detector confidently predicts a "person" in regions that do not correspond to a real human. Notably, only one of these false positives overlaps with a labeled person-like distractor, while the remaining false positives arise from background clutter or ambiguous regions. This indicates that, at the optimized threshold, Faster R-CNN is largely effective at suppressing high-confidence detections on explicitly annotated person-like objects, but still produces occasional spurious detections in visually complex scenes.

The Average Precision results reinforce this interpretation. AP@0.50 of 0.947 on the test set and 0.934 on validation indicate that Faster R-CNN ranks true person detections well above incorrect ones across a wide range of confidence thresholds. In practical terms, this means that most false positives are assigned lower confidence scores than true positives, enabling effective control of false alarms through thresholding.

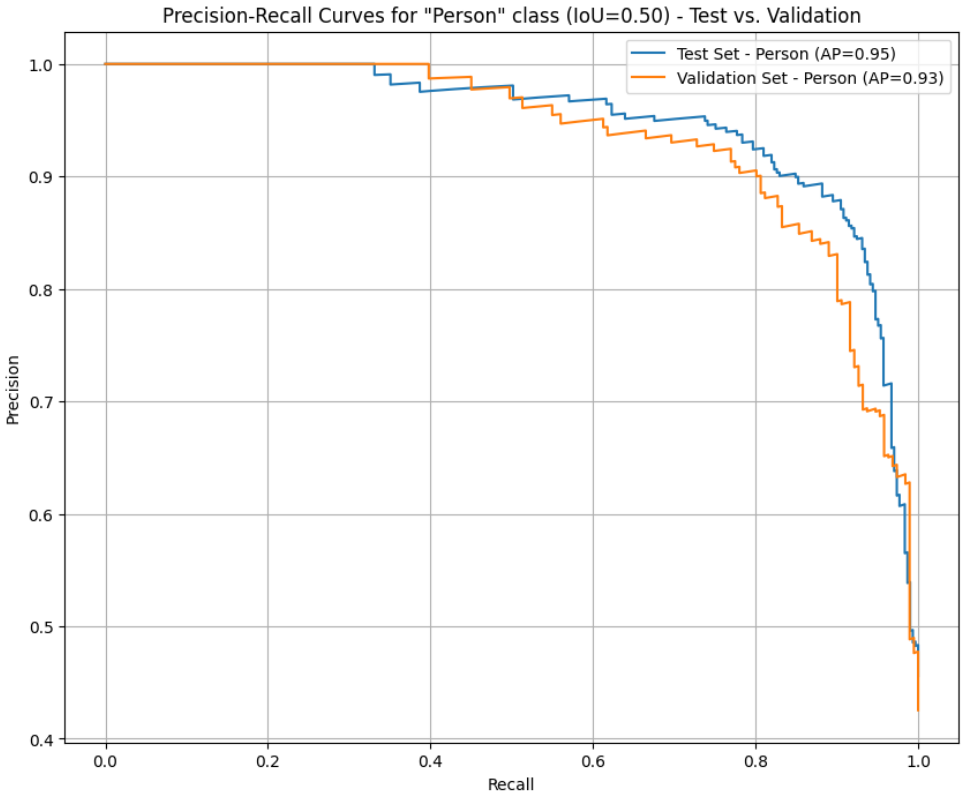


Figure 10: ***Precision–Recall Curve: Precision–recall curves for the "person" class at IoU = 0.50 on the test and validation sets.***

This trend is clearly visible in the Precision-Recall curves (Fig. 10). Precision remains high (≈0.90–0.98) across much of the recall range, indicating that confident predictions are usually correct. Overall, these results suggest that Faster R-CNN is robust in the sense that it rarely produces confident false positives on person-like distractors, and when errors occur, they are largely confined to lower-confidence predictions that can be filtered without severely degrading recall.

4.3. DETR

4.3.1. Evaluation Overview. We evaluate DETR using both standard object detection metrics and robustness-focused analyses tailored to the PnPLO dataset. Because DETR is a full detection model rather than a binary classifier, we report mean average precision across multiple IoU thresholds, per-class average precision, and class-specific precision, recall, and F1 scores at a fixed confidence threshold. In addition, we analyze normalized confusion matrices, predicted probability distributions, and qualitative detection examples to understand failure modes and robustness behavior.

4.3.2. Qualitative Prediction Analysis. We begin by analyzing qualitative DETR predictions on representative test images.

