Integrating YOLOv8 and IoT in a Computer Vision System for Child Detection in Smart Cities

Modhawi Alotaibi*, Atheer Alruwaythi, Sara Alenazi, Maisaa Alsaedi
College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia

Abstract—**In an era marked by technological advancements aimed at establishing smart cities, technology increasingly focuses on enhancing aspects related to crowd management. The widespread deployment of CCTV systems, combined with the integration of computer vision, has enabled accurate insights into crowd density estimation. Our research highlights the potential benefits of child detection across various domains that serve governments and business decision-making. Leveraging Internet of Things (IoT) devices to collect real-time data and employing artificial intelligence (AI) based on deep learning through computer vision is powerful in such domains. In this paper, we propose an IoT architecture that facilitates intelligence and decision-making in two phases: 1) a deep learning model with object detection and image segmentation capabilities using YOLOv8, and 2) a tracking/counting algorithm for estimating child density based on DeepSORT. Our implementation efficiently identified and classified children in extracted images with an accuracy rate of up to 98%. Also, our model outperformed the other two solutions proposed by previous studies in terms of mAP@50, Precision, and Recall metrics. The results provide valuable insights for businesses aiming to refine site selection and guide governments in improving urban planning and safety, thereby fostering sustainable and intelligent urban development.**

Keywords—Computer vision; Internet of Things; deep learning; YOLOv8; DeepSORT

I. Introduction

With the widespread deployment of Closed-circuit Television (CCTV) surveillance systems, computer vision plays a vital role in various applications based on crowd and flow density analysis. In an era characterized by technological shifts to smart cities, there is a focus on developing and enhancing aspects related to crowd management. One interesting application is the use of crowd analysis for child detection. The research field of child detection can be beneficial in various domains, such as advertising and business site selection. Governments can also gain significant advantages from child detection in domains such as public safety and law enforcement, as well as from collecting demographic information accurately for enhanced urban planning. Child detection can play a crucial role in addressing critical safety issues, such as supporting efforts to locate lost children [1, 2] and monitoring traffic to detect front-seat child occupancy in vehicles [3]. Additionally, governments can deploy child detection for smart urban planning to optimize site selection for various reasons [4], including areas targeted towards children (e.g. child centers, schools, and parks).

Despite the growing interest in crowd analysis, there remains a notable research gap in the accurate detection and tracking of children within dynamic urban environments using integrated IoT and AI technologies. Existing studies often focus on general crowd density, leaving child-specific analysis underexplored. This paper aims to bridge that gap by proposing a specialized framework for child detection and density estimation.

Distinguishing children from adults in digital images has become increasingly important for various applications, such as adjusting electronic device controls and tailoring advertisements [5], as well as informing business decisions. In today's competitive business landscape, selecting the right location is a make-or-break decision for new ventures. Businesses can utilize crowd and flow density analysis as a strategic method to gain valuable insights by accurately estimating foot traffic patterns and behaviors in a given area. Consequently, entrepreneurs can determine the optimal area for their target customers (including children), ultimately leading to enhanced commercial outcomes.

Boosting economic growth and safety through crowd data analysis to identify optimal areas for targeted audiences or enhancing the human experience lays the foundation for sustained growth and long-term success [6]. Emerging technologies like IoT and AI can play vital roles in such decision-making processes. To demonstrate the impact of crowd analysis, specifically child detection, on enhancing smart urban planning, we integrate IoT, computer vision, and deep learning technologies, focusing on areas where children are the primary beneficiaries.

IoT has become pervasive and intuitive in the technical, economic, and social fields of smart cities. It works by incorporating internet connectivity with powerful data analysis capabilities [7]. IoT deploys sensing devices to collect data and then transforms raw data into actionable insights. This process requires sensors that generate data to be collected, analyzed, and visualized to obtain useful information [8]. The data is then interpreted to produce decisions in a deliverable form [9]. The massive volume of data produced by IoT devices highlights the importance of AI, which efficiently manages and processes this vast data [10].

