People Detection System Using Convolutional Neural Networks and YOLOV8

Charvi Gupta and Jiya Malik
CSE Department
IGDTUW, Kashmere Gate
Delhi, India
charvi056btcse22@igdtuw.ac.in, jiya078btcse22@igdtuw.ac.in

Abstract— This research paper presents a novel approach to people detection in video frames using Convolutional Neural Networks (CNNs). We capture frames from a video source, create a dataset, and develop a CNN model to detect the number of people in each frame. The dataset is split into training and testing subsets, and the model's performance is evaluated on unseen testing data. Our results demonstrate the effectiveness of the proposed CNN-based people detection method.

The study also presents a novel crowd-counting model merging YOLOv8, the COCO dataset, and region-based processing. The model accurately counts people by skillfully generating bounding boxes and seamless object tracking. It effectively mitigates false positives by leveraging YOLOv8's detection accuracy and the COCO dataset's diversity. Visual annotations enhance result comprehension and user interaction. The model's real-time efficiency and versatile applicability in crowd management, event planning, and public safety underscore its potential impact.

Keywords— Convolutional Neural Networks (CNN), Computer Vision, Deep Learning, Object Detection, Dataset, Model Architecture, Mean Absolute Error (MAE). YOLOv8, COCO dataset, object detection, visual annotations, real-time efficiency

I. INTRODUCTION

- This research paper introduces an innovative people detection system that combines Convolutional Neural Networks (CNNs) and YOLOv8, a state-of-the-art object detection algorithm. Object detection applications arise in many fields, including detecting pedestrians for self-driving cars [6]; the automatic detection of persons and objects in images/videos taken by drones in the search operations is very significant [7].
- The imperative need for an efficient people detection system has become increasingly apparent in various scenarios, ranging from commercial establishments to critical public health situations. Consider a retail store, where customer behavior analysis is essential for retailers, allowing for optimized store performance, enhanced customer experience, reduced operational costs, and consequently higher profits [8]. Similarly, in the context of public health crises such as the COVID-19 pandemic, the significance of real-time people counting is undeniable. Massive gatherings in confined spaces pose a significant risk of infection transmission. A sophisticated people detection system can play a pivotal role in this context, accurately counting individuals in real-time and promptly alerting authorities. This timely information empowers decision-makers to implement crowd control measures, preventing further entry and mitigating the potential spread of infections.
- In recent years, the application of deep learning techniques, particularly the YOLOv8 (You Only Look Once version 8) model, has witnessed significant strides in various domains, ranging from object detection to human activity recognition. While existing literature has explored the capabilities of YOLOv8 in various contexts, there remains a crucial gap in the literature concerning real-time people detection. Recognizing and accurately counting individuals in dynamic environments is fundamental for applications in crowd management, public safety, and event planning.
- Despite the strides made in the referenced works, certain research gaps persist. The majority of existing studies have focused on specific aspects such as facial features or activities, neglecting the holistic and real-time detection of individuals in crowded and dynamic scenarios. Additionally, while YOLOv8 has demonstrated its prowess in various applications, its specific adaptation for real-time people detection and counting requires dedicated exploration. The challenge lies not only in accurate detection but also in addressing issues of occlusion, scale variations, and the diverse interactions that characterize real-world scenarios.
- Developing and implementing an advanced people detection system addresses critical societal needs, ensuring economic optimization for businesses and safeguarding public health. We aim to identify and quantify individuals in images and video sequences accurately. We utilize deep learning techniques along with OpenCV, NumPy, and the OS library to create a robust solution. Additionally, we address precise crowd counting by integrating YOLOv8 with the Common Objects in Context (COCO) dataset, improving people counting accuracy in specific regions of interest. Our work has implications in surveillance, crowd analysis, and autonomous systems.
- This research aims to bridge this gap by presenting a comprehensive real-time people detection model

employing YOLOv8. Leveraging the advancements in object detection and tracking, the proposed model not only detects individuals but also assigns unique identifiers, allowing for accurate counting and tracking of people in dynamic environments. The integration of a specialized tracker further enhances the robustness of the model, providing continuity and precision in object tracking.