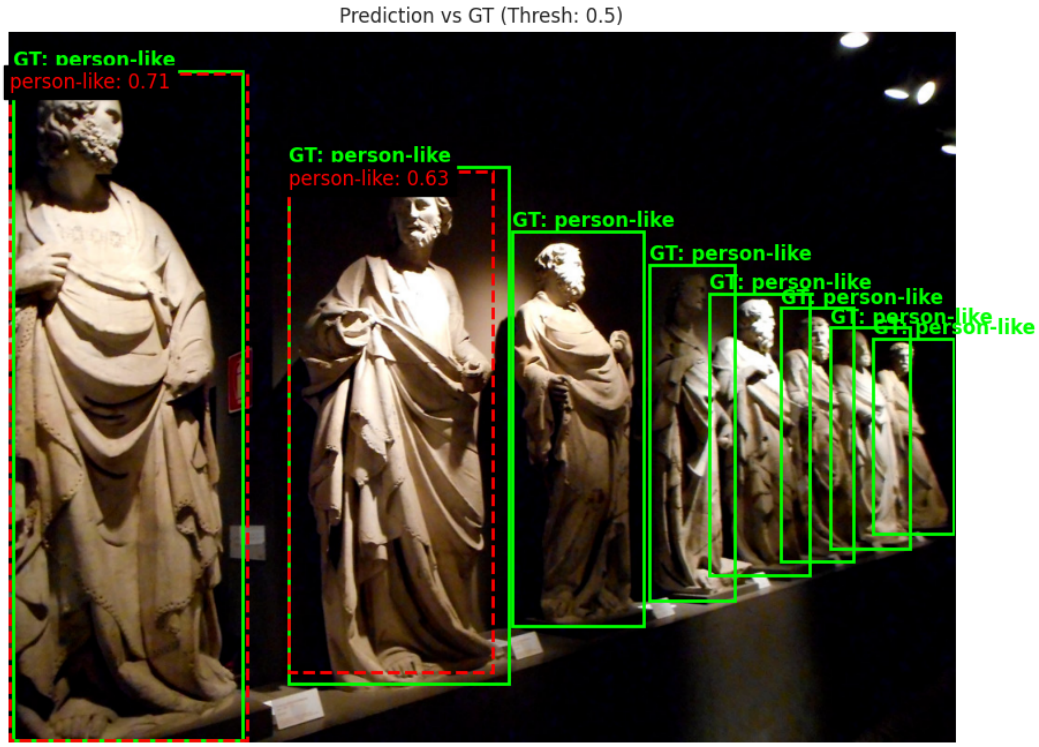
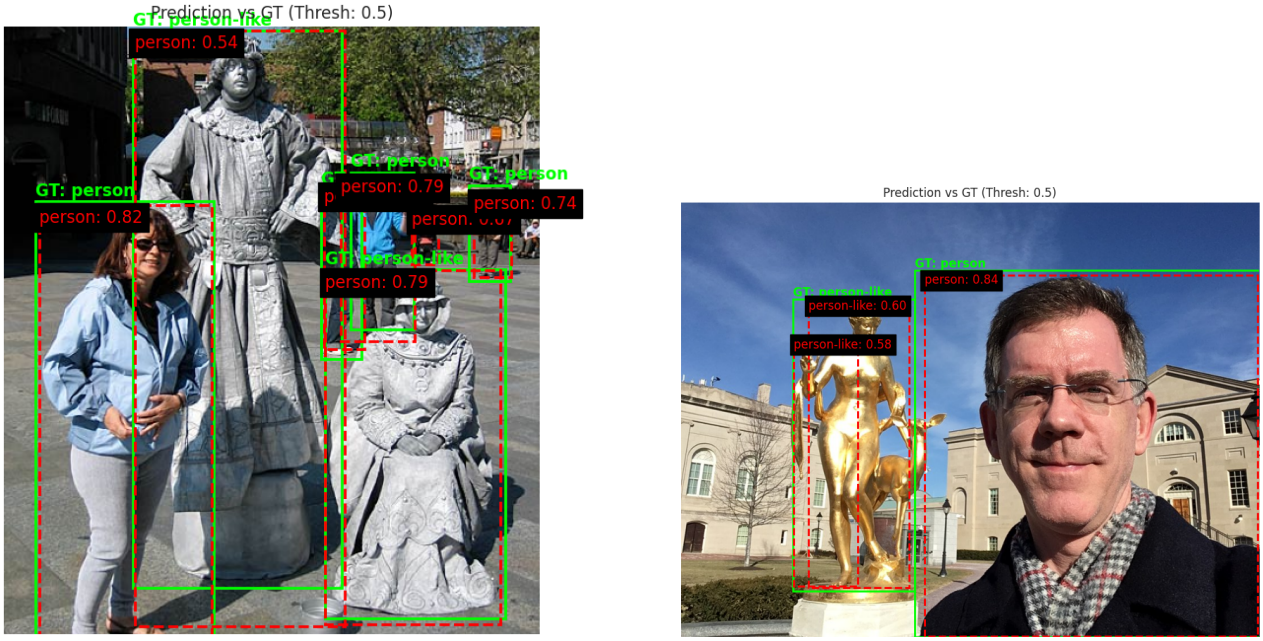


Figure 11: ***Representative qualitative DETR predictions on grouped statues.***

As show in Fig 11 with an image of multiple statues arranged in groups, DETR correctly identifies two statues as person-like class, but fails to detect majority of statues in the image. The model does produce correct prediction with good approximation of bounding boxes, but lacks to confidently predict multiple objects in the scene for the statue gallery example.