AI is a field of computer science that studies technologies allowing computers to intelligently simulate human cognitive functions [11]. AI encompasses various components, including deep learning, a subset of machine learning (ML) that enables machines to learn by analyzing inputs using neural networks inspired by the structure of the human brain's neurons. These

*Corresponding Author.

networks consist of layers of interconnected neurons arranged in a specific pattern and are specialized for processing a grid of values. Cognitive computing emulates human cognitive abilities, including sensing, perception, learning, decision-making, and action When integrated with IoT, cognitive computing can enhance surveillance systems to enable automated and intelligent crowd management, which understands context better and responds efficiently [12].

In the context of crowd analysis, leveraging IoT and AI, primarily through computer vision, enables accurate crowd detection and analysis. Computer vision techniques help analyze surveillance videos to extract valuable insights. It utilizes cameras, data, and algorithms to process information quickly and efficiently [13]. Computer vision techniques provide solutions to current challenges in crowd analysis, such as tracking accuracy and density estimation [14]. Our research discusses how these techniques detect children within crowd videos, which can benefit various applications across multiple domains.

Our main contribution in this paper is leveraging AI tools, IoT, and computer vision to improve urban site selection decisions for child-targeted areas. By adopting IoT devices such as cameras, we collect vast amounts of real-time data from various locations and process it with advanced AI techniques to accurately interpret crowd characteristics and behaviors. Computer vision plays a crucial role in this process, providing precise visual data analysis and applying tracking algorithms that detect, track, and estimate child density in specific areas. We propose leveraging these technological advancements to make smarter decisions, emphasizing the significant impact of our study on optimizing urban planning for sites targeted toward children. This work has significant implications for governments, urban planners, and businesses, enabling data-driven policies, enhanced safety measures, and improved infrastructure planning.

In addition to the introduction in Section I, this paper offers five more sections. Section II reviews related work on deploying IoT and intelligent methods, such as computer vision and deep learning in the crowd analysis domain. Section III describes the proposed IoT architecture. In Section IV, we discuss in detail the methodology. We present the results of our proposal in Section V. Finally, Section VI concludes the paper.

II. Related Work

In this section, we review other studies relevant to our research. The domain of crowd analysis has been addressed from various angles using different enabling tools. One primary tool is IoT. Solmaz et al. [15] presented a study on managing crowd mobility in smart cities using IoT technologies. They introduced the federated interoperable semantic IoT (FIESTA-IoT) platform, which enables information sharing among stakeholders for efficient crowd management. They addressed challenges such as the scalability of connected systems, information transparency between different systems, data federation, and information privacy. In another study, Varghese et al. [12] discussed the application of cognitive computing for smart crowd management in a surveillance system. The authors showcased two case studies: smart crowd management during Hajj (an Islamic mega event) and smart crowd management during the Tour de France (an annual bicycle race). Their IoT-based intelligent crowd-management system utilized mobile sensor data along with video data for efficient monitoring. They employed various AI algorithms, such as deep learning, to analyze IoT big data for real-time insights. In the context of mega gatherings, crowd data should be collected and utilized for better management. The authors [16] proposed an intelligent IoT approach for crowd management during Hajj to avoid congestion. Their solution was built based on the characteristics of Mina (a holy area in Makkah). In this proposed approach, they constructed the IoT sensing layer using Radio Frequency Identification (RFID) cards and surveillance cameras. Then, based on calculated risk factors, a routing algorithm guides pilgrims through safe paths. However, this proposal lacks the use of AI tools, which could elevate the solution and its outcomes. Mohd et al. [8] discussed current IoT research technologies and applications for serving Hajj and Umrah pilgrimages based on crowd analysis. They proposed practical suggestions that could employ technologies to facilitate services provided to pilgrims, Umrah performers, visitors, and service providers. The first service they suggested was a smart-screen device to prevent cases of lost pilgrims and Umrah performers. The second was a monitoring system with automated alerts to detect and respond quickly to unusual events. Third, they proposed a smart parking system using devices at each parking lot to help people locate their vehicles via plate numbers and interactive maps.

