ISY5002 Pattern Recognition Systems (PRS)

Project Report

Street Owl

Night's Watch Team Members

Tan Eng Hui (A0291201W)

Wang Tao (A0291189R)

Hu Lei (A0120681N)

Ho Zi Hao Timothy (A0150123B)

Content

1. Executive Summary

Project Overview

Retailers, event organisers, and municipal planners often lack tools to analyse street-level crowd patterns, density, and movement in real-time, which could provide valuable insights for optimising operations, marketing, and resource allocation. Although many public spaces are monitored by street-view CCTV systems, these are typically used only for security, leaving their full potential for business and operational insights untapped. Street Owl transforms live video feeds from street-view CCTV into actionable crowd analytics, enabling stakeholders to understand crowd dynamics and respond effectively to on-the-ground conditions.

Solution Benefits

Street Owl provides a unique solution by offering:

1. **Real-time crowd detection and density classification**, allowing retailers and other stakeholders to monitor public foot traffic, peak times, and congestion points outside their premises.
2. **Seamless integration with existing street-view CCTV systems**, avoiding the need for additional equipment and making advanced analytics accessible through current infrastructure.
3. **Actionable, location-based insights** that support smarter decision-making for staffing, marketing efforts, event planning, and public safety, enabling stakeholders to improve both customer engagement and operational efficiency.

Market Landscape and Opportunity

In Singapore's highly urbanised setting, efficient use of public space and insight into pedestrian flow are essential for both businesses and public safety. With extensive street-level CCTV coverage and constant activity in retail and commercial areas, Singapore provides an ideal environment for Street Owl's crowd analytics platform. By analysing foot traffic and crowd patterns, the platform enables retailers, event organisers, and city officials to make informed decisions on staffing, marketing, and resource management, enhancing customer experiences and public engagement.

Street Owl's technology is also applicable to other high-density urban areas worldwide, helping stakeholders globally unlock the potential of existing surveillance systems for strategic urban management and improved customer interactions.

Conclusion and Future Prospects

Street Owl aims to broaden its capabilities by integrating with a wider range of surveillance systems and refining its analytics platform. This will support diverse stakeholders in accessing reliable crowd insights from street-view data, ultimately fostering more responsive, data-driven urban spaces.

2. Project Background

2.1. Problem Statement

In densely populated urban environments, such as Singapore, understanding crowd dynamics is critical for businesses, event organisers, and municipal planners. Despite widespread use of CCTV systems in public spaces, these systems are underutilised, primarily serving security purposes rather than providing actionable business insights. The lack of integrated real-time analytics tools on crowd density, movement patterns, and behaviours limits stakeholders' ability to optimise operational decisions, manage resources efficiently, and improve customer experience.

Current solutions are primarily designed for traditional security monitoring, offering basic crowd detection features without the depth of analysis required for business-oriented insights. This limits their capacity to provide the actionable intelligence needed for real-time operational adjustments, leaving gaps in foot traffic analysis, resource planning, and tailored marketing strategies.

To address these gaps, the project proposes to leverage AI-driven video analytics to transform CCTV footage into a platform for real-time crowd insights. This enables businesses to harness data for strategic decision-making, offering benefits like optimised staffing, targeted marketing, and better crowd management, especially in high-traffic areas.

2.2. Market Research

Based on the market research, the global video analytics market is expected to grow from USD8.3 billion in 2023 to USD 22.6 billion by 2028, at a compound annual growth rate of 22.3% during the forecast.[1] By region, the North American crowd analytics dominates this market, the rapid advancements in automation and digitalization in the area and developing technologies are the key factor of market expansion. In the meanwhile, the Asia Pacific Crowd Analytics market is expected to grow at the fastest CAGR from 2023 to 2032. [9] With the wide adoption of high-resolution camera and sensors to capture more clear and detailed footage, enabling more accurate analysis and interpretation of data.

The growing demand for high-resolution cameras drives the rapid growth of video analytics demands. AI Video analytics includes using algorithms to read the video content for several purposes such as security surveillance, crowd analysis, retail and market analysis, facial recognition, license plate recognition, traffic monitoring, production & maintenance and so on. Categorized by different applications, the crowd

analysis is expected to hold a large market share. There is scope for our project in the current market climate to add business insights to surveillance data.

2.3. Market Competitors

We believe that our proposed solution would help Singapore businesses make better data-driven decisions to enhance customer engagement and operational efficiency. Existing projects do not identify crowd distinct features, utilize near real–time data. This is the summary of the capabilities of crowd detection systems from major competitors:

Company Name	Features	Use Cases
Vizsafe	Real-time crowd monitoring, geolocation, safety alerts	Public safety, event management, stadiums, cities
CrowdVision	Automated crowd density analysis, real-time video feed	Airport s, transport hubs, large venues
Camio	Smart video monitoring, AI video analytics, alerts	Corporate offices, retail, public spaces
DataFromSky	Object tracking, behavior analytics, movement prediction	Smart cities, traffic management, stadiums, events
Xovis	People flow analytics, sensor-based crowd measurement	Retail, airports, transit hubs
iOmniscient	Facial recognition, behavior analysis, density management	Smart cities, large venues, airports, public events

2.4. Value Proposition

The key value add of this project crowd detection system is to analyse valuable business-value information for retail businesses, with regards to crowd footfall. The system aims are summarised below:

- **Enhance Real-Time Crowd Monitoring:** Enable accurate observation and analysis of crowd dynamics in various settings.
- **Optimise Business Strategies:** Utilise crowd data and insights to support informed business decisions tailored to observed patterns and conditions. Understand the profile of the crowd by querying using LLM the clothes colour of crowd members.
- **Support Strategic Planning:** Provide actionable insights to various stakeholders, including business owners, investors, and other stakeholders, to facilitate better decision-making

and resource allocation. This is facilitated by crowd density classification using the classification model.

2.5. Business Model

The business model for the project would be SaaS (Software as a Service) platform for crowd detection. The SaaS is a infrastructure-light strategy so that new Street Owl customers are able to benefit from the video analytics platform without relying on installing technical hardware infrastructure. A user-friendly dashboard would allow them to generate insight on their surveillance data, including crowd density, crowd detection and crowd feature recognition.

Our target business customers' profile includes retail shops, event management businesses, transportation hubs, and smart city. These various organisations have key need to gain deeper knowledge over their crowd information. We will offer different subscription tiers to businesses, to cater different functions for their specific needs. These tiered capabilities range from basic crowd density prediction to more advanced features like crowd feature recognition. Our platform will also be adjustable to include API access to integrate into the business' existing dashboard or data platforms for greater synergy.

2.6. Pricing Strategy

The pricing strategy will offer three types of plans for business to choose from, which are the Basic, Pro, and Enterprise tiers. The Basic tier, priced at an affordable rate, will include standard crowd analytics features like footfall counting and crowd density analysis, suitable for small retail businesses and community events. The Pro tier, targeting medium-sized businesses and larger retail chains, will include additional features such as real-time demographic analysis, time graph of foot traffic trends, and basic API access.

The Enterprise tier of the crowd system will target larger scale organisations, such as government agencies, MNCs or smart city projects. Furthermore, the project deliverables for Enterprise tier customers may include further finetuning of the vision model on their specific types of crowds along their physical location – to increase customisation to their business needs. Also, Street Owl system can be integrated to the Enterprise tier business platforms via API access, provide technical support customer service and enable them to harness the AI-driven crowd system platform for operational planning.

3. System Overview

3.1. Project Scope

This project involves the development of an advanced real-time crowd monitoring system using computer vision technologies.