(a) (b)

Figure 12: ***Representative qualitative DETR predictions on mixed scenes containing both people and statues.***

In mixed scenes containing both people and statues, DETR generally succeeds at detecting real people with high confidence. In the image containing a person standing near a statue, the human subject is detected with a high person confidence, while the statue receives a slightly lower but still substantial person-like score. This behavior demonstrates that DETR incorporates global context and relative appearance compared to faster R-CNN model as it does a relatively good job handling mixed scenes.

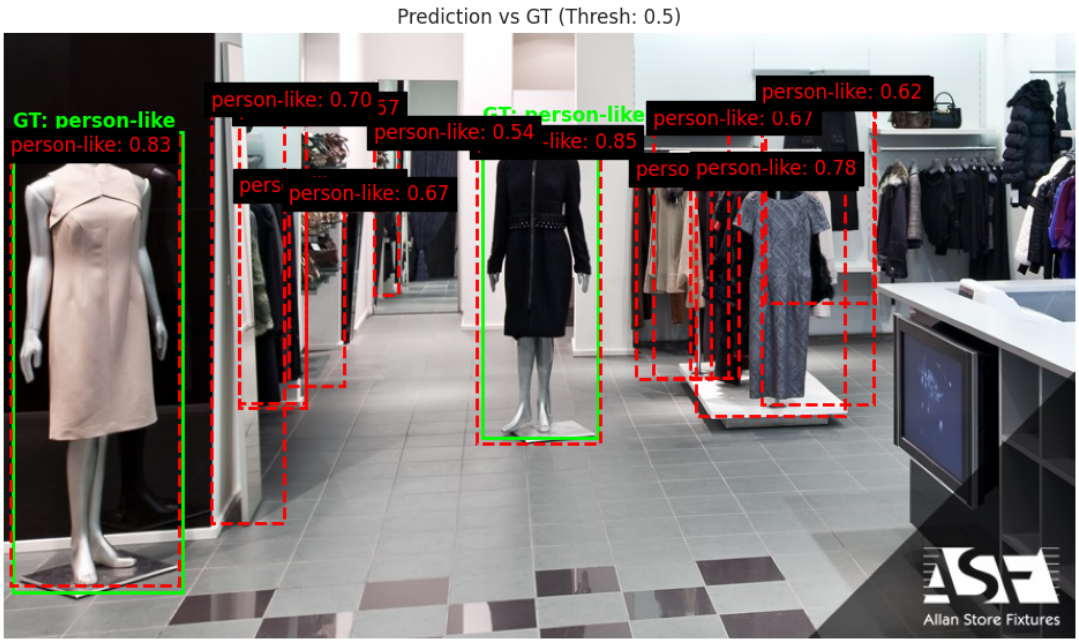


Figure 13: ***Representative qualitative DETR predictions on indoor retail scenes filled with mannequins.***

In indoor retail scene filled with mannequins, DETR produces a dense set of detections. Many mannequins are correctly labeled as person-like, but the model also predicts the background assignments as person-like with high confidence. This showcases that DETR struggles with background clutter image classification examples, its possible that this could have been produced by the lack of "background" class in the PnPLO dataset. Overall, semantic ambiguity persists for person-like objects, particularly statues and mannequins with realistic proportions and poses, but the model does performs relatively well in mixed scenes and classification of objects.

4.3.3. Confusion Matrix. The normalized confusion matrix for DETR includes three prediction categories: person, person-like, and background. Although the PnPLO dataset contains only two semantic classes, the background category arises from the object detection formulation used by DETR. Thus, background does not represent an additional semantic class, but rather the absence of a matched detection.

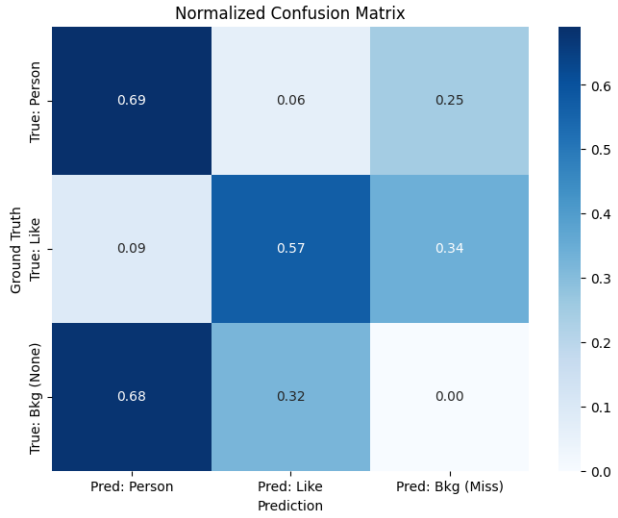


Figure 14: ***Normalized confusion matrix for DETR, including background.***

The confusion matrix shows, for true person-like objects, a substantial fraction are either predicted as person-like, predicted as person, or missed entirely and assigned to background. Misses typically occur when objects are visually repetitive, or surrounded by similar objects, such as mannequins arranged in rows. Similarly, for the person class, there's a substantial portion of misclassifications as background class. Although the PnPLO dataset contains only two semantic classes, the background category arises from the object detection formulation used by DETR. Thus, background does not represent an additional semantic class, but rather the absence of a matched detection. These incorrect classifications could be result of bad handling

4.3.4. Quantitative Analysis and Class-Specific Behavior Table 3 provides a detailed breakdown of DETR's per-class performance at a fixed confidence threshold of 0.5. Several trends emerge that clarify how DETR trades off precision, recall, and abstention through background assignments.