A second enabling tool is computer vision in conjunction with other technologies. Li et al. [17] examined crowd density in a tourist area using video management dynamic information analysis. They proposed a counting model that could detect crowd density through cameras. Additionally, they developed inverse-scale perception modules to support the extraction of multi-scale information. They used their model to generate a distribution-density map, counting the population based on convolutional neural networks (CNN). The authors of [18] proposed a framework that integrates CCTV video footage and spatial information, using deep learning models, geometric transformations, and tracking algorithms to extract data on crowd congestion. This approach—using deep learning techniques to accurately analyze and process visual information—allows real-time monitoring and prediction of crowd congestion, assisting urban planners and infrastructure operators in managing crowd congestion and improving public safety. Besides deploying crowd analysis for detecting abnormalities, it can also be used for flow prediction. Interestingly, Zhou et al. [19] focused on crowd-flow prediction for planned transportation sites, a critical challenge for urban planners and administrators. They offered a data-driven technique for predicting future population flows for newly planned sites.

In the domain of child safety, numerous studies have focused on techniques for child detection in images [2,3,20]. For instance, Basaran et al. [2] proposed the ChildSafe system, a solution for classifying individuals as children or adults using 3D depth cameras, aimed at enhancing child safety applications. The methodology involved collecting data from 193 participants, including 150 children and 43 adults aged 13 to 50, using Microsoft Kinect sensors. Participants performed eight predefined actions, and the system extracted body lengths, facial metrics, and relative ratios to mitigate errors from absolute measurements. A bin-boundary-based classifier was applied to process these features, mapping them into discrete bins and calculating weights based on the likelihood of feature occurrences. The system achieved a high classification accuracy of up to 97% with low error rates. However, the solution faces challenges, particularly its reliance on controlled settings with predefined actions, which limits its effectiveness in real-world environments with dynamic lighting and unscripted behaviors. Balci et al. [3] presented a study on an automated system for detecting child occupancy in the front seat of vehicles on roadways, addressing a critical safety issue in traffic enforcement. They employed deep learning techniques to analyze images captured from an overhead gantry, achieving an overall accuracy of 90%. The methodology includes various approaches for face detection and passenger classification.

In other studies, targeting social security and crime prevention applications, the authors of [21,22] focused on classifying pedestrians as children or adults using biometric features from video data. Both approaches used the head-to-body length ratio for classification but differed in methodology, accuracy, and how they addressed challenges. In [21], a method for classifying pedestrians as children or adults using biometric features derived from long-distance video data was proposed. The study employed Haar-like features with Adaboost for full-body detection and utilized static thresholds for head-to-body ratio-based classification. The method achieved 74.7% accuracy for adults and 68.1% for children. However, challenges arose from inconsistent head detection, especially when head orientations deviated from the frontal view, and the reliance on static thresholds limited its adaptability in dynamic scenarios. Similarly, the authors of [22] proposed a framework that utilized Haar-like features with Adaboost for full-body detection and Local Binary Patterns (LBP) for head detection. This study introduced a moving average algorithm to dynamically adjust classification thresholds, addressing variations in pedestrian positions. The approach achieved 100% accuracy for children and 64.5% for adults, with lower accuracy for adults attributed to deviations in body proportions. On the other hand, [23] presented a method for detecting underage individuals using facial image classification. It utilized a dataset categorized into three age groups: children, adults, and the elderly. The methodology employs two main feature extraction techniques—Histogram of Oriented Gradients (HOG) and Haar features—to analyze facial characteristics, including wrinkles, geometric features, and facial ratios. The implementation uses the Viola-Jones algorithm for face detection, processing real-time images at two frames per second, making it suitable for various real-world applications such as online content restriction and age-age-restricted purchases.

One major benefit of crowd analysis is the ability to detect and track objects. In [24], the authors introduced DeepSORT, a tracking system that uses a neural network to enhance object tracking by combining motion and appearance information. DeepSORT aims to improve tracking accuracy by reducing identity switches and prioritizing frequently seen objects. DeepSORT achieves excellent results in tracking multiple objects and performs well across various scenarios, including our research focus on smart urban site selection for children's areas.