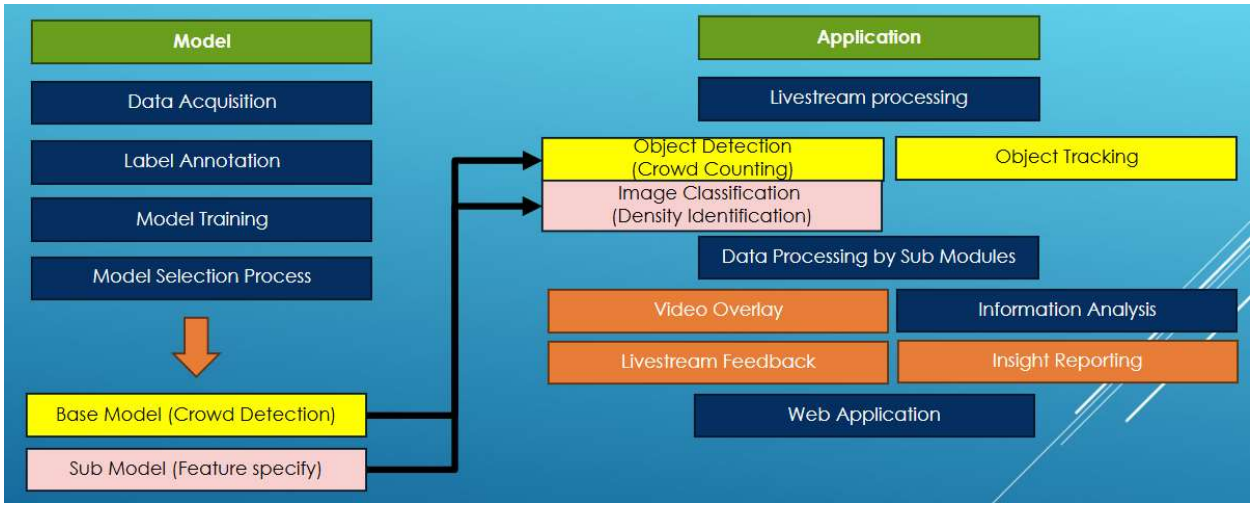
The three key features would be the following:

- **Develop a Near Real-Time Monitoring System**: The system should be capable of ingesting and processing live video feeds, such as live streams of various settings (e.g. public areas like street views).
- **Analyse Crowd Data**: Process and interpret real-time data to generate insights and reports on crowd frequency, congestion levels, and specific behaviours that would provide insight for businesses to act on.
- **Provide Data-Driven Insights:** Offer relevant recommendations to support decision-making for businesses and stakeholders, enabling them to optimise operations, investment opportunities, and strategic planning based on the analysis outcome.

The system with the use of artificial intelligence (AI) models shall be able to process live streams from surveillance cameras to produce useful and actionable insights on the dynamics and behaviour of the traffic.

3.2. System Design

The system is divided into two main components: Model and Application, each responsible for distinct tasks within the real-time crowd monitoring system.



Model Component

The Model component focuses on the development and selection of the AI models that would power the crowd monitoring system. This section describes the generic approach that we use to develop the two main models (Object Detection/Tracking Model for Crowd Count and Image Classification for Density Identification):

- **Data Acquisition:** Collecting of relevant datasets, such as video footage from surveillance cameras or other sources, to train the models.
- **Label Annotation:** The acquired data is labelled and annotated to identify different elements, such as individuals, objects, or specific behaviours. This step is essential for supervised learning techniques, allowing the model to recognize crowd patterns and behaviours accurately.
- **Model Training:** The labelled data is then used to train AI models. In this step, deep learning techniques, such as convolutional neural networks (CNNs), are applied to learn and detect human object
- **Model Selection Process:** After training, several models are evaluated based on performance metrics such as fitness and accuracy. Preliminary models would be selected and tested to assess the actual performance with the actual livestreams. This process is done iteratively to select the best model.

Application Component

The Application component is responsible for the real-time deployment and execution of the trained models to analyse live video feeds, provide feedback and display to the web application for consumption by the users:

- **Livestream Processing**: Handle and process live video feeds into processable frame in near real-time, ensuring continuous monitoring of various settings.
- **Object Detection:** To detect instances of human in the scene within the video stream.
- **Object Tracking:** Once detected, the system tracks the movement and behaviour of individuals over time.
- **Density Identification:** To detect density level based on the video frame presented to the model.
- **Data Processing by Sub Modules:** Specific sub-modules are responsible for processing the detected objects and tracking data.
- **Video Overlay:** Information would be overlay on the video frame to highlight detected objects such as object ID, previous coordinate and tripwire (left and right exit).
- **Livestream Feedback:** The video with overlay would be generated.
- **Information Analysis:** Collected data from object detection and tracking is processed to extract meaningful insights.
- **Insight Reporting:** Processed data are generated as chart and statistics to be displayed within the web application
- **Web Application:** The web application is accessible through a user-friendly web application, allowing stakeholders to interact with the application and make their business decision based on the trend.

3.3. Tools and Techniques

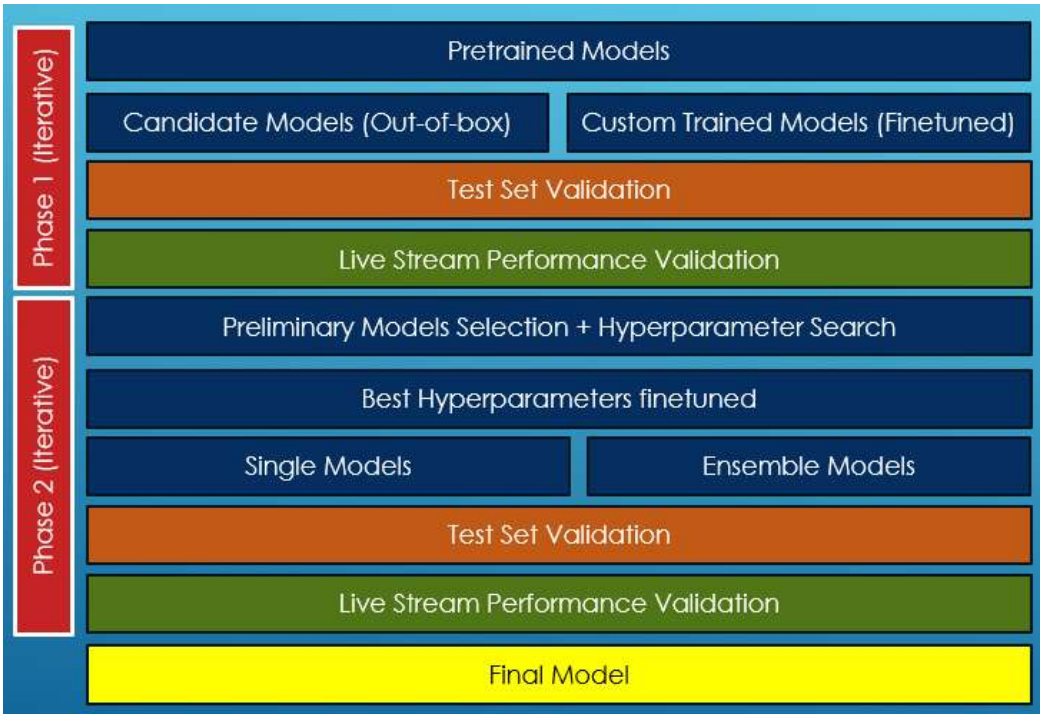
Tool	Purpose
YouTube Live Feed	Source of real-time surveillance footage from storefront environments for use in crowd analysis.
Streamlink	Capture and convert YouTube live streams into MP4 video format for further processing.
FFMPEG	Extract individual frames from MP4 videos as JPG images, forming the base dataset for model training.
Roboflow	Annotate extracted images with bounding boxes and object detection classes, configured for YOLOv8 compatibility.
YOLOv8*	Utilised as the pretrained model to detect and identify crowd objects within video frames. *For Yolo architecture, please refer to appendix
Streamlit	Provides an interactive interface for displaying live detection and analytics results to end-users.
PyTorch	Deep learning framework used for training, finetuning, and real-time inference of both the pretrained YOLOv8-based object detection model and the density classification model.