Metric	Person-like	Person
mAP (IoU = 0.50:0.95)	0.3055	0.2921
F1 Score	0.5918	0.6041
Precision	0.6172	0.5368
Recall	0.5683	0.6909
True Positives (TP)	158	219
False Positives (FP)	98	189
False Negatives (FN)	120	98

Table 3: *Per-class DETR evaluation metrics at confidence threshold 0.5.* Metrics are reported on the PnPLO test set using IoU thresholds from 0.50 to 0.95 for mAP.

For the `person` class, DETR achieves a recall of 0.6909, but this comes at the cost of lower precision (0.5368), with 189 false positives relative to 219 true positives. Many of these false positives correspond to statues or mannequins that closely resemble human form, as observed in the qualitative results. The resulting F1 score of 0.6041 shows that DETR prioritizes detecting people when possible, even if this admits a nontrivial number of ambiguous detections.

For the `person-like` class, DETR exhibits the opposite behavior. Precision is higher at 0.6172, while recall drops to 0.5683. This asymmetry indicates that DETR is more conservative when assigning the person-like label, frequently abstaining by assigning background instead. This behavior is reflected in the larger number of false negatives (FN = 120) relative to true positives (TP = 158). In practical terms, many person-like objects are not explicitly labeled as such, but are instead missed entirely. The near-equal F1 scores for `person` (0.6041) and `person-like` (0.5918) suggest that DETR does not strongly favor one class over the other, but rather balances its errors through a combination of misclassification and background suppression.

The mean average precision (mAP) values further reinforce this interpretation. The mAP scores for `person-like` (0.3055) and `person` (0.2921) are nearly identical, indicating similar ranking quality across classes. The relatively low absolute values reflect the difficulty of the task rather than poor localization. As shown qualitatively, DETR often produces well-aligned bounding boxes but struggles with semantic separation and consistent matching in cluttered scenes. Because mAP penalizes both misclassification and missed detections across multiple IoU thresholds, these results highlight the combined impact of background assignments and class ambiguity. We note that two mAP variants are reported for DETR. The per-class values in Table 3 correspond to COCO-style mAP averaged over IoU thresholds from 0.50 to 0.95, which is a stringent metric that penalizes both localization errors and missed detections. For cross-model comparison presented later in Table 5, we additionally report mAP@0.50, which measures detection performance at a single, more permissive IoU threshold. As expected, mAP@0.50 (0.4913) is substantially higher than mAP@0.50:0.95 (0.30), reflecting DETR's ability to produce reasonably aligned detections while still struggling with semantic ambiguity and matching consistency in crowded scenes.

4.3.5. Summary of DETR Performance. Overall, DETR's quantitative behavior reveals a structured abstention strategy rather than indiscriminate false positives. Compared to HOG and Faster R-CNN, DETR more frequently chooses to suppress detections as background instead of forcing a potentially incorrect class decision. While this reduces recall for person-like objects, it limits the most costly failure mode in robotics and surveillance systems: confident false positive person detections. This behavior is particularly valuable for mitigating false positives in safety-critical contexts, even though it results in lower overall recall when faced with the inherent semantic

similarities between humans and high-fidelity person-like objects. Nevertheless, the remaining false positives on highly realistic statues and mannequins indicate that global attention alone is insufficient to fully resolve person-like ambiguity.

DETR does not fully solve the robustness challenge posed by person-like objects. Statues and mannequins that closely resemble humans continue to generate confident detections, and dense scenes introduce matching and recall limitations. These results suggest that while transformer-based detection is a strong step forward, additional semantic reasoning or multimodal cues may be necessary to achieve reliable robustness in real-world robotic and surveillance applications.

4.4. SAM2

4.4.1. Evaluation Overview. The evaluation of the Det-SAM2 pipeline was conducted on the held-out test dataset. We matched predicted boxes to ground-truth boxes using an IoU threshold of 0.5 and report mean Average Precision (mAP@0.5), per-class AP, Precision-Recall curves, and confusion matrix including background.

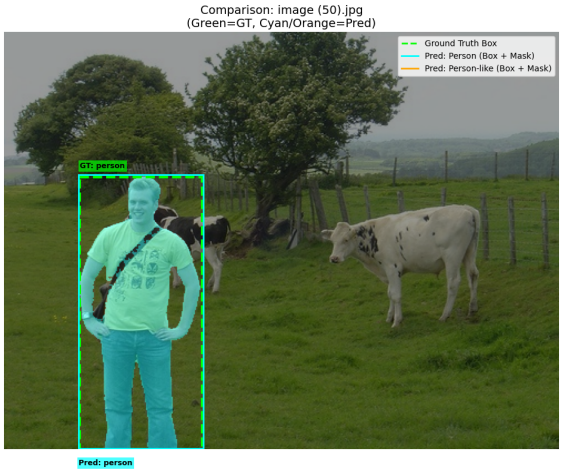
4.4.2. Qualitative Analysis. Figure 15 highlights representative qualitative behaviors of the Det–SAM2 pipeline across a range of scenes. Visual inspection confirms that SAM2 produces clean, high-resolution masks when the detector's bounding box is accurate. Failure cases typically originate upstream; for instance, if YOLO proposes a slightly misaligned or spurious box, SAM2 faithfully segments the prompted region even when it does not correspond to a true object. One notable false-positive pattern observed in our dataset is that the detector occasionally classified horses as humans, even in images where actual humans were present. Overall, SAM2 improves spatial accuracy but does not correct semantic errors introduced by the detector.