Selecting an appropriate site for a particular purpose is one of the many applications of crowd and flow analysis [25–28]. The purposes vary to serve different sectors, such as education, transportation, agriculture, entertainment, business, and health. Focusing on the concept of site selection from a business perspective, many authors strongly connect location selection to business success [29–31]. In [29], the author emphasized that investors must evaluate new and existing locations to determine the optimal sites for starting a business. Akpan et al. [30] studied the impact of retail outlet location and accessibility on business performance. They gathered data from businesses through a structured survey of 90 of 120 retail outlets considering functioning in an area. They analyzed the collected data using two statistical techniques: Structural Equation Modelling (SEM) and Relative Importance Index (RII). SEM is used to study the connection between retail location and profitability, while RII ranks the elements according to each factor's value in relation to other factors impacting company performance. In another study, Yakovlev et al. [31] examined the business-site-selection problem, explicitly addressing the maximum coverage location problem (MCLP). The MCLP is an optimization challenge that assists businesses in identifying the most suitable locations for their facilities to maximize service coverage. The study focused on the continuous formulation of the MCLP, which considers the arbitrary spatial shape of the demand zone (where services are needed) and the service areas (where facilities will be placed). Due to limitations in processing spatial and geometric data for the demand area, the demand zone is represented as a set of points. The study involved developing mathematical models and approaches to solve the planar MCLP using modern computer technologies. The authors emphasized the importance of considering the size and shape of the demand zone and service areas in the decision-making process for optimal facility placement.

The reviewed literature highlights substantial advancements in crowd management, child detection, and site selection through IoT, computer vision, and deep learning. However, several limitations persist. Many crowd management systems lack robust AI integration, reducing their adaptability and predictive capabilities in dynamic environments. Child detection methods often depend on controlled settings or static thresholds, which limit their effectiveness in real-world scenarios characterized by variable lighting, movement, and behavior. Tracking algorithms like DeepSORT demonstrate strong performance in object tracking, yet few studies have applied them specifically in child density estimation for urban planning. Additionally, while site selection models emphasize business optimization, they rarely incorporate demographic-specific crowd insights—such as child presence—which could significantly enhance planning for child-focused infrastructure [29–31].

III. IoT Architecture

In this section, we present our proposal: the IoT architecture, which consists of four layers, as shown in Fig. 1.

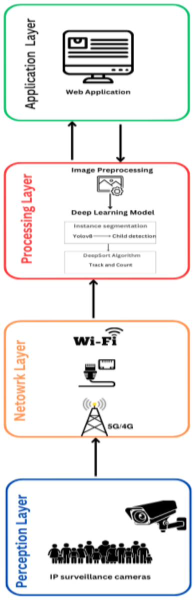


Fig. 1. IoT architecture.

First, the perception layer serves as the base-sensing layer, consisting of sensors that capture real-time environmental data and transmit it to the network layer via various communication protocols. In our work, we exploit the fact that surveillance cameras are widely deployed in public spaces, utilizing them as sensing devices to obtain video frames.

Second, the network layer functions as a communication bridge, connecting layers through internet protocols and network technologies like Wi-Fi, cellular networks, and Ethernet. Wi-Fi offers secure, reliable, and high-speed wireless connectivity, while cellular networks provide wide-area connectivity, supporting high data rates for IoT devices. Ethernet is a wired network that suits real-time applications due to its security and reliability, serving as a backbone network for interconnecting IoT gateways [32]. These features are crucial for sensors in IoT systems. In this work, data are collected from the sensors/cameras and transmitted to the processing layers via the network layer.

Third, the data processing layer processes the collected data from the perception layer through the network layer. This layer is where the intelligence resides, intertwining computer vision and deep learning technologies. Initially, our model preprocesses the frames extracted from the cameras and transmits them to our proposed model. The model then performs analysis and estimation by detecting children, tracking, and counting them accurately. The results generated from the model are then passed to the application layer for practical use.

Finally, the application layer presents the results to the end user through an interactive user interface. In our envisioned application, the results are displayed as a heatmap that ranks different sites based on varying child densities using easy-to-interpret color-coded visualizations.

IV. Methodology

In this section, we outline our methodology to demonstrate how the proposed system was designed, implemented, and evaluated. We started with deploying AI and computer vision tasks, including classification, object detection, and instance segmentation to achieve our objectives. We elaborate on our proposed model as well as other deep learning models in terms of model setup, building, and evaluation. We examine four You Only Look Once (YOLOv8) models, and the highest-performing model is compared against other proposed solutions [1, 39]. We then discuss the tracking and counting mechanism we adopt for estimating child density in crowds.

A. Deep Learning Model Selection

In this part, we start by showcasing the three main elements of any deep learning model for detection: setup, building, and evaluation.