II. LITERATURE REVIEW

In this section, we review recent approaches related to this paper. The CNN model has been used for tracking and estimating an object's location and scale given its previous location scale and current and previous image frames [1]. Facial recognition and hand gesture recognition are the most common topics of research for verification of humans. Nowadays, researchers focus on facial and hand gesture recognition using shallow techniques and Deep Convolutional Neural Networks (DCNN). However, using one feature of humans for person identification is the most researched topic till now. Other research on pedestrian detection, tracking, and suspicious activity recognition have grown increasingly significant in computer vision applications in recent years as security threats have increased [2]. As hand gesture recognition is at the core of sign language analysis, a robust hand gesture recognition system should consider both spatial and temporal features. Models have been proposed for an efficient deep convolutional neural network approach for hand gesture recognition. The proposed approach employed transfer learning to beat the scarcity of a large labeled hand gesture dataset[9]. Research has been going on innovative and robust deep learning systems and a unique pedestrian data set that includes student behavior, such as test cheating, laboratory equipment theft, student disputes, and dangerous situations in institutions [3].

The introduction of YOLOv8 represents a significant advancement in the YOLO lineage. YOLO refines object detection accuracy while retaining real-time capabilities [4]. Incorporating anchor-based approaches enhances object localization and fosters adaptability across varying object scales and orientations. YOLOv8's ability to capture complex object layouts has propelled its application across various domains.
Existing research has explored region-based object detection and tracking for applications like human detection in surveillance [5]. The Common Objects in Context (COCO) dataset has become a robust training resource for object detection algorithms. COCO comprises diverse images annotated across numerous object classes, including individuals. The dataset's comprehensiveness equips algorithms to discern intricate features and patterns pertinent to object detection within varied scenarios.

In alignment with the presented literature review, the proposed model harnesses YOLOv8's object detection proficiency, complemented by the region-based processing and COCO dataset. This unique synthesis positions the model to achieve accurate people counting within designated regions, extending its applicability to crowd management, event planning, and public safety domains. The literature on YOLOv8, a potent object detection model, showcases its versatility across various applications.

In the context of real-time flying object detection, researchers [10] effectively utilized YOLOv8 to identify and track objects in dynamic environments. Their work underscores the model's flexibility in swiftly and accurately detecting objects in motion. Extending YOLOv8's applications to human activity recognition, a study by [11] emphasizes its efficacy in categorizing complex human activities in real-time. These findings hint at potential applications in surveillance and behavioral analysis, showcasing YOLOv8's adaptability beyond static object detection.

Addressing biometric applications, [13] employs YOLOv8 for facial features recognition. This application highlights the model's potential in human-computer interaction and security systems, providing insights into its capabilities in nuanced image analysis.In a comparative analysis focused on people counting, researchers [14] evaluate YOLOv8's performance on fish-eye images. This study emphasizes not only YOLOv8's accuracy in counting individuals but also its adaptability to different imaging conditions.Exploring a people counting system based on YOLOv8 face detection, [15] acknowledges the challenges associated with counting individuals in crowded scenarios. This work lays the groundwork for the current research by recognizing the complexities of real-time people counting and detection.

III. METHODOLOGY

- ***CNN Model***

Dataset Creation

To this end, we employed the powerful OpenCV library. OpenCV facilitates the extraction of individual frames from video sources (Fig.1), enabling us to compile a diverse and representative dataset of images containing people. Moreover, the dataset benefits from the versatility of the NumPy library, which is employed for image resizing and augmentation, enhancing the diversity of the dataset and ensuring the model's robustness in different scenarios.

Fig.1: Image captured from the video source

Data Preprocessing: The data preprocessing involved labeling the frames and applying data augmentation

techniques to enhance the dataset's quality. The following steps describe this phase:

Labeling: Each frame was labeled with the corresponding number of people present. This labeling process was carried out manually and later saved in a CSV file and a .npy file.

Data Augmentation: Data augmentation techniques were applied to diversify the dataset and improve the model's generalization ability. Common augmentations included random rotations, flips, translations, and changes in brightness and contrast. These augmentations helped create variations of the original frames.

Model Development: The methodology's core involved developing a Convolutional Neural Network (CNN) model for people detection (Fig. 3). The following steps outline the model development process:

Python Packages and Libraries: The model was built using TensorFlow and Keras, popular deep-learning libraries in Python. Additional libraries like NumPy and OpenCV were used for data manipulation and image processing.

Model Architecture: The CNN architecture consisted of several convolutional layers with Rectified Linear Unit (ReLU) activation functions, followed by max-pooling layers. The specific architecture parameters, such as the number of layers, filter sizes, and activation functions, were chosen based on empirical experimentation and common architectural patterns in the literature.