Fig. 15a, YOLO correctly identifies the foreground human as `person` and produces a well-aligned bounding box. Given this accurate prompt, SAM2 generates a clean segmentation mask that closely follows the human silhouette with minimal background leakage, demonstrating the pipeline's best-case performance when detection is reliable.

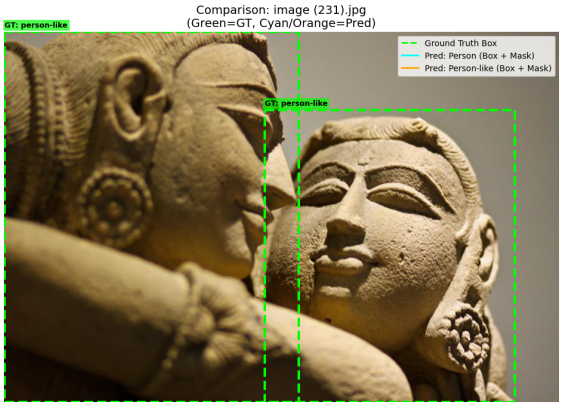
Fig. 15b shows a case involving rigid `person-like` statues. The detector does not predict the `person-like` class with high confidence, and SAM2 does not produce any segmentation. This illustrates the objects for which the pipeline fails to classify correctly.

Fig. 15c depicts a challenging indoor scene containing mannequins, where the model fails to correctly distinguish multiple person-like objects from a single true person. Notably, a mannequin posed in a running position is misclassified as a human, suggesting that dynamic, human-like postures could bias the classifier toward the person class. This failure mode indicates that while SAM2 improves spatial mask precision, it does not mitigate semantic errors introduced by upstream classification, particularly in cases where pose cues dominate over material or contextual evidence.

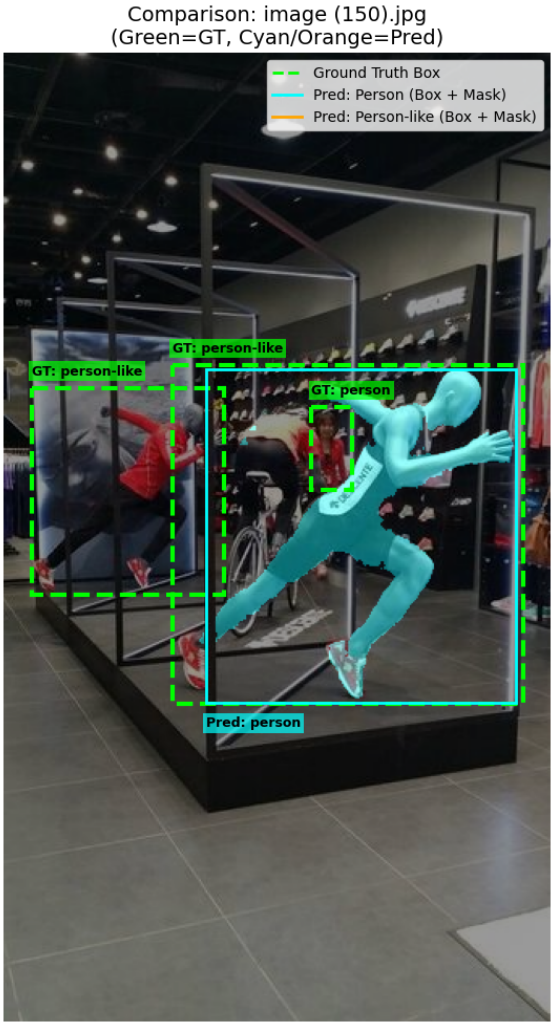
Finally, Fig. 15d shows a mixed scene containing both a real person and a visually similar person-like object. The detector assigns distinct class labels to each instance, and SAM2 produces corresponding masks, illustrating correct classification which could be driven by subtle texture or material cues.