4. Model Component

For Street Owl, we have developed two AI models for effective crowd monitoring: **Crowd Object Detection Model** and **Density Classification Model**. These models enable real-time insights into crowd size and density by processing live video feeds.

Crowd Object Detection Model - we followed a structured and iterative approach that leverages pretrained YOLOv8 models, followed by evaluating both out-of-box and fine-tuned versions on our annotated dataset.

The initial model candidates underwent rigorous **Test Set Validation** and **Live Stream Performance Validation** to ensure their suitability for real-time applications. In Phase 2, we conducted hyperparameter tuning to optimise model accuracy and efficiency, comparing single and ensemble models to determine the best configuration for live performance. The final model was selected after multiple rounds of validation to achieve a balance between detection accuracy and processing speed, ensuring optimal performance in live streaming environments.



Density Classification Model - we developed a custom Convolutional Neural Network (CNN) from scratch to classify crowd density levels—sparse, dense, and crowded. While not following the structured processes, it was tested on live stream performance to validate its actual performance.

4.1. Crowd Object Detection Model for Crowd Counting

The Crowd Object Detection Model is designed to facilitate accurate, real-time crowd counting in public spaces by identifying and locating individuals within live video feeds.

Using the YOLOv8 architecture—a cutting-edge deep learning model trained for object detection—the model is then fine-tuned to detect human presence in each video frame. This model is used as the baseline for crowd counting for the Street Owl system, to enable crowd monitoring processing for further gaining of insights.

4.1.1. Dataset for Finetuning

For the Crowd Object Detection Model, the dataset consists of video frames captured from surveillance feeds, processed as images with annotations marking human presence. Key dataset statistics are:

Classes	No. of Images	No. of Instances	Avg Image pixel/size
1 (Human)	103	1006	2.06mp

Images are uploaded to the Roboflow platform and manually annotated via polygon around human figures so it can be used for segmentation/object detection purposes during exploration. The dataset is then versioned and exported in the format compatibility with the YOLOv8 model. The default ratio for training, validation and testing dataset is set at

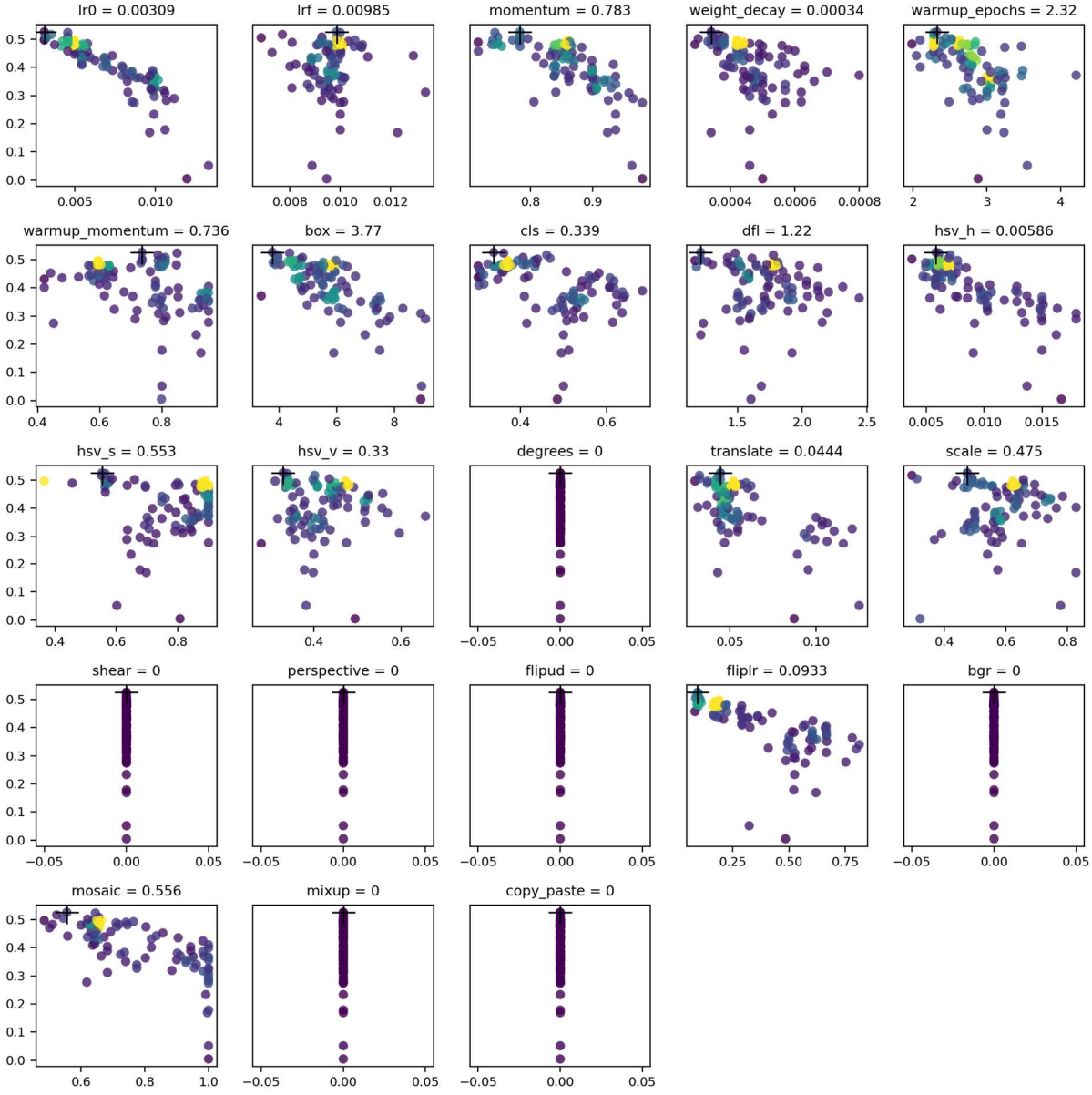
4.1.2. Models Training and Performance

In addition to using the YOLOv8 models as-is, we fine-tuned the YOLOv8 pretrained variants (e.g., YOLOv8n, YOLOv8s, YOLOv8l) using our annotated dataset. Each model then underwent hyperparameter tuning, followed by performance comparisons based on fitness metrics. To ensure a comprehensive evaluation, additional metrics were also reviewed to identify any extreme results, such as significant drops in precision or recall.