1) Setup: Initially, we searched for a suitable dataset by exploring various open-source platforms. We found a dataset on Kaggle that contained 800 pictures [33]. However, it was not ideal as it did not meet our model's requirements and yielded poor results. Consequently, we created a new dataset by collecting data from public places and live streaming videos from a YouTube channel named "The Real Samui Webcam" [34]. Finally, we cleaned the data and extracted the frames that met our model's requirements.

After collecting the data, we began preprocessing it for the model. First, we classified the dataset into "child" and "non-child," as our focus lies between these two classes. Subsequently, we utilized Roboflow to apply data augmentation techniques, such as resizing images to 640 × 640 pixels and grayscale conversion for 15% of the images to help the model generalize across different lighting conditions. To further enhance the model's adaptability, we increased the saturation of the images by up to 30%, allowing it to learn effectively in various lighting scenarios throughout the day. Finally, to ensure data robustness, we split the dataset into 80% for training, 10% for validation, and 10% for testing.

2) Model building: Our strategy was to deploy different types of models to ensure optimal results. We began with a traditional CNN and applied the classification task among the different layers. However, the results were unsatisfactory. We then shifted our focus to transfer learning, a machine learning approach that uses previously acquired knowledge to solve a related problem. We searched for a suitable pretrained model and chose YOLO, a real-time object detection system [35]. YOLO is a widely used object detection and image segmentation model known for its speed and accuracy [36]. However, it has many versions, and we opted for YOLOv8, which introduced new features and improvements to enhance performance, flexibility, and efficiency.

YOLOv8 allows for multiple computer vision tasks, such as object detection and instance segmentation. We began with object detection and annotated the images in the dataset with bounding boxes. Then we imported our dataset into the model. Finally, we tested our model to assess the outcomes and discovered that they needed improvement. We attempted to enhance the results by increasing the number of images in the dataset, but the results remained unsatisfactory.

Delving deeper into YOLOv8, we decided to experiment with other computer vision tasks. The first was semantic segmentation, which assigns a class label to pixels using a deep learning algorithm. It is one of three subcategories in the overall process of image segmentation that helps computers understand visual information. However, after trying it, we found that the task's process could not accomplish our goals because it aims to accurately classify each pixel in an image according to the relevant object category, meaning it cannot distinguish between distinct instances of a single object class [37]. Therefore, we decided to try the second task, instance segmentation. This method identifies and separates individual objects within an image, including detecting the boundaries of each object, assigning a unique label to each object, and distinguishing between distinct instances of a single object class. Due to their different performance, we precisely tuned two types of YOLOv8-seg—YOLOv8s-seg and YOLOv8x-seg—to choose the best performer among them. YOLOv8s-seg is smaller and used for faster inference, containing 261 layers. In contrast, YOLOv8x-seg is a larger, more complex model designed to achieve higher segmentation accuracy, with 401 layers, making it slower in inference speed and more costly. By incorporating it into our process, we achieved the best results.

For the training process, we unified the hyperparameters as shown in Table I. The epochs parameter is set to 100, defining the maximum number of iterations during training. The learning rate (lr0) is set to 0.00001, influencing the rate at which the model updates its parameters based on the training data. The batch size is set to 16, determining the number of training samples processed before the model's internal parameters are updated. Using these configurations along with other hyperparameters, we initiated training, ensuring a consistent and effective evaluation process.

TABLE I. HYPERPARAMETERS SETTINGS

Number of Iterations	100 Epochs
learning rate	0.00001
batch size	16
image size	640 × 640 pixels

3) Evaluation: We built multiple models to achieve optimal performance. Table II showcases the diverse results of five model variations in terms of four performance metrics. The mAP@50 is a metric used to evaluate the performance of object detection and instance segmentation models. It measures how well a model can identify and localize distinct objects within an image. The mAP@50 is calculated as the mean of the average precision values across all object classes in the dataset, providing a comprehensive measure of a model's ability to detect and classify distinct object types accurately. Precision, measures the accuracy of the model's positive predictions; therefore, the higher the value, the better the result, indicating that the model accurately detects the object within its correct class and avoids misclassification. Lastly, the recall metric refers to the ratio of correct values predicted relative to all correct values, clarifying the model's completeness in identifying the most relevant instances. As for the models, we first built a CNN with a classification architecture in which the nm3467output is a set of probabilities for each class (child and non-child), with the highest probability indicating the predicted class within a separate frame.

