Object Detection Task

1 Zero-shot Object Detection

In this section, the accuracy of three pretrained models(Grounding DINO Tiny (VLM), YOLOv8s, and YOLOv8n) is analyzed and compared on random frames from the test video. This evaluation highlights their relative performance and detection capabilities in zero-shot object detection.

1.1 Vision Language Models(VLMs)

This part demonstrates zero-shot object detection using a vision language model, Grounding DINO Tiny, which integrates transformer-based detection with grounded pretraining on large-scale image–text pairs.

Figure 1: Output of the Grounding DINO model using the prompt ”person.”

As illustrated in the figure 1, the VLM effectively detects and labels people, showing significantly superior performance in the zero shot setting compared to the YOLO models that were specifically trained on human detection datasets.

1.2 YOLO Models

The accuracy of the YOLO models with pretrained weights is illustrated in Figures 2 and 3.

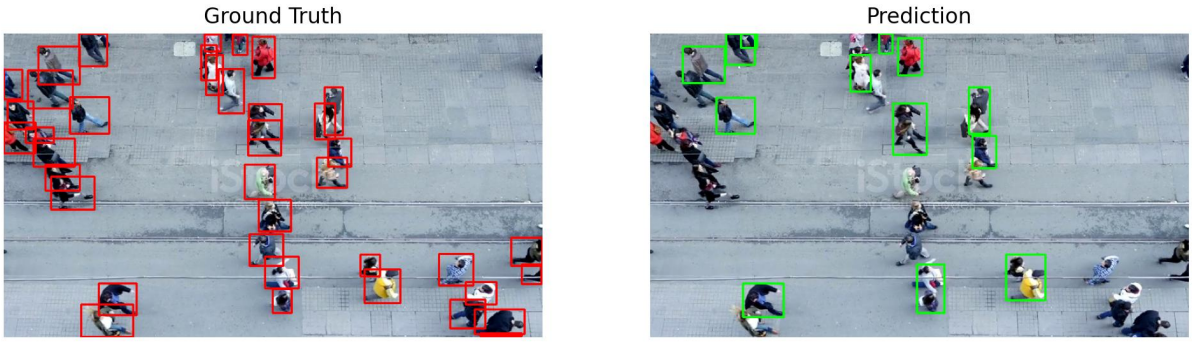

Figure 2: Output of the YOLOv8n model, showing that the pretrained YOLO fails to detect all persons in the frame.

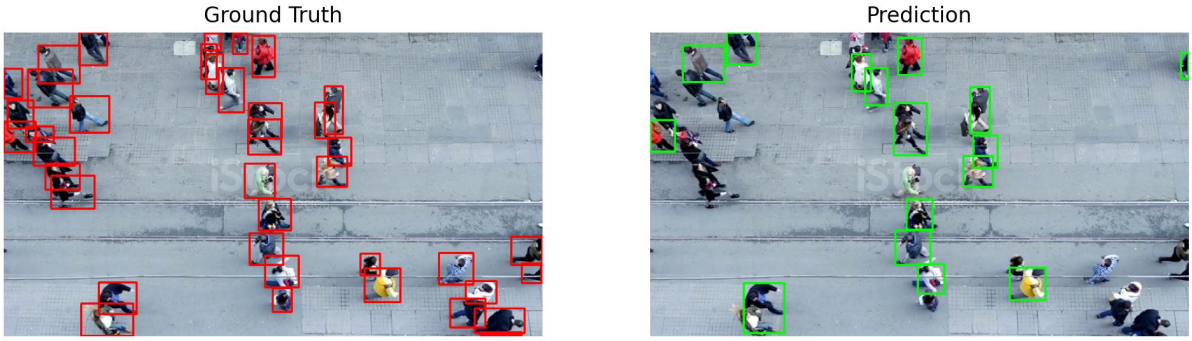

Figure 3: Output of the pretrained yolov8s model.

Model	Precision	Recall	F1	Time (ms)
VLM (Grounding-DINO)	0.844	**0.948**	**0.893**	212
YOLOv8n	0.889	0.413	0.565	**18**
YOLOv8s	**0.964**	0.543	0.691	23

Table 1: Comparison of object detection performance between Grounding-DINO (VLM), YOLOv8n, and YOLOv8s. Precision, recall, and F1 scores are computed over four randomly selected frames thatlabeled manually using LabelImg. The reported time corresponds to the average inference latency per image (in milliseconds).

Although the VLM model achieves higher accuracy, Table 1 shows that its inference time is significantly longer than that of the YOLO models, making it unsuitable and inefficient for real time applications.

2 Custom Approach

As shown in Table 1, VLMs achieve high accuracy but require a considerable amount of time for inference. in contrast, YOLO models perform inference much faster. Therefore, a knowledge distillation approach is employed to finetune the YOLO model to acquire knowledge similar to that of the VLM while maintaining high inference speed.