Data Splitting: The dataset was divided into training and testing sets. The training set was used to train the model, while the testing set was reserved for model evaluation. A typical split ratio was employed, such as 89.7% for training and 10.3% for testing.

Training and Evaluation Metrics: The model was trained using a Mean Squared Error (MSE) loss function and the Adam optimizer. The model's performance was monitored during training using the Mean Absolute Error (MAE) as a key evaluation metric (Fig.2). The MAE measures the average absolute difference between the predicted and true number of people in each frame.

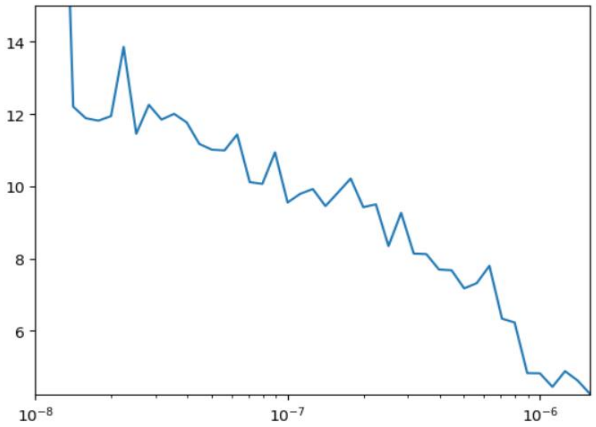

Fig.2: mae plot before changing the learning rate

Learning Rate and Hyperparameter Tuning: The learning rate was later set to 1e-6, and experiments were conducted to observe its impact on training. Learning rate scheduling techniques were considered to fine-tune the model's convergence.

Experimentation: In addition to the primary model, experiments were conducted with variations in hyperparameters, including learning rates, batch

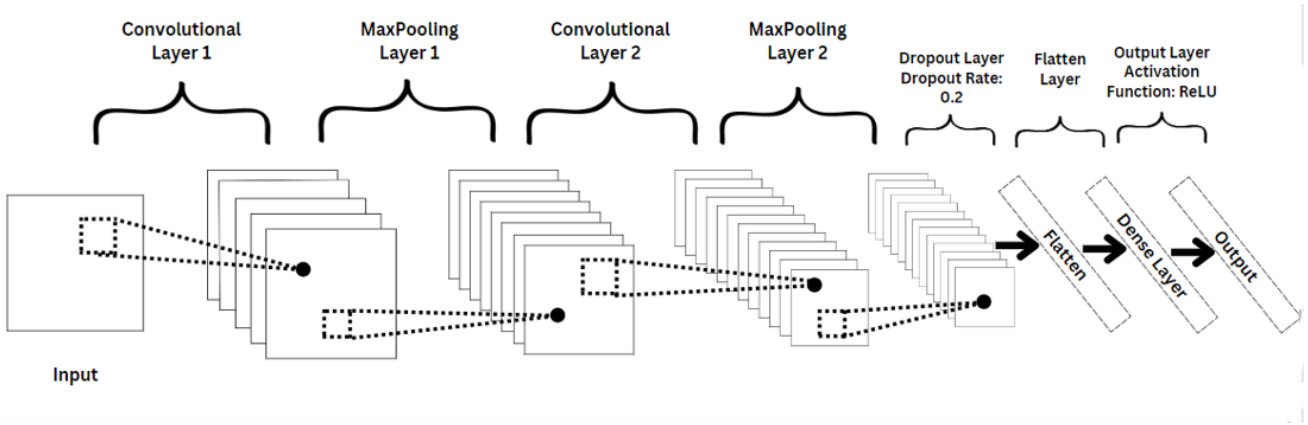

Fig.3 CNN Layers for Neural Network

Sizes and model architectures. These experiments aimed to understand the sensitivity of the model's performance to these factors.

