Open-Vocabulary vs Real-Time Object Detection: A Comparative Benchmark Study

Jahnvi Paliwal
Email: paliwaljnv08@gmail.com

***Abstract*—This study presents a controlled benchmark comparison between OWL-ViT, an open-vocabulary vision-language transformer, and YOLOv8, a real-time convolutional object detector. The evaluation measures inference latency, frame rate performance (FPS), detection frequency, confidence distribution, prompt sensitivity, and statistical comparison. Results show that YOLOv8 achieves approximately 14× higher throughput, while OWL-ViT demonstrates semantic flexibility but strong sensitivity to prompt-scene alignment. The findings highlight the computational and stability trade-offs inherent in open-vocabulary detection systems.**

I. Introduction

Object detection has progressed from fixed-category convolutional architectures to transformer-based vision-language models capable of open-vocabulary inference. Traditional detectors such as YOLOv8 operate on predefined class spaces and are optimized for real-time deployment. In contrast, OWL-ViT enables zero-shot object detection through dynamic textual prompts.

This work quantitatively analyzes the trade-off between efficiency and semantic flexibility.

II. Model Architectures

A. *YOLOv8 Architecture*

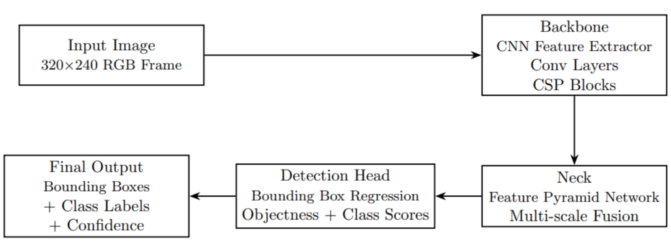

Fig. 1: **YOLOv8 Architecture Diagram.** Replace this with your CNN pipeline diagram (Backbone → Neck → Head).

YOLOv8 follows a single-stage detection paradigm consisting of:

- Convolutional backbone for feature extraction
- Feature pyramid aggregation
- Detection head for bounding box regression and classification

B. *OWL-ViT Architecture*

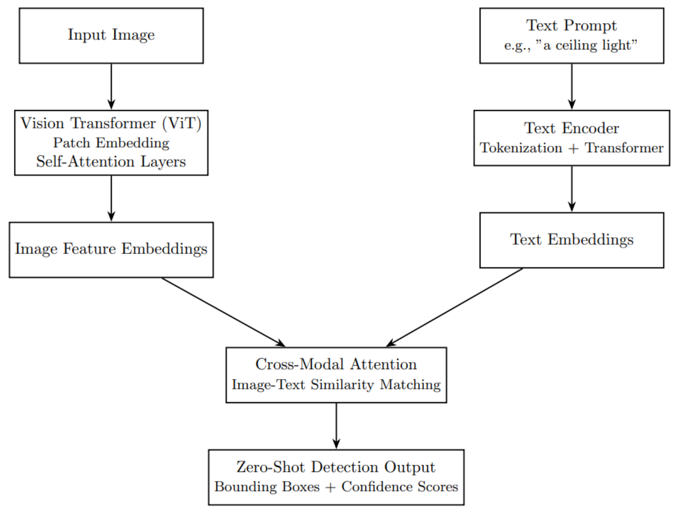

Fig. 2: **OWL-ViT Architecture Diagram.** Replace with Vision Transformer + Text Encoder + Cross-Attention diagram.

OWL-ViT integrates:

- Vision Transformer backbone
- Text encoder
- Cross-modal attention mechanism

III. Methodology

A. *Experimental Setup*

- Resolution: 320×240
- Frame Skip: 5
- Confidence Threshold: 0.25
- Hardware: Google Colab GPU

B. *Evaluation Metrics*

- Average inference time
- Frames per second (FPS)
- Detection count per class
- Confidence distribution
- Statistical significance (t-test)

IV. Performance Results

TABLE I: **Inference Performance Comparison**

Metric	OWL-ViT	YOLOv8
Avg Inference Time (s)	0.2268	0.0158
FPS	4.41	63.12

YOLOv8 achieved approximately **14.35× higher throughput**.