Finetuning and Hyperparameter search

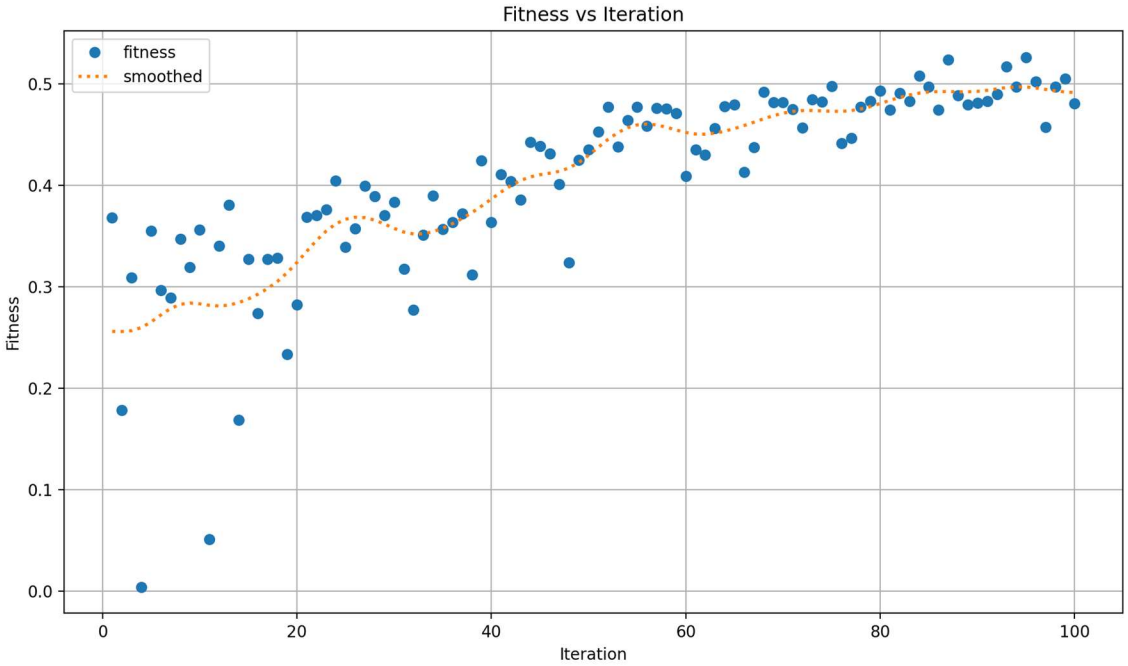
Each tuning experiment trains the specific model for 30 epoch and 100 iterations based on fitness function.

The following screenshot is an example of the tuning process and visualization.



Data Augmentation Parameters	
hsv_h	Adjusts hue in HSV colour space.
hsv_s	Adjusts saturation in HSV colour space.
hsv_v	Adjusts value (brightness) in HSV colour space.
degrees	Rotation augmentation.
translate	Translation augmentation.
scale	Scaling augmentation.
shear	Shearing augmentation.
perspective	Perspective augmentation.
flipud	Vertical flip augmentation.
fliplr	Horizontal flip augmentation.
bgr	Specifies if the image uses BGR colour order.
mosaic	Combines four images into one for training.
mixup	Combines two images and labels for augmentation.
copy_paste	Copies objects from one image to another for augmentation.

Fitness is then calculated using a weighted combination of several key performance metrics (mean of top-1 and top-5 accuracies as fitness score)



Performance Comparison Based on Metrics

All the YOLOv8 models in both original and fine-tuned states were compared using the identified metrics.

Hardware limitations also presented challenges in model training for lower-spec compute devices, particularly with models like YOLOv8x and YOLOv8l. Given the requirements for processing of live video, processing time on video feeds was also a critical factor in model selection.

While overall fitness is a useful metric, observations indicated that recall is particularly critical for our specific crowd monitoring requirements. Therefore, during preliminary models' selection, we prioritised models demonstrating a balance between precision and recall with small or moderate parameters/size.

	yolov8l	yolov8x	custom yolov8l	yolov8m	custom yolov8s	custom yolov8m	custom yolov8n	yolov8s	yolov8n
metrics/precision(B)	0.832891	0.747905	0.932068	0.836011	0.838637	0.812256	0.752601	0.721584	0.543725
metrics/recall(B)	0.686047	0.709302	0.638188	0.569767	0.725228	0.704336	0.686047	0.558140	0.457301
metrics/mAP50(B)	0.813421	0.793831	0.788736	0.764786	0.794818	0.799388	0.752222	0.672199	0.490784
metrics/mAP50-95(B)	0.534223	0.532265	0.518156	0.504664	0.496595	0.495705	0.414315	0.400311	0.277658
fitness	0.562143	0.558421	0.545214	0.530676	0.526418	0.526073	0.448106	0.427499	0.298971

Metric	Description
precision	Correct detections out of all positive predictions (true positive + false positive).
recall	Correct detections out of all actual objects (true positive + false negative).
mAP50	mean Average Precision at IoU threshold of 0.5.
mAP50-95	mean Average Precision across IoU thresholds from 0.5 to 0.95.
fitness	Composite score combining precision, recall, and mAP for overall performance.
IoU	Intersection over Union (IoU) measures overlap between the predicted and actual bounding boxes, showing how well the object is detected. A higher IoU means better localisation.

4.1.3. Evaluation of ensemble model

In the exploration process, various models, including YOLO v8s, YOLO v8m, and ensemble configurations, were tested, but no significant improvements were observed with the ensemble approach*. While stacking models can potentially enhance

detection accuracy, the ensemble model introduced added computational complexity with only minimal performance gains, making it unsuitable for real-time applications.

*Non-Maximum Suppression (NMS) was used to ensemble the models result which main purpose is to filter out redundant or overlapping bounding boxes. Results may differ for different IoU threshold applied. Too low/high are not good for performance but in general the algorithm might not perform well if multiple objects are very close together and bounding boxes overlap significantly which might be the situation for this project.

Model	Result
Custom Model YOLO v8s	27 humans detected
Custom Model YOLO v8m	21 humans detected
Ensembled Model*	31 humans detected
Ground Truth	23 human detected

This represents a single observation; multiple experiments with various combinations have been conducted to derive the overall findings and conclusions.

The single model finetuned YOLO v8m provides a balanced trade-off between detection accuracy and computational efficiency, making them suitable for our use case.

4.1.4. Evaluation and Outcomes

- **YOLOv8m** was selected as the optimal model for real-time crowd counting due to its effective balance of accuracy, processing speed, and manageable model size. This would meet the specific requirements of real-time video processing.
- Ensemble models were explored but were excluded because they added complexity without substantial performance gains. Hence, single model approach is preferred.
- Live performance evaluations demonstrated that YOLOv8m performed better than the default YOLOv8 models, confirming its robustness and suitability for real-world, continuous crowd monitoring in real-time application.

4.2. Density Classification Model for Density

The Density Classification Model was developed to classify crowd density levels in public spaces, distinguishing between sparse, dense, and crowded environments.