TABLE II. MODELS COMPARISON

Model	Accuracy	mAP@50	Precision	Recall
CNN Classification	0.24	-	0.537	0.598
Object Detection by (YOLOv8s-detect)	-	0.911	0.821	0.873
Object Detection by (YOLOv8x-detect)	-	0.903	0.837	0.907
Instance Segmentation by (YOLOv8s-seg)	-	0.985	0.967	0.954
Instance Segmentation by (YOLOv8x-seg)	-	0.99	0.972	0.988

However, it does not detect two classes in one frame, leading to inaccurate results. Moreover, as shown in Table II, it has poor accuracy and precision of 0.24 and 0.537, respectively. The recall of 0.598 indicates that the model cannot accurately detect and distinguish the classes. In contrast, the object detection model is commonly used for tasks involving detecting multiple classes of objects in the same frame. Thus, we decided to build a model based on object detection. However, detecting objects with a bounding box showed better results than CNN. We tried multiple pretrained models for YOLOv8 object detection that Ultralytics offers [38]. We started with a small pretrained model (YOLOv8s-detect), which is lightweight and makes the trained model fast and efficient due to its limited computational resources. The result was still not as expected because the boxes could not precisely detect the full individual body shape in all positions. The mAP@50 was 0.911, precision increased to 0.821, and recall reached 0.873. Subsequently, we tried the extra-large model (YOLOv8x-detect), which is slower than the small model but more accurate and complex, as it has more layers and achieves higher performance. The result from the YOLOv8x-detect for the mAP@50 was 0.903, and precision and recall increased to 0.837 and 0.907, respectively, compared to the previous model.

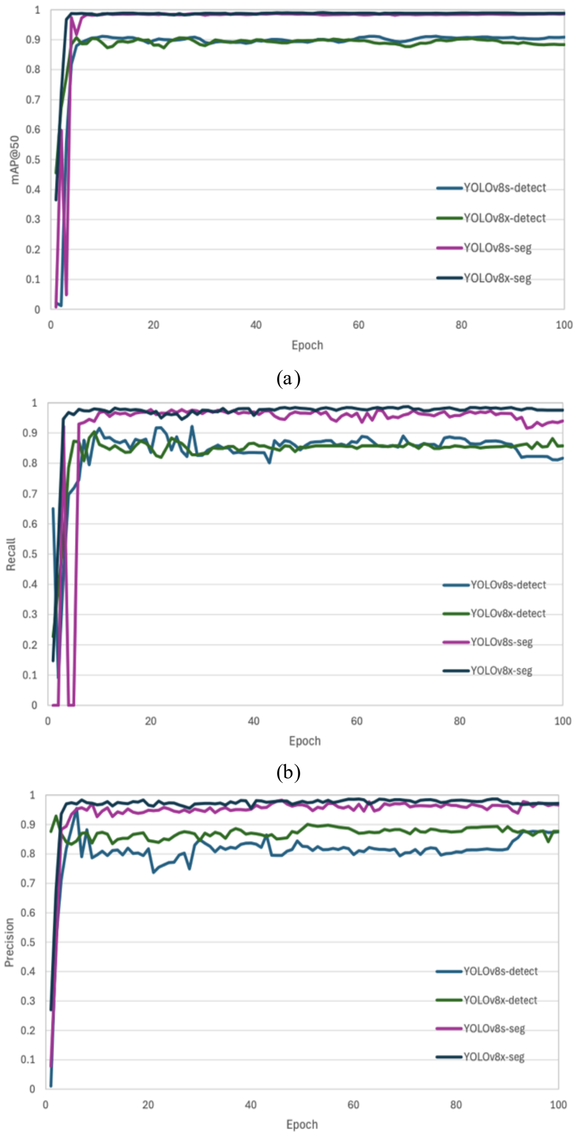
To explore deeper knowledge and acquire more accurate results, we tried another computer vision task that assigns a class label to pixels using a deep learning algorithm, which is instance segmentation. Instance segmentation provides fine-grained localization, enabling precise differentiation between individuals in various positions. This method provides detailed and accurate information, making it easier to distinguish between different objects within an image.