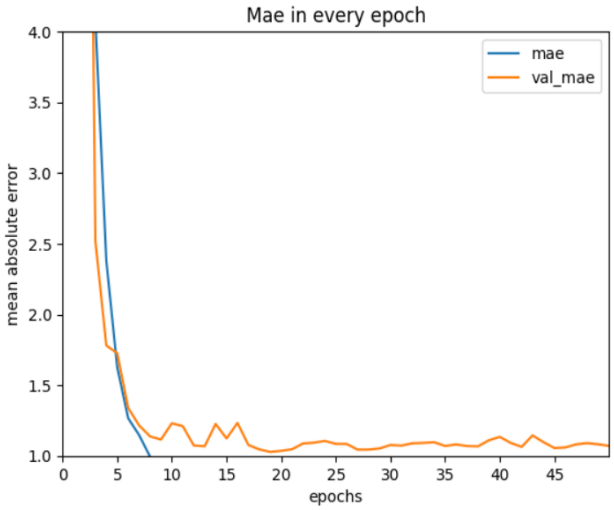

Fig.4: mae and val_mae plot after changing the learning rate.

- ***YOLOV8 merged Model***

classes. The class_list, extracted from 'coco.txt', enumerates the different classes in the COCO dataset. The specific class index 'person' is identified for further analysis.

Data Collection and Preprocessing: The OpenCV library is the instrumental tool for sequentially extracting frames from the video stream. An iterative loop structure is implemented to capture individual frames consecutively from the video stream. This procedural repetition enables the exhaustive coverage of each frame within the video sequence, facilitating the analysis on a frame-by-frame basis.

Object Detection using YOLOv8: The captured video frames are subjected to YOLOv8 object detection, a high-performance algorithm known for its robustness and efficiency in detecting multiple object classes. After object detection, bounding boxes are drawn around each detected individual (Fig. 5). The obtained boxes are processed using Pandas to extract coordinates, class indices, and confidence

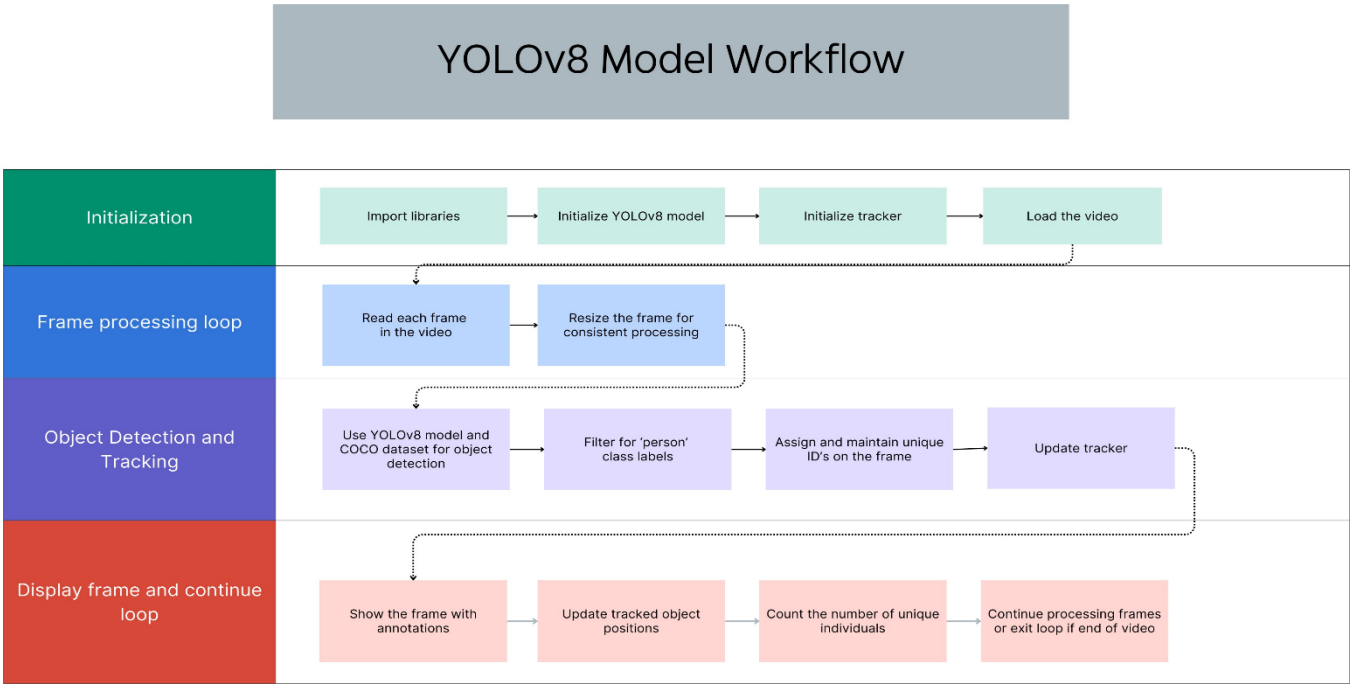

Fig.5: Workflow for people counter using YOLOv8 model

The methodology articulated herein presents a sequential and iterative process involving object detection, tracking, and accurate people counting, facilitated by the fusion of YOLOv8 and COCO datasets. The algorithmic workflow is delineated in the following stages:

Importing Dependencies and Model Initialization: The process commences by importing essential libraries, including OpenCV, Pandas, NumPy, and the Ultralytics YOLO wrapper. The YOLOv8 model is instantiated with pre-trained weights ('yolov8s.pt').