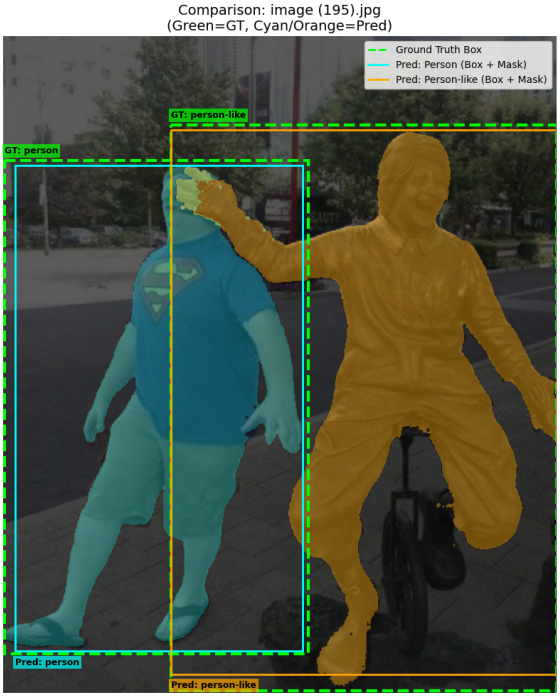
(a) Correct person detection; clean mask.



(b) Person-like statues; no good prediction.



(c) Challenging indoor mannequin/person-like overlap.



(d) Mixed scene; person vs. person-like separation.

Figure 15: ***Qualitative Det–SAM2 results on PnPLO.*** **Each row shows representative test images illustrating segmentation masks produced by SAM2 when prompted by YOLO detections.**

4.4.3. Quantitative Results. Across both classes, the detector achieves mAP@0.5 = 0.848, with mean precision = 0.788 and mean recall = 0.788 on the test set. Performance is slightly higher on the `person-like` class (AP@0.5 = 0.8603) than on the `person` class (AP@0.5 = 0.8363), which is consistent with the dataset's "person vs. person-like" distribution where many distractors are visually human-shaped, making them more likely to be scored confidently.

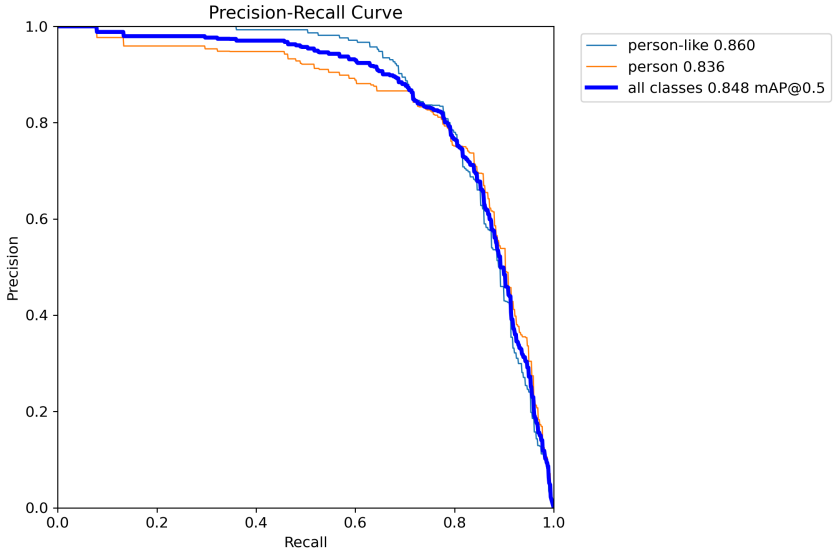


Figure 16: ***Precision-Recall curves for the `person-like` class, the `person` class, and the mean across classes at IoU*** = 0.5.

The Precision-Recall curves show high precision at low to moderate recall, followed by a sharp decline beyond approximately 0.8 recall, indicating a surge in false positives when recall is aggressively optimized. This suggests that most errors stem from ambiguous background regions and borderline person-like objects rather than systematic mislabeling.

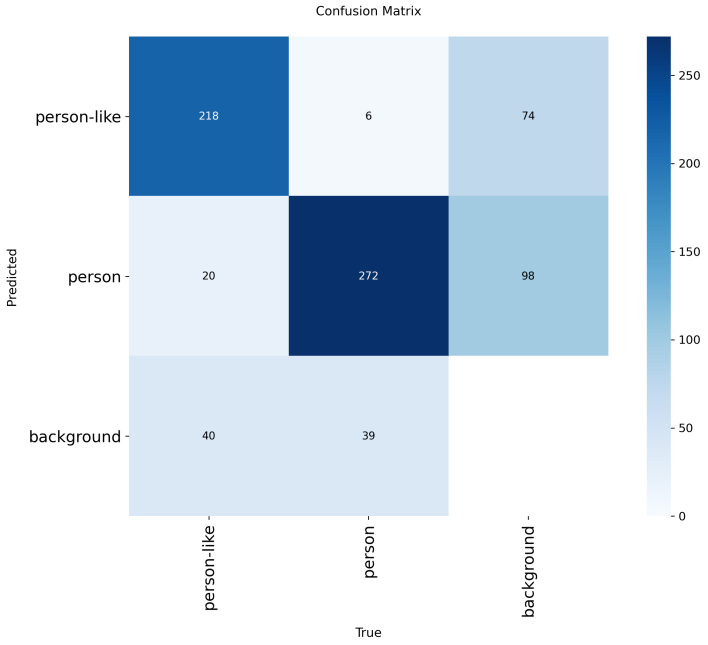


Figure 17: ***Confusion matrix including background.*** **Confusion matrix for the YOLO detector evaluated on test set.**

Strong diagonal entries indicate correct classification of foreground objects, while off-diagonal entries reveal cross-class confusion and ghost detections on background regions. Here, ghost detections refer to false positive predictions that arise from visually salient background regions where no annotated object exists, rather than from misclassifications between foreground classes. Cross-class confusion is relatively small (e.g., `person-like` $\rightarrow$ `person`), indicating that pure label flipping is not the dominant failure mode. More significant are ghost detections on background regions (e.g., `background` $\rightarrow$ `person`), highlighting a key limitation of one-stage detectors: visually salient background structures can trigger confident predictions even when no object is present.