We also tried multiple pretrained models for YOLOv8 instance segmentation. First, we started with a small pretrained model (YOLOv8s-seg), which is lightweight and makes the trained model fast and efficient. Fortunately, the metrics achieved drastic improvements compared to the previous models, with the mAP@50 increasing to 0.985, and precision and recall reaching 0.967 and 0.954, respectively. Then we moved to the extra-large model (YOLOv8x-seg), which is more accurate and complex. The results of the YOLOv8x-seg model yielded the best results, with a mAP@50 of 0.99 and a precision of 0.972. The recall showed the most substantial improvement, reaching 0.988, confirming that the extra-large model outperformed all prior variations.

Fig. 2 depicts the performance evaluation of the four YOLOv8 model variations over a hundred epochs based on three key metrics—precision, recall, and mAP@50. Generally, the YOLOv8x models demonstrated more consistent and reliable performance compared to the YOLOv8s models across both the object detection and instance segmentation tasks. The advantage of YOLOv8x is its architectural design. The YOLOv8x model has more layers and a larger set of parameters compared to the YOLOv8s model. This increased complexity enables the YOLOv8x model to learn more effectively, allowing for a better understanding of the underlying data. However, it is important to consider that these complex computations consume more resources. For that reason, we initially used the YOLOv8s model, which proved to be a solid alternative.

YOLOv8s-detect and YOLOv8x-detect models achieve precision, recall, and mAP@50 around 0.7 to 0.8, 0.6 to 0.8, and 0.8 to 0.9, respectively. Among the two, YOLOv8x-detect's results were much closer but smoother. On the other hand, YOLOv8s-seg and YOLOv8x-seg achieved precision, recall, and mAP@50 up to 0.9 in all metrics. Moreover, the YOLOv8x-seg exhibited a consistent level without significant fluctuations. Overall, the segmentation model YOLOv8x-seg surpasses the detection models by achieving higher and more stable levels across all metrics, making it more reliable for our tasks, which require precise object identification to distinguish between child and non-child.

Fig. 3 depicts the relationship between precision and confidence during validation while comparing YOLOv8x-detect, YOLOv8s-detect, YOLOv8s-seg, and YOLOv8x-seg models. Initially, YOLOv8x-detect exhibits a precision of approximately 0.38, which progressively ascends, stabilizing around 0.8, and later increases to approximately 0.92. YOLOv8s-detect starts with a precision range of 0.2 to 0.3. This model then demonstrates a consistent rise to approximately 0.8 to 0.89. There is an apparent rise in precision to the range of 0.8 to 0.9, followed by a minor drop and a subsequent clear increase to around 0.95. From these observations, we infer that YOLOv8s-detect has higher confidence values but slightly lower precision compared to YOLOv8x-detect. In contrast, YOLOv8s-seg surpasses both detection models in precision and confidence, starting at approximately 0.56 and steadily increasing to nearly 0.88, followed by a gradual increase to nearly 0.98. Finally, YOLOv8x-seg outperformed all other models, particularly in terms of precision. It started at approximately 0.8, followed by a gradual increase to 0.9, stabilizing between 0.85 and 0.95, and then ultimately reaching a precision of nearly one.



(a)

(b)

(c)

Fig. 2. YOLOv8 metrics evaluation: (a) mAP@50, (b) Precision, (c) Recall.