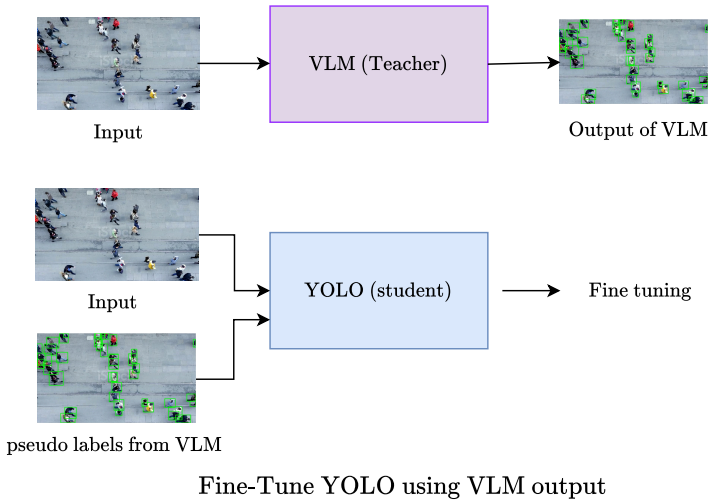

Figure 4: Teacher–student framework showing knowledge distillation from VLM to YOLO.

For the knowledge distillation process, a set of 175 unlabeled aerial images was used. The

VLM model automatically generated the pseudo-labels for these images. Figure 5 shows three examples of the images along with the labels produced by the VLM.

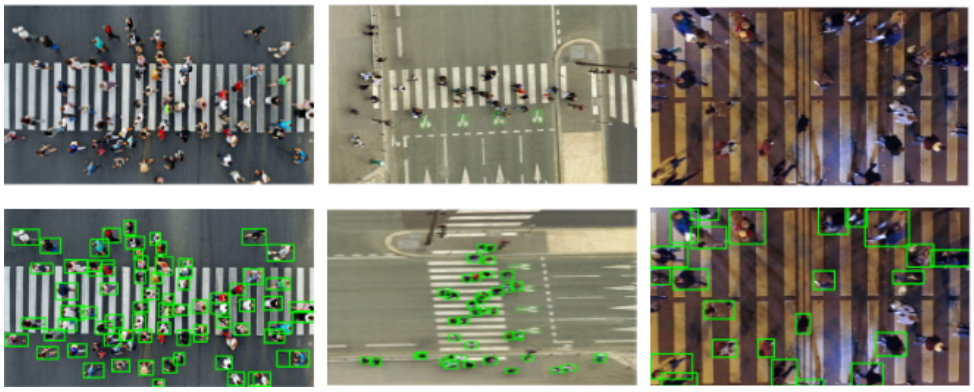

Figure 5: Examples of pseudo-labels generated by the VLM for a subset of the 175 unlabeled images used in the knowledge distillation process.

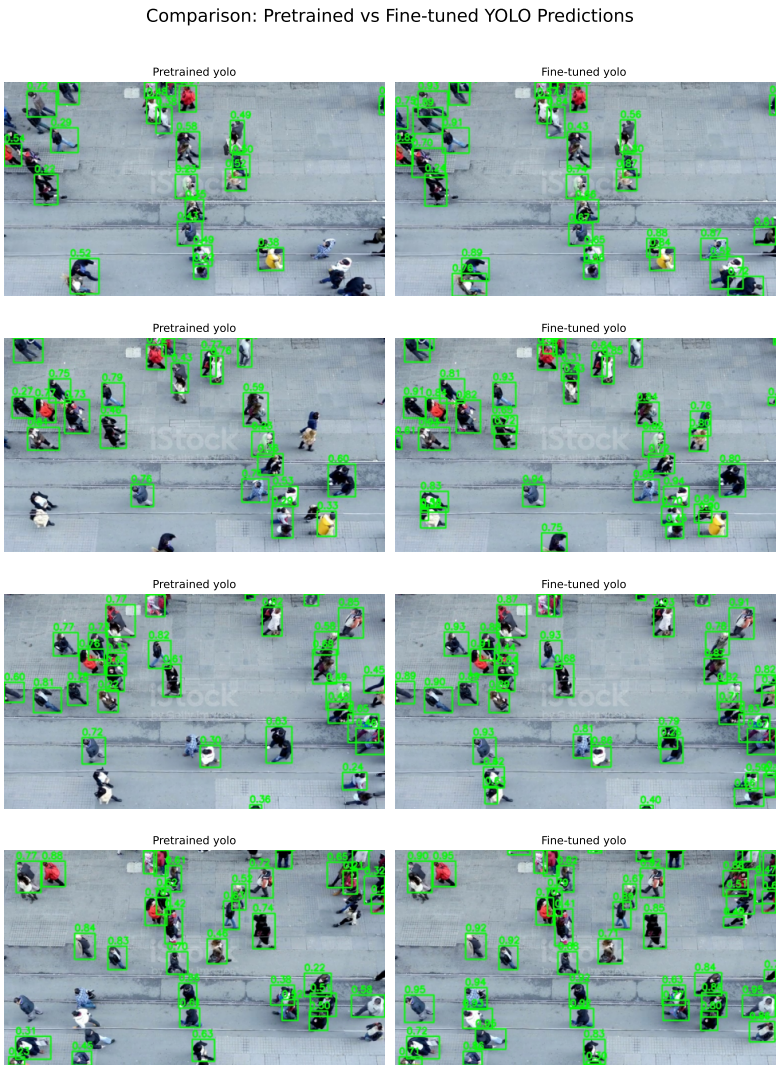

Figure 6: Output of the YOLOv8s model after distillation.