Metric	**Person**	**Person-like**
True Positives (TP)	272	218
Total False Positives (FP)	45	60
Misclassification from other class	6 (statues)	20 (real people)
Ghost detections (background)	39	40
False Discovery Rate (FDR)	14.20%	–
Precision	85.80%	–

Table 4: ***False positive breakdown for the SAM2 pipeline on the PnPLO test set.*** **Ghost detections correspond to predictions on background regions without ground-truth objects, while misclassification errors arise from confusion between `person` and `person-like` classes.**

A closer examination of false positive behavior reveals that SAM2's errors are dominated by *missed detections* rather than semantic confusion between foreground classes. For the `person` class, the model produces 45 false positives, of which 39 (86.7%) originate from empty background regions rather than misclassification of person-like objects. Only 6 false positives arise from statues being mislabeled as people. This results in a false discovery rate of 14.2%, indicating that when SAM2 predicts a person, it is correct in the large majority of cases.

A similar pattern is observed for the `person-like` class, where 60 false positives occur, with 40 (66.7%) corresponding to background detections and only 20 arising from confusion with real people. These results indicate that SAM2 exhibits relatively strong semantic separation between `person` and `person-like` objects, and that its false positive rate is driven primarily by detector over-sensitivity to visually salient background regions rather than systematic misclassification of human-shaped distractors. From a robustness perspective, this behavior is preferable to HOG-style silhouette confusion, as background ghost detections can often be mitigated through temporal filtering or downstream validation, whereas persistent misclassification of static person-like objects is more difficult to suppress.

4.5. Summary of Comparative Analysis

Model	Type	F1 Score	False Positive Rate	mAP
HOG + SVM	Classifier	0.748	0.306	0.7904
Faster R-CNN	Detector	0.873	0.124	0.7010
DETR	Detector	0.604	0.063	0.4913
SAM2	Detector	0.773	0.051	0.8483

Table 5: ***Comparison of pedestrian detection models on the PnPLO test set.*** **F1 scores are reported at the selected operating point for each model. False positive rate measures the fraction of nonhuman objects incorrectly classified as persons (lower is better). Mean Average Precision (mAP) is reported at IoU $= 0.50$ for detector-based models.**

5. Conclusion

This study evaluated the robustness of four distinct computer vision paradigms, HOG, Faster R-CNN, DETR, and YOLO+SAM2, specifically focusing on their ability to distinguish real pedestrians from person-like distractors in the PnPLO dataset. Our results demonstrate that while modern deep learning architectures significantly outperform classical gradient-based methods, robustness to visual ambiguity is not a direct byproduct of model complexity, but rather a result of specific architectural priors. Our benchmarking revealed a clear hierarchy of performance and failure modes:

- **HOG+SVM:** Proved inadequate for robust deployment, with a high False Positive Rate (30.6%) driven by a reliance on local gradient silhouettes that cannot distinguish between the geometry of a statue and a human.
- **Faster R-CNN:** Provided a strong balance of precision and recall ($F_1 = 0.873$), indicating that region-based semantic features are effective at filtering out background clutter, though it remains susceptible to "person-like" objects with high-quality textures.
- **DETR:** Exhibited a "cautious" robustness. While it had the lowest *mAP* (0.49) due to matching difficulties in crowded scenes, it achieved a remarkably low False Positive Rate (6.3%), suggesting that global self-attention allows the model to "abstain" from making overconfident predictions under high ambiguity.
- **SAM2 (YOLO-driven):** Emerged as the most precise tool for localization (*mAP* 0.84). However, it revealed a "semantic bottleneck": while SAM2 generates near-perfect geometric masks, its robustness is entirely dependent on the upstream detector, which occasionally misclassified objects like detailed statues or running mannequins as humans.

6. Limitations

Several practical constraints impacted the scope and depth of this benchmarking study. First, the lack of a dedicated `background` class in our primary dataset forced models to rely on implicit negative sampling, which likely contributed to the elevated false positive rates in the CNN, SAM2, and DETR models. Second, computational limitations significantly restricted the training phase; due to the resource quotas of a Google Colab Student account, both Faster R-CNN and DETR could only be trained for a limited number of epochs. This constraint likely prevented the models from reaching full convergence, potentially masking their peak performance capabilities. Finally, the inherent complexity of these architectures made fine-tuning on our specific dataset a significant hurdle, as the high parameter count often led to inadaquete training given our limited dataset size.