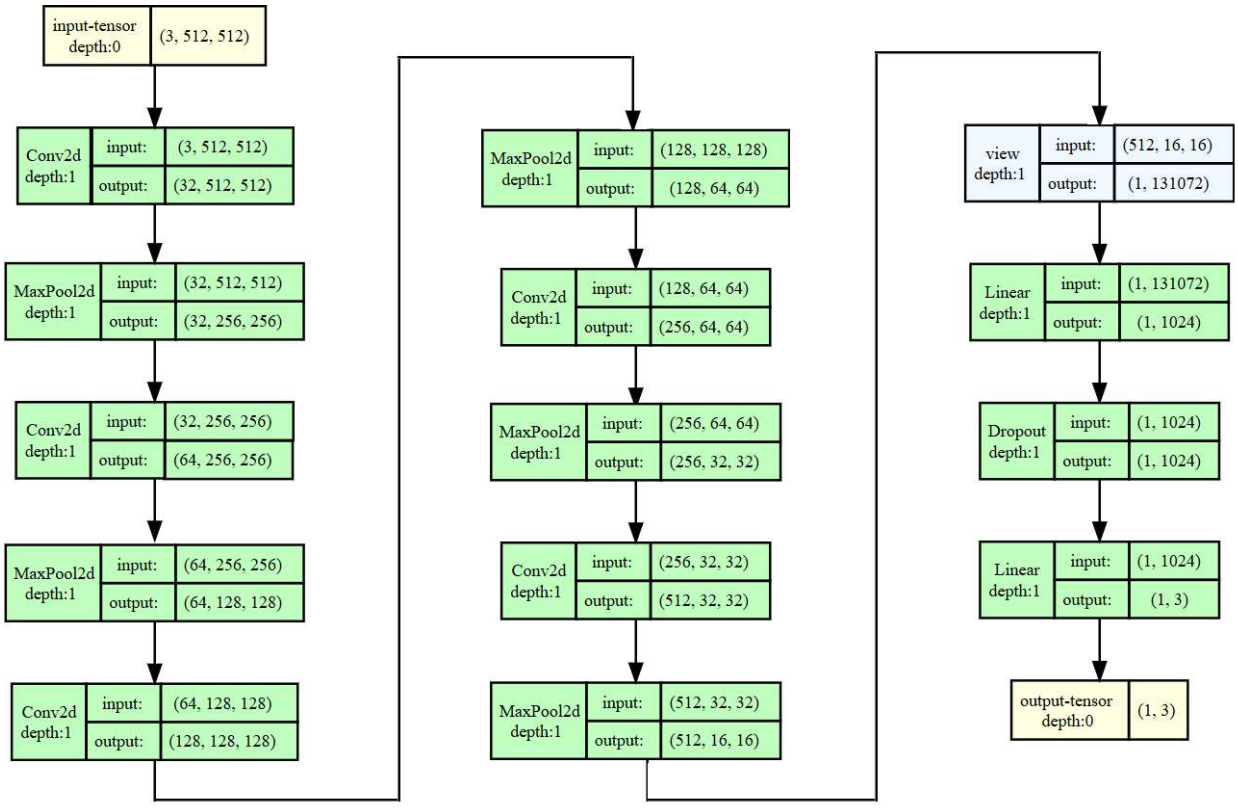
4.2.1. Models Classification Classes

The model uses three classes based on crowd density:

Classes	No. of Images	Avg Image pixel/size
0 (Sparse)	285	2.06mp
1 (Dense	44	2.06mp
2 (Crowded)	100	2.06mp

4.2.2. Model Training and Performance

The Density Classification Model is based on a convolutional neural network (CNN) architecture optimised for crowd density classification.



Model Summary

Input Layer: Processes high-resolution images with dimensions (3, 512, 512) (RGB channels).

Convolutional Layers:

- Layers with increasing depth (32 to 512 channels) to progressively capture more complex features.
- Essential for identifying patterns ranging from basic edges to intricate crowd density characteristics.

Max Pooling Layers:

- Applied after each convolutional layer to down sampling spatial dimensions, reducing computational requirements.
- Provides translation invariance, making the model robust to variations in crowd positioning.

Flattening Layer:

- Converts the final output of convolutional layers into a 1D vector for the fully connected layers.

Fully Connected (Linear) Layers:

- Combines extracted features to perform classification.
- Dimensionality is reduced from 131,072 to 1024 for computational efficiency.

Dropout Layer:

- Applied after the first fully connected layer to prevent overfitting by randomly zeroing out activations.

Output Layer:

- Maps to three classes (sparse, dense, crowded) for density classification.

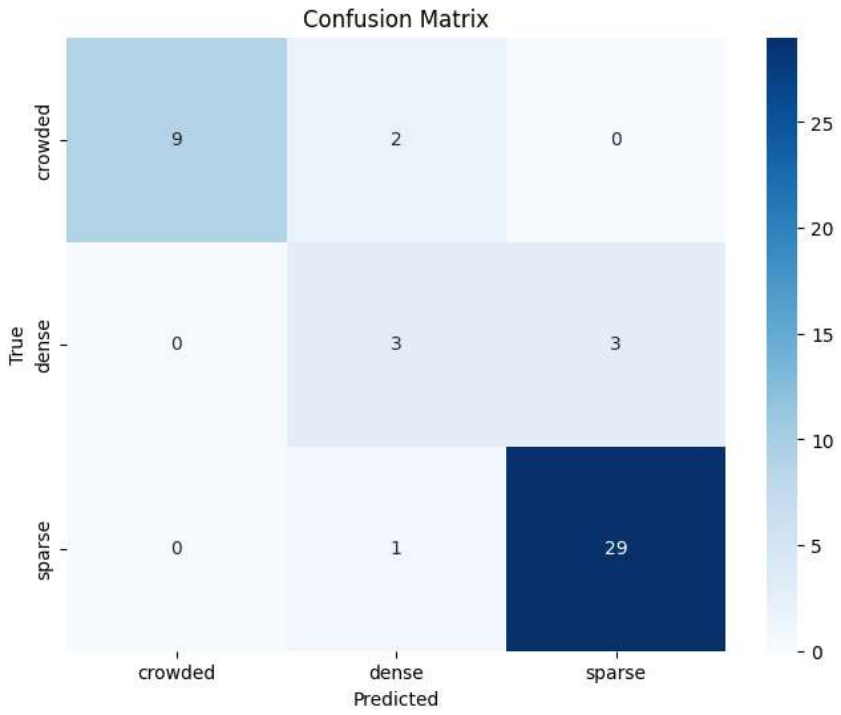
Model Training

CrossEntropyLoss was used as the loss function to measure the difference between the model's predictions and actual labels.

To ensure optimal model performance, we implemented a **best model selection technique** based on validation accuracy:

- At the end of each epoch, the model's validation accuracy (epoch_acc_val) is evaluated.
- If the validation accuracy for the current epoch surpasses the previous best accuracy (best_val_acc), the model is saved. This technique ensures that only the version of the model with the best validation performance is preserved, helping us retain the model with the highest generalisation capability.

Result based on test dataset

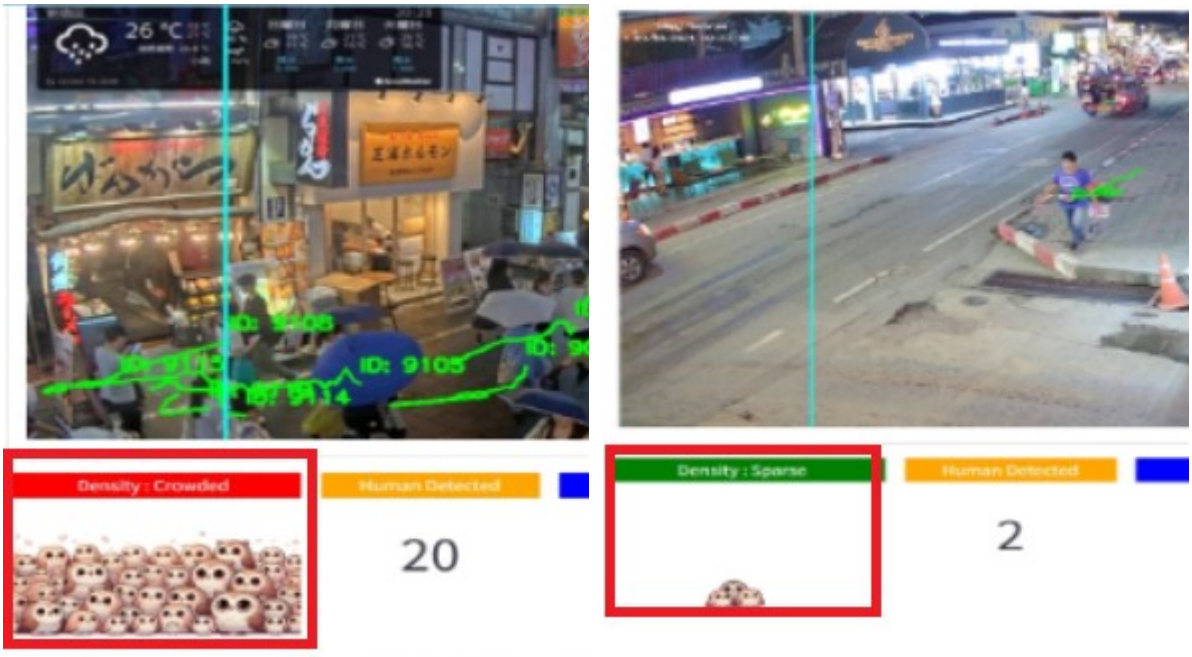


Class	Precision	Recall	F1-Score	Support
Crowded	1.00	0.82	0.90	11

Dense	0.50	0.50	0.50	6
Sparse	0.91	0.97	0.94	30
Accuracy			**0.87**	**47**
Macro Avg	0.80	0.76	0.78	47
Weighted Avg	0.88	0.87	0.87	47