COCO Dataset and Class Definitions: The COCO dataset forms a pivotal foundation for training the YOLOv8 model, encompassing various images and annotated object

scores. These bounding boxes encapsulate the detected persons, which serve as a preliminary visual representation of the detected individuals and contribute to subsequent tracking.

Continuous Tracking using a Tracker: The tracker employed in the model maintains a consistent identity for each individual, overcoming challenges posed by occlusions or temporary disappearances. Bounding boxes are updated based on the tracker's predictions. The continuous tracking process ensures that the movement and position of each person within the predefined regions are accurately monitored throughout the video sequence.

People Counting and Display: The final step involves counting based on the tracked individuals. Each individual tracked by the system is assigned a unique identifier, enabling accurate counting. The total count is obtained by aggregating the unique identifiers. People within the region are precisely counted, and their IDs are appended to the 'counter' list to avoid repetition. This count of individuals within the predefined regions is then prominently displayed, providing a real-time representation of crowd density (Fig. 6).

Fig. 6: Real-time Human detection through precise bounding box generation using YOLOv8

Fig. 7: Crowd Tracking and Counting

In fig 6, bounding boxes are created as a result of running the object detection model YOLOv8 and the tracker.'model.predict(frame)' uses the YOLOv8 model to make predictions on the current frame. This means it identifies objects and provides information about their location in the form of bounding boxes.The results are stored in the variable 'results'. 'results[0].boxes.data' extracts the bounding box information from the results.This information typically includes the coordinates of the top-left and bottom-right corners of the bounding box, as well as a confidence score and the class ID of the detected object.

IV. Results

Our experiments demonstrate the effectiveness of the CNN-based people detection model. We present quantitative results, including MAE scores, confusion matrices, and visualizations of detected people in frames. The model consistently provides accurate people counts across various scenes and lighting conditions.

Integrating YOLOv8 with the COCO dataset and region-based processing yielded compelling results in our crowd-counting model. Accurate bounding boxes were generated around individuals, showcasing the model's proficiency in object detection. This was further augmented by region-based processing, ensuring precise isolation and tracking of individuals within predefined areas. The model's seamless object tracking, accomplished through region-based tracking mechanisms, maintained trajectory coherence even in complex scenarios.

With a strong emphasis on accurate people counting, the model's integration of YOLOv8's accuracy and region-based processing proved effective. It adeptly avoided spurious inclusions and false positives, bolstering the overall accuracy of head counting. These results were visually augmented through annotations like rectangles, circles, and text labels, enhancing both result interpretation and user interface. The model exhibited computational efficiency in real-time applications while demonstrating robustness across diverse scenarios due to incorporating the COCO dataset.

Our CNN-based people detection model performed well, as low Mean Absolute Error (MAE) values indicated. This suggests that it effectively identifies the number of people in video frames, aligning with our research objectives.

Limitations include manual labeling and dataset diversity. Future work should explore automated annotation, diverse datasets, and hyperparameter tuning for model improvement.

Our study presents a promising CNN-based people detection model with practical implications. Addressing limitations and further refinement can enhance its applicability and accuracy.

The results validate the efficacy of the proposed model in achieving accurate and efficient crowd counting while overcoming challenges posed by occlusions, scale variations, and dynamic scenarios. Integrating YOLOv8, the COCO dataset, and region-based processing forms a potent synergy, addressing limitations observed in traditional crowd-counting methods. The practical implications extend to crowd management, event planning, and public safety, underscoring the model's potential to contribute to decision-making processes and resource optimization.

V. Conclusion

Our research presents a significant advancement in real-time people detection, encompassing two distinct models: CNNs and YOLOv8. These models offer practical applications in surveillance, crowd management, and autonomous navigation by providing enhanced accuracy and security measures. Future work should focus on addressing limitations like manual labeling and dataset diversity.