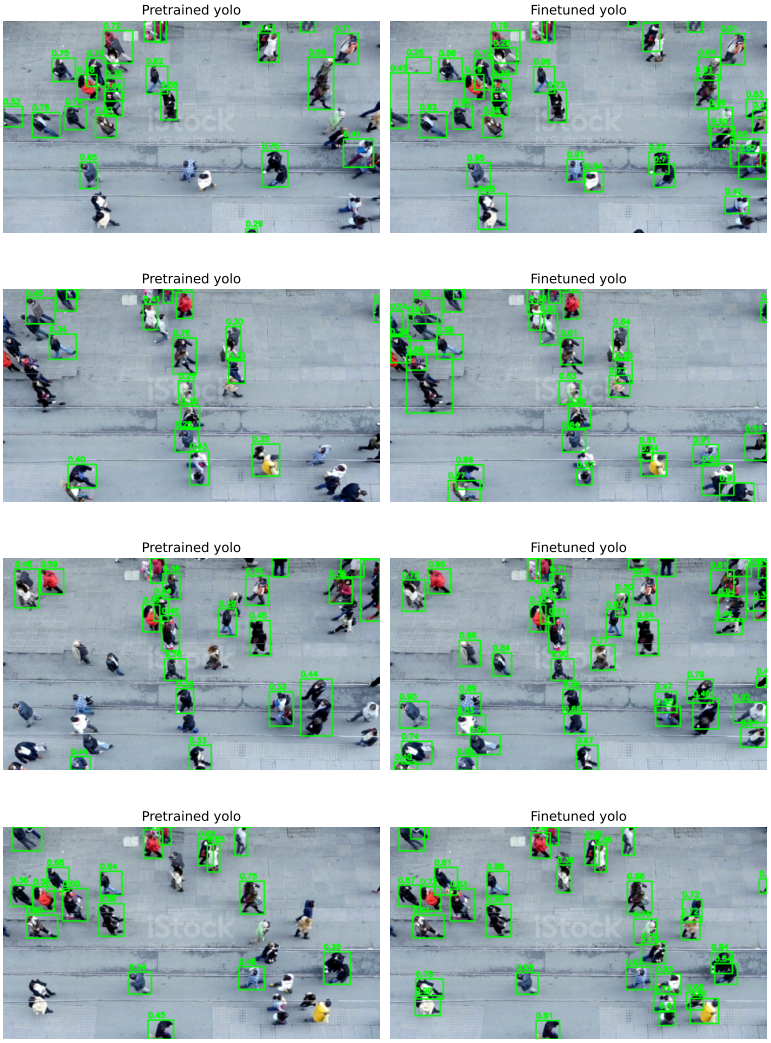

Figure 7: Output of the YOLOv8n model after distillation.

In Figures 6 and 7, it can be observed that the yolov8s and yolov8n models achieve improved detection performance after being fine-tuned using the pseudo labeled data generated by the VLM. a more detailed comparison is presented in Table 2. This approach leverages the knowledge of the VLM without requiring any manual annotation cost, while preserving the high inference speed of yolo, making it well-suited for real-time applications.

3 Real-time people Tracking and Counting system

Based on three models (VLM, pretrained yolov8n, and finetuned yolo (using knowledge distillation from VLM to yolo)) the system performs person detection and counting, and the results are saved in the output-videos folder. According to the Frames Per Second (FPS) evaluation, VLM models demonstrate substantially lower speed, obtaining **6.88 FPS**, while yolo reaches **58.90** FPS. This indicates that VLM is approximately ten times slower than yolo.

Model	Precision	Recall	F1	Time (ms)
VLM (Grounding-DINO)	0.844	0.948	0.893	212
YOLOv8s (pretrain)	0.964	0.543	0.691	23
YOLOv8s (finetuned using the VLM model)	0.907	0.824	0.862	23
YOLOv8n (pretrain)	0.889	0.413	0.565	18
YOLOv8n (finetuned using the VLM model)	0.913	0.753	0.824	18

Table 2: Comparison of object detection performance between Grounding-DINO-Tiny (VLM), YOLOv8n, and YOLOv8s. After performing knowledge distillation from the VLM into YOLO, the student YOLO models achieve detection performance close to the teacher VLM model, without using any manually annotated training data. The reported precision, recall, and F1 scores are calculated on four randomly selected frames that were manually annotated only for evaluation. Inference time corresponds to the average latency per image (in milliseconds).