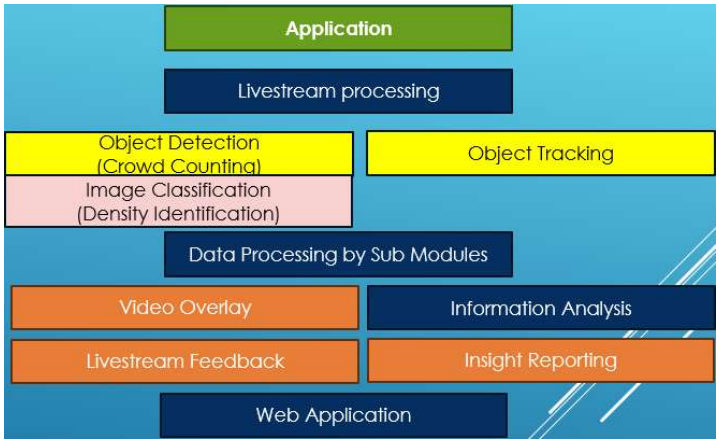
4.2.3. Evaluation and Outcomes

- The Density Classification Model effectively categorises sparse, dense, and crowded environments, achieving an accuracy of 87% on the test data and a macro F1 score of 0.78.
- Although it performs well overall, accuracy in the dense class is lower (0.5), indicating that the model could benefit from additional labelled data to improve differentiation between dense and crowded classes.
- Overall, in the actual environment with live video feed, the model provides consistent crowd density classification across varied scenes, offering valuable real-time insights for public space crowd management.



5. Application Component

The application component is integral to the real-time utilisation of trained models, enabling live video feed analysis, data processing, and insight generation within a user-friendly interface. This section details the web application development and system performance capabilities for the Street Owl system.



5.1. Web Application Development and Implementation

The Street Owl web application is developed using **Streamlit**, an interactive web-based framework enabling live streaming, real-time crowd detection, and analytics visualisation.

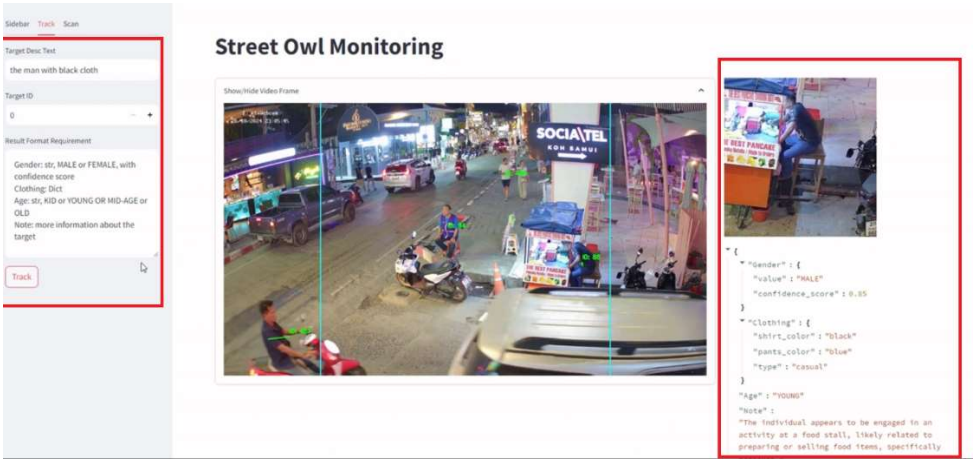
5.1.1. Application Key Features

Insight Report Generation:

- The insights gathered from LMM-driven tracking and detection are compiled into the **Insight Report**, which presents visual and statistical data on crowd dynamics.

LMM Dynamic Object Tracking:

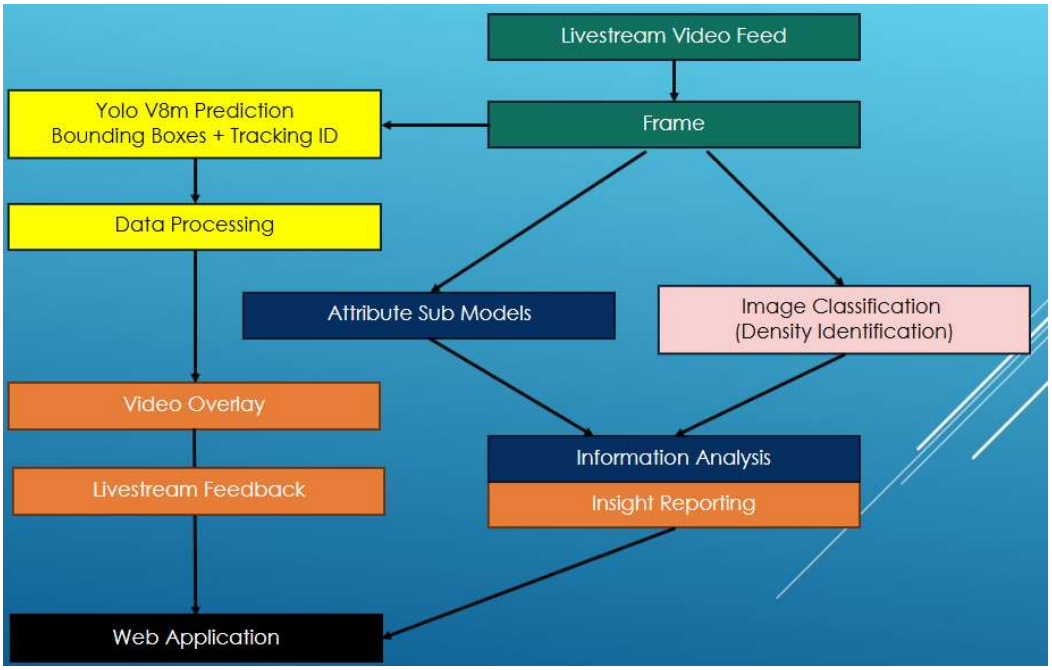
- In the **Track Mode**, users can define specific objects or individuals to analyse by providing two inputs: a description of the target and the required attributes. The workflow is as follows:
    - Detection results are processed by the LMM according to the user's description, allowing the system to identify and extract details for a single matching object based on specified attributes.
- This dynamic tracking allows for real-time adjustments, helping users adapt to changing conditions and analysis requirements without reliance on a pre-defined model output.



LMM Zero-Shot Detection with Scan Mode:

- In Scan Mode, the LMM employs OWL-ViT (an open-vocabulary zero-shot detection model) to detect new categories and objects based on user input.
    - Users specify target categories via natural language, and the model identifies multiple objects or categories in a single frame, supporting scenarios where predefined categories are insufficient.
- Advanced optimisations such as token-drop and instance selection enhance detection speed and efficiency.

5.1.2. Backend Processing



Livestream Video Feed:

- The video feed is captured in real time via Streamlink, providing a continuous source of crowd video feed.

Frame Processing:

- Each frame is passed through the **YOLO v8m model** for inference and processing, where bounding boxes and tracking IDs are generated to detect and follow individuals.

Data Processing:

- The output from YOLO is then processed to create overlays for tracking visuals on the video.

Split Paths for Further Analysis:

- The frame splits into two additional processing paths:
    - **Attribute Sub-Models**: Analyse specific behavioural traits of individuals within the crowd.
    - **Image Classification (Density Identification)**: Classifies crowd density as sparse, dense, or crowded.