V. Detection Coverage Analysis

TABLE II: **Detection Count Per Class**

Label	Model	Count
a ceiling light	OWL	1
book	YOLO	3
bottle	YOLO	6
bowl	YOLO	13
cat	YOLO	8
cup	YOLO	25
dining table	YOLO	6
knife	YOLO	1
laptop	YOLO	1
person	YOLO	45
potted plant	YOLO	97
refrigerator	YOLO	1
vase	YOLO	31

A. *Detection Count Visualization*

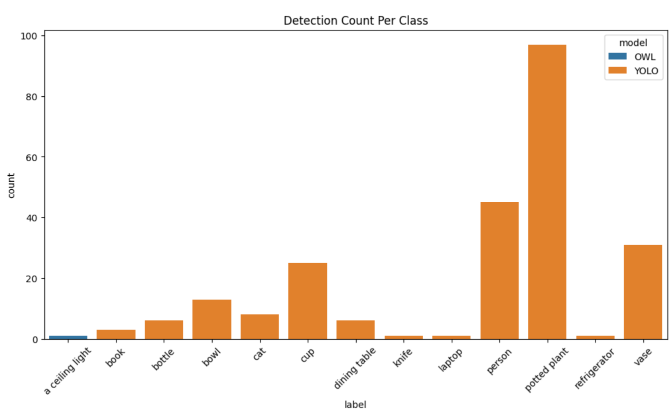

Fig. 3: **Detection Count Per Class (OWL-ViT vs YOLOv8)**

VI. Confidence Distribution Analysis

TABLE III: **Confidence Score Statistics**

Model	Count	Mean	Std	Max
OWL-ViT	1	0.2596	–	0.2596
YOLOv8	237	0.4795	0.1839	0.8766

Two-sample t-test result:

$$p = 0.234$$

A. *Confidence Distribution Visualization*

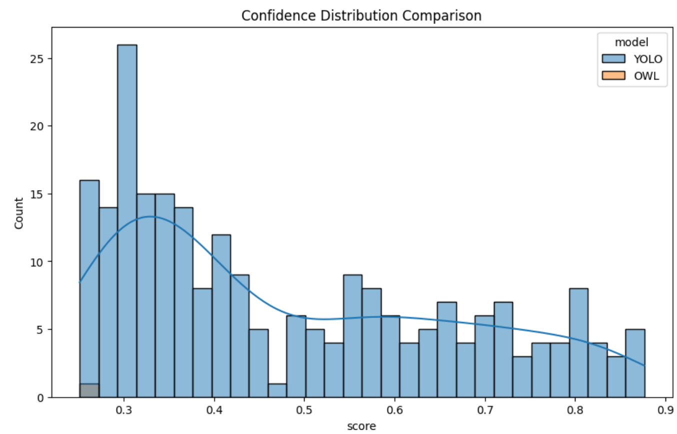

Fig. 4: **Confidence Distribution Comparison**

VII. Discussion

The experiment reveals a clear trade-off:

- **YOLOv8**: High-speed, robust, production-ready detection.
- **OWL-ViT**: Semantic flexibility, zero-shot capability, computationally intensive.

OWL-ViT's limited detections may result from resolution constraints, prompt phrasing, and scene alignment.

VIII. Conclusion

This benchmark confirms:

- YOLOv8 achieves real-time performance (63 FPS).
- OWL-ViT operates at 4 FPS.
- Open-vocabulary detection introduces flexibility at significant computational cost.
- Statistical confidence difference was not significant, but detection imbalance limits inference strength.

Future work should include ground-truth annotation and multi-scene evaluation for mAP-based benchmarking.