The fusion of YOLOv8, the COCO dataset, and region-based processing has produced an advanced crowd-

counting model. This model excels in object detection, tracking, and precise people counting within specific regions, showcasing its robustness even in dynamic scenarios. It minimizes false positives and ensures accurate people counting by leveraging the COCO dataset's comprehensive annotations and YOLOv8's capabilities.
With real-time efficiency as a hallmark, our model demonstrates practical utility in crowd management, event logistics, and public safety, making it a valuable asset in various domains. This innovative synthesis of cutting-edge technologies contributes to the field's advancement and facilitates informed decision-making.

VI. REFERENCES

[1] Jialue Fan, Wei Xu, Ying Wu, & Yihong Gong, (2010), Human Tracking Using Convolutional Neural Networks, IEEE Transactions on Neural Networks, 21(10), 1610–1623.

[2] Mysha Sarin Kabisha, Kazi Anisa Rahim, Md. Khaliluzzaman, Shahidul Islam Khan. Face and Hand Gesture Recognition Based Person Identification System using Convolutional Neural Network. IJISAE, 2022, 10(1), 105–115.

[3] Ujwalla Gawandea, Kamal Hajarib, Yogesh Golhar, Real-Time Deep Learning Approach for Pedestrian Detection and Suspicious Activity Recognition, Procedia Computer Science 218 (2023) 2438–2447.

[4] Mubin Modi, Zaid Marouf, Anvay Wankhede, Dr. Shabina Sayed (2021) A Real-time Crowd Counting application with Live analytics and selective detection using YOLOv4(ISSN-2349-5162).

[5] Ansari, M.A., Singh, D.K. Human detection techniques for real-time surveillance: a comprehensive survey. *Multimed Tools Appl* 80, 8759–8808 (2021)10.1007/s11042-020-10103-4.

[6] Szarvas, M., Yoshizawa, A., Yamamoto, M., & Ogata, J. (2005). *Pedestrian detection with convolutional neural networks. IEEE Proceedings. Intelligent Vehicles Symposium, 2005.* doi:10.1109/ivs.2005.1505106

[7] Sambolek, S., & Ivasic-Kos, M. (2021). *Automatic Person Detection in Search and Rescue Operations Using Deep CNN Detectors. IEEE Access, 9, 37905–37922.* doi:10.1109/access.2021.3063681

[8] Nogueira, V., Oliveira, H., Augusto Silva, J., Vieira, T., & Oliveira, K. (2019). *RetailNet: A Deep Learning Approach for People Counting and Hot Spots Detection in Retail Stores. 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI).* doi:10.1109/sibgrapi.2019.00029

[9] Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., & Mekhtiche, M. A. (2020). *Hand Gesture Recognition for Sign Language Using 3DCNN. IEEE Access, 8, 79491–79509.* doi:10.1109/access.2020.2990434

[10] Dillon Reis*, Jordan Kupec, Jacqueline Hong, Ahmad Daoudi(2023), *Real-Time Flying Object Detection with YOLOv8,Cornell University,* arXiv:2305.09972 [cs.CV]

[11] Nilesh Parmanand Motwani, Soumya S, *Human Activities Detection using DeepLearning Technique- YOLOv8*, ITM Web of Conferences 56, 03003 (2023), ICDSAC 2023, https://doi.org/10.1051/itmconf/20235603003

[12] Zheng Wang, Dong Xie, Hanzhi Wang, Jiang Tian, *An Effective Two-stage Training Paradigm Detector for Small Dataset* (2023), arXiv:2309.05652

[13] D. Al-obidi and S. Kacmaz, "Facial Features Recognition Based on Their Shape and Color Using YOLOv8," 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkiye, 2023, pp. 1-6, doi: 10.1109/ISMSIT58785.2023.10304905.

[14] J. Telicko and A. Jakovics, "Comparative Analysis of YOLOv8 and Mack-RCNN for People Counting on Fish-Eye Images," 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Canary Islands, Spain, 2023, pp. 1-6, doi: 10.1109/ICECCME57830.2023.10252265.

[15] T. -Y. Chen, C. -H. Chen, D. -J. Wang and Y. -L. Kuo, "A People Counting System Based on Face-Detection," 2010 Fourth International Conference on Genetic and Evolutionary Computing, Shenzhen, China, 2010, pp. 699-702, doi: 10.1109/ICGEC.2010.178.