Information Analysis:

- Insights from data processing, attribute models, and density classification are combined to generate comprehensive crowd analytics.

Insight Report Generation:

- The final insights are compiled into the Insight Report display via visualisation such as charts and statistics.

Web Application Display:

- All insights are displayed in the web application, providing users with real-time, visualised data and actionable insights for immediate decision-making.

5.2. System Performance

This section outlines the system's performance in terms of processing speed, detection accuracy, resource utilisation, and overall stability during real-time crowd monitoring.

Real-Time Processing Speed

- **Frame Rate**: The system was able to process live video at an average rate of 10~30 frames per second (fps) on Nvidia 3060, providing near real-time crowd insights. This may vary based on Video Quality of the live feed.
- **Overlapping and Occluded Objects**: Detection accuracy decreased in scenarios with overlapping individuals or heavy occlusion, particularly in dense crowd settings.
- **Hardware Constraints**: The system's performance was constrained on lower-spec hardware, requiring reduced video resolution or slower processing rates to maintain stability.

Detection and Classification Accuracy

- **Object Detection Accuracy**: Using YOLOv8 for crowd detection, the system achieved a precision of 88%, recall of 87%, and a mean Average Precision (mAP) of 87% on the test dataset. These metrics reflect the model's performance in identifying and counting individuals in varied environments.
- **Density Classification Accuracy**: The density classification model achieved an accuracy of 87% in distinguishing between sparse, dense, and crowded environments. The model maintained high accuracy in sparse environments but encountered challenges with distinguishing dense from crowded environments, with the dense class achieving a lower F1 score of 0.50.

Observed Live Video Feed Performance

- **General Detection Accuracy**: The system effectively detected and tracked individuals in real time, demonstrating strong overall performance.
- **Density Classification**: The model reliably distinguished between sparse and crowded scenes but had lower precision distinguishing dense from crowded settings. With combination using rubric, the performance is satisfactory to good.
- **Real-Time Responsiveness**: The system maintained real-time processing speeds, meeting requirements for typical crowd monitoring scenarios although the frame rate is not high.

6. Findings and Discussions

This section provides an overview of insights derived during the development of the models and application. Key findings reflect on model performance, real-world applicability, and technical considerations that influenced the final design and implementation of the system.

- **Metrics vs. Live Performance**: While offline metrics (precision, recall, mAP50) indicated high accuracy, live-stream deployment revealed drops in performance, especially in dense, overlapping crowds. This highlights the importance of validating models in real-world conditions.
- **Model Efficiency**: The YOLO v8m model was chosen for its balance between accuracy and processing speed. Ensemble models offered slight metric improvements but were computationally intensive, unsuitable for real-time applications.
- **Hardware Constraints**: High-performance models required lower resolutions and batch sizes to achieve real-time processing on standard hardware, reducing precision. YOLO v8m offered a practical trade-off, balancing accuracy and speed for live usage.
- **Dense Crowd Challenges**: Overlapping bounding boxes in dense crowds reduced detection accuracy. Non-Maximum Suppression (NMS) minimised redundancy, but closely spaced individuals continued to pose detection challenges.
- **Density Classification**: The density model reached 90.4% accuracy during validation but dipped to 87% on test data, particularly struggling with “dense” versus “crowded” classifications. Nevertheless, it reliably classified density in varied street scenes.
- **Dataset Quality**: Model reliability improved when live feed characteristics matched training data. A more diverse dataset would further improve the model's adaptability to varied conditions in real-world applications.

7. Challenges and Constraints

Throughout the project, several challenges arose in data gathering, labelling, and system compatibility. Below are the primary issues:

- **Data Gathering**
    - **Issue**: Sourcing high-quality street-level surveillance video feed of similar quality such as distance, zoomed-out perspectives, blockade. Some of the livestream also have unpredictable live streaming periods and limited public availability.
    - **Course of Action**: Leveraged live feeds from other countries such as Thailand and Japan, which provided suitable high-resolution images for real-time analysis to enrich our dataset.
- **Data Labelling**
    - **Issue**: The labelling process requires manual creation of bounding boxes following specific formatting based on framework for pretrained models, making it tedious, time-consuming and error prone.
    - **Course of Action**: We have explored multiple annotation tools and derive to the utilisation of Roboflow Platform, which allows easier labelling with UI and automated conformance of format for common framework such as YOLO. Saving us some time and effort in preparing data for training and deployment.
- **System Compatibility**
    - **Issue**: Ensuring compatibility across various operating systems (Windows, Mac, Linux) was challenging due to team members' diverse setups. The system relies on Torch for model training, Anaconda for environment management, and Streamlit for web interface output, requiring consistency across platforms.
    - **Course of Action**: Anaconda was used to standardise environment setup, simplifying dependency management across different OSs. Torch was selected for its compatibility with various Python versions and CUDA GPU support on Windows, unlike TensorFlow, which lacks Windows GPU support. Streamlit's cross-platform functionality enabled seamless deployment across all development environments.

8. Risks and Concerns

This section highlights potential risks and areas for improvement within the system. Given additional time and resources, we would like to address these issues to enhance the system functionality, scalability, and data privacy compliance.

- **Data Privacy Concerns and Solutions**

  As the system capture and present identifiable facial details within video feeds, which raises privacy concerns if end-users can view sensitive visual data. This may lead to privacy breaches, particularly in regions with strict data protection regulations.

  - **Issue:** Presenting identifiable facial details in video feeds risks violating privacy laws and could lead to misuse if end-users access sensitive visual information.
  - **Potential Solution:** Implement automatic face-blurring to anonymise individuals before presenting footage, allowing necessary details (e.g., clothing colour, carried items) to remain visible without revealing personal identifiers.

- **Scaling Challenges and Solutions**

  The current system's standalone setup requires each client to have dedicated processing resources, limiting scalability and introducing redundancy. Supporting multiple distributed clients demands a more centralised, flexible infrastructure.

  - **Issue:** Increased computational load for real-time analysis of multiple video streams may lead to latency and slow system performance.
  - **Potential Solution:** Integrate a Video Management System (VMS) to manage video streams effectively and balance the load across multiple servers and have a centralised control so processing of the video and data can be done more efficiently
  - **Issue:** Challenges in scaling infrastructure to support distributed clients may require additional resources and complex management.
  - **Potential Solution:** Employ cloud-based infrastructure with scalable resources to allow dynamic resource allocation based on client demand.

9. Conclusion

We believe that Street Owl has successfully shown its ability to deliver real-time crowd detection and density analysis, helping businesses optimise decisions. This MVP has met the initial goals of providing valuable insights through a web application with live stream processing, featuring crowd object detection, density classification, and profile analysis using LMM to capture individual details. These capabilities enable businesses to better understand and respond to crowd dynamics, enhancing operational efficiency, customer targeting, and overall strategic planning.