7. Challenges

The primary challenge encountered during development was the preprocessing and standardization of the dataset. We faced significant data inconsistencies regarding annotation formats (varying between COCO, JSON, and Pascal VOC XML), which necessitated the development of custom scripts to ensure a uniform pipeline. Furthermore, getting the dataset environment correctly configured was a non-trivial task; the models expected highly specific input dimensions and normalization constants that varied across the different frameworks. Managing these disparate input requirements while maintaining a consistent evaluation metric across the hierarchy of models proved to be the most time-consuming aspect of the implementation.

8. Future Work

While this study focuses primarily on robustness to false positives from person-like objects, a natural next step is a deeper analysis of false negatives, particularly missed detections of real pedestrians. In safety-critical applications such as robotics and surveillance, false negatives can have severe consequences, and understanding the conditions under which modern detectors fail to detect people is essential for reliable deployment. Future work would include a systematic breakdown of false negatives by pose, occlusion, scale, and scene context to complement the false-positive-focused analysis presented here. Expanding the PnPLO dataset would also enable more comprehensive robustness evaluation, which could include targeted data augmentation strategies such as viewpoint variation, lighting changes, and synthetic insertion of person-like objects to better stress-test model generalization. Finally, a more granular exploration of confidence thresholding and calibration methods, including class-specific thresholds and uncertainty-aware decision rules, could provide deeper insight into how operating points affect downstream system behavior. Together, these extensions would help bridge the gap between benchmark performance and reliable real-world deployment.

9. Acknowledgments.

COS 429 Support. We would like to express our sincere gratitude to the course instructors and TAs for their guidance, feedback, and support throughout the duration of this project!

Code Availability. The full implementation, evaluation scripts, and reproducibility instructions for this project are publicly available at: https://github.com/s-bhatia1216/pedestrian-detection.git.

Generative AI Usage. Lastly, we acknowledge the use of generative AI for debugging at instances throughout the data pre-processing, training, and evaluation pipelines of the four model architectures. It also served as a learning tool to better understand the conceptual mechanics of the models evaluated in this study. We also used generative AI to resolve LaTeX syntax configuration bugs while setting up our report.

References

[1] A. Alotaibi, H. Alatawi, A. Binnouh, L. Duwayriat, T. Alhmiedat, and O. M. Alia, "Deep learning-based vision systems for robot semantic navigation: An experimental study," *Technologies*, vol. 12, no. 9, p. 157, 2024. [Online]. Available: https://doi.org/10.3390/technologies12090157

[2] S. Boddu and A. Mukherjee, "Lightweight object detection using quantized YOLOv4-Tiny for emergency response in aerial imagery," arXiv preprint, 2025. [Online]. Available: https://arxiv.org/abs/2506.09299

[3] M. M. Bouchene, "Bayesian optimization of histogram of oriented gradients (hog) parameters for facial recognition," *The Journal of Supercomputing*, vol. 80, no. 14, pp. 20 118–20 149, 2024.

[4] H. Bristow and S. Lucey, "Why do linear SVMs trained on HOG features perform so well?" arXiv preprint, 2014. [Online]. Available: https://arxiv.org/abs/1406.2419

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://research.facebook.com/file/359707902426148/End-to-End-Object-Detection-with-Transformers.pdf

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[7] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[8] N. J. Karthika and S. Chandran, "Addressing the false positives in pedestrian detection," in *Lecture Notes in Electrical Engineering*, 2020, pp. 1083–1092.

[9] E. Kee, J. J. Chong, Z. J. Choong, and M. Lau, "Low-cost and sustainable pick and place solution by machine vision assistance," in *Proceedings of the IEEE Conference*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9997663

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick, "Segment anything," arXiv preprint, 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[11] B. Lafreniere, T. R. Jonker, S. Santosa, M. Parent, M. Glueck, T. Grossman, and D. Wigdor, "False positives vs. false negatives: The effects of recovery time and cognitive costs on input error preference," in *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2021, pp. 54–68.

[12] Y. Park, C. Sobolewski, and N. Azizan, "Identifying reliable predictions in detection transformers," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2412.01782v1

[13] N. Picello, F. Herrero, S. Hernández, A. López, and A. Santamaria-Navarro, "Leveraging pedestrian detection and tracking in robotics navigation: A survey with practical illustrations," *IEEE Access*, vol. 13, pp. 158 926–158 937, 2025.

[14] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," arXiv preprint, 2015. [Online]. Available: https://arxiv.org/abs/1506.01497

[16] Scikit-learn Developers, "1.4. support vector machines," Scikit-learn documentation, n.d., accessed 2025. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html

[17] N. Souli, P. Katzanis, P. Kardaras, Y. Grigoriou, S. G. Stavrinides, P. Kolios, and G. Ellinas, "An onboard uav multi-task system for trajectory prediction and state estimation employing transformer- and reservoir-based networks," *ACM J. Auton. Transport. Syst.*, vol. 2, no. 4, Jun. 2025. [Online]. Available: https://doi.org/10.1145/3725893