Future Roadmap

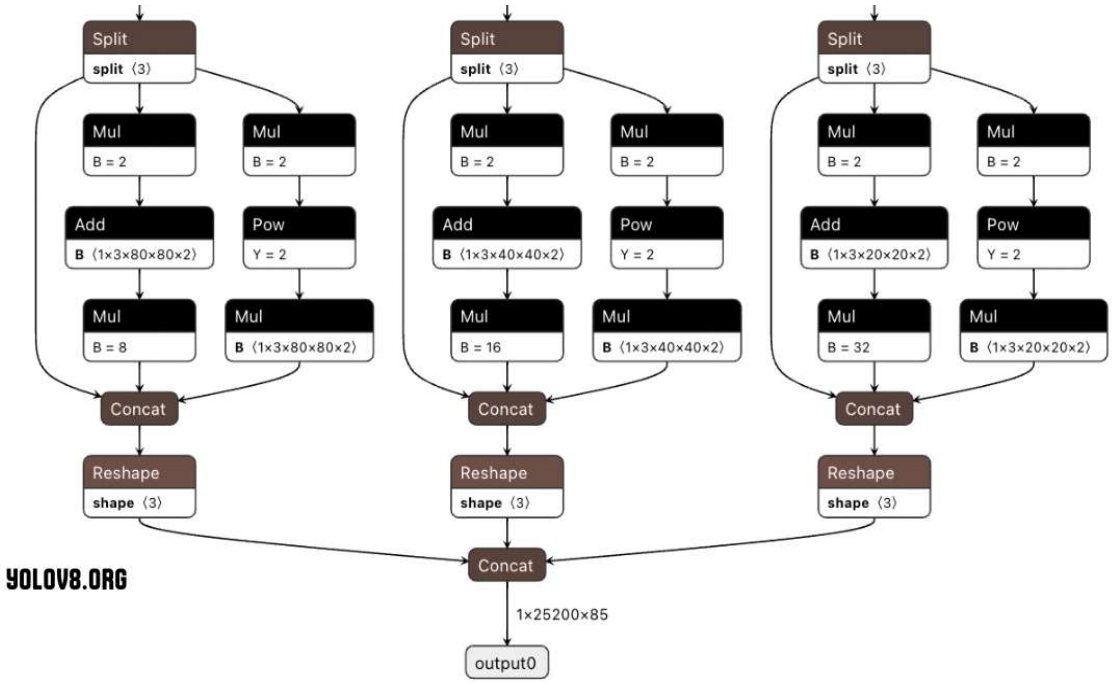
To enhance Street Owl's capabilities, future directions will refine insights and expand functionality, offering businesses more targeted and predictive crowd analytics. These upgrades aim to give a deeper, forward-looking understanding of crowd dynamics, supporting strategic planning and customer engagement.

- **Enhanced Crowd Feature Detection**: By identifying additional crowd traits like group size, based on indicators such as holding hands, similar attire, or synchronised movement, businesses can gain a clearer view of demographic details. This can support targeted marketing initiatives, such as group-based promotions, by recognising crowd compositions like student groups or couples, allowing more personalised engagement.
- **Enhanced Density Classification Model:** Improving the density classification model using an autoencoder or transfer learning from pretrained models like ResNet would allow for more accurate mapping of live inputs into a detailed feature space. This enhancement would diversify model capabilities, enable potential ensemble applications, and create a more adaptable system for complex crowd density scenarios.
- **Time-Series Prediction of Crowd Levels**: Predictive analytics, potentially using models like LSTM, would allow Street Owl to forecast crowd flow based on factors like time, day, or recent trends. This capability would provide businesses with forward-looking insights, helping them plan resources and staffing proactively during peak times, such as holidays or events, and enabling smoother operations and customer experience.

These enhancements are crucial for scaling Street Owl's impact and improving service offerings, keeping us competitive and responsive to the changing needs of our users.

Appendix

A. Model Architecture for Crowd Object Detection Model



YOLO v8 Model Overview (source: yolov8.org and yolov8.org/yolov8-architecture)

Backbone: This is the convolutional neural network (CNN) responsible for extracting features from the input image. YOLOv8 uses a custom CSPDarknet53 backbone, which employs cross-stage partial connections to improve information flow between layers and boost accuracy.

Neck: The neck, also known as the feature extractor, merges feature maps from different stages of the backbone to capture information at various scales. YOLOv8 Architecture utilizes a novel C2f module instead of the traditional Feature Pyramid Network (FPN). This module combines high-level semantic features with low-level spatial information, leading to improved detection accuracy, especially for small objects.

Head: The head is responsible for making predictions. YOLOv8 employs multiple detection modules that predict bounding boxes, objectness scores, and class probabilities for each grid cell in the feature map. These predictions are then aggregated to obtain the final detections.

B. Installation/Deployment Guide

Prerequisite:

Graphic card with CUDA support
Anaconda (python 3.10)

Installation and Deployment Instructions:

1. Navigate to the SystemCode folder in the anaconda terminal:
   cd SystemCode
2. Create the environment with the requirements:
   conda env create -f environment.yml
3. Activate the Conda environment:
   Conda activate torch
4. [Optional] Setup your OpenAI Key*:
   export OPENAI_API_KEY="xxxxxxxxxxxxxxx"
   * *Note: Only LMM feature that relies on the OpenAI cannot be used without the key.*
5. Launch the Streamlit application:
   streamlit run app.py
6. Navigate to **http://localhost:8501/**

C. Matrix for Project Module Mapping

Requirements	System/Feature Mapping
Supervised Learning	Supervised model training with annotated labels for finetune pretrained models Supervised image classification for density classification model
Deep Learning techniques	Training multiple YOLO v8 with techniques such as data augmentation, hyperparameter search, evaluation of models Convolutional Neural Network (CNN)-based Density classification model training
Ensemble approach	Using NMS to filter bounding boxes and integrate results from multiple models into single output
Intelligent Sensing/Sense making	Using live stream video feed images to extract useful information and understanding underlying events to produce insights for decision making

D. Reference

1. United Overseas Bank (UOB). (n.d.). Digital assets, Web3, and AI strategy. UOB Group. Retrieved from https://www.uobgroup.com/techecosystem/sff/digital-assets-web3-ai-strategy.html
2. Ivey Publishing. (n.d.). United Overseas Bank: Branch crowd analytics. Ivey Publishing. Retrieved from https://www.iveypublishing.ca/s/product/united-overseas-bank-branch-crowd-analytics/01t5c00000DMLWWAA5
3. Senstar. (n.d.). Crowd detection: Keep people safe with intelligent video analytics. Senstar. Retrieved from https://senstar.com/products/video-analytics/crowd-detection/
4. Bosch Security. (n.d.). Crowd detection with Bosch video analytics. Bosch Security. Retrieved from https://www.boschsecurity.com/sg/en/solutions/video-systems/video-analytics/video-analytics-types/crowd-detection/
5. Fortune Business Insights. (n.d.). Video analytics market size, share & industry analysis, 2024-2032. Retrieved from https://www.fortunebusinessinsights.com/industry-reports/video-analytics-market-101114
6. YOLOv8. (n.d.). YOLOv8 architecture. Retrieved from https://yolov8.org/yolov8-architecture/
7. MarketsandMarkets. (n.d.). Intelligent video analytics market by application, vertical, and geography - global forecast to 2026. Retrieved from https://www.marketsandmarkets.com/Market-Reports/intelligent-video-analytics-market-778.html
8. Ultralytics. (n.d.). Hyperparameter tuning. Ultralytics. Retrieved from https://docs.ultralytics.com/guides/hyperparameter-tuning/
9. Crowd Analytics Market Research Report. Retrieved from https://www.marketresearchfuture.com/reports/crowd-analytics-market-1850
10. Hugging Face. (n.d.). OWL-ViT model documentation. Retrieved from https://huggingface.co/docs/transformers/en/model_doc/owlvit
11. Hugging Face. (n.d.). OWL-V2 model documentation. Retrieved from https://huggingface.co/docs/transformers/en/model_doc/owlv2

Other References:

- Vizsafe - Safety and Security Solutions | Vizsafe | United States
- CrowdVision - Automated people tracking using video analytics
- Intelligent video monitoring using AI and standard cameras | Camio
- Deep traffic video analysis - DataFromSky
- People Flow and People Counting solutions | Xovis

- iOmniscient - a World Leader in AI based Video Analytics
- Caught on camera: How 90,000 police cameras across Singapore help solve crimes
- Surveillance camera statistics: which are the most surveilled cities?
- Retail Redefined – Singapore – MTI