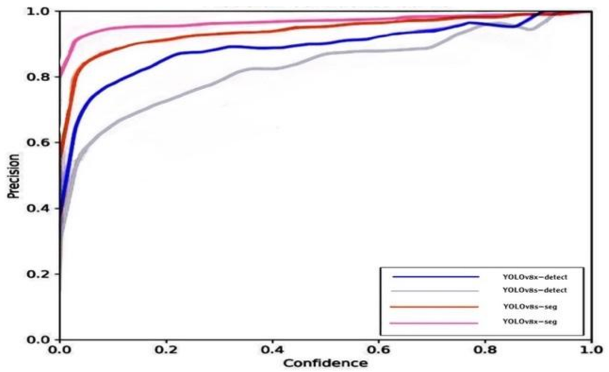


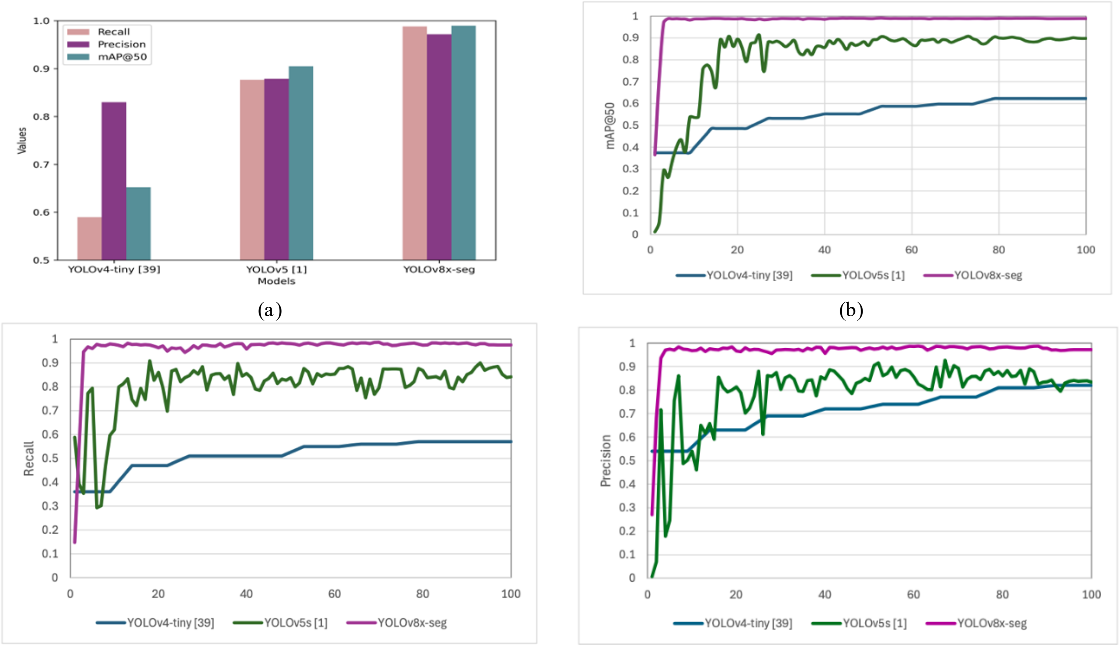
Fig. 3. Precision and confidence in validation.

4) Performance comparison with other YOLO-based models: We compare our highest achieving model (YOLOv8x-seg) as explained previously to other solutions proposed by Tahir et al. [1] and Lin et al. [39]. Fig. 4 compares the performance of these three YOLO-based models: YOLOv4-tiny [39], YOLOv5s [1], and YOLOv8x-seg.

Subfigure (a) illustrates the overall comparison across the mAP@50, Precision, and Recall metrics, while (b–d) provide a detailed view of each metric. Among the models, YOLOv8x-seg achieves the best performance, with a mAP@50 of 0.99, Precision of 0.972, and Recall of 0.988. These exceptional results are attributed to its advanced architectural design and segmentation optimization, which enhance both accuracy and detection capabilities.

The YOLOv5s [1] model also demonstrates high performance, achieving a mAP@50 of 0.905, Precision of 0.879, and Recall of 0.877. While it performs well overall, it remains slightly behind YOLOv8x-seg. Additionally, its effectiveness decreases under suboptimal camera angles, limiting its ability to detect objects accurately in certain scenarios. In contrast, YOLOv4-tiny [39] delivers the lowest scores, with a mAP@50 of 0.652, Precision of 0.83, and Recall of 0.59. These results reflect its performance limitations, especially in challenging tasks such as detecting small objects like babies in strollers, where it struggles to generate precise bounding boxes.

Overall, YOLOv8x-seg stands out as the most advanced and accurate model, making it the ideal solution for applications where performance and precision are critical.



(a)

(b)

(c)

(d)

Fig. 4. Comparison of three YOLO-based models: (a) All metrics, (b) mAP@50, (c) Recall, (d) Precision.

B. *Tracking and Counting Algorithm*

The process of tracking objects across multiple frames is known as object tracking. In object tracking, a unique ID is assigned to each detected object and tracked across multiple frames. Object tracking has gained popularity with the evolution of computer vision and deep learning techniques [40]. Once we successfully applied our model and detected children in an area, we used object- tracking algorithms to follow them. We employed the highly effective DeepSORT algorithm [24], which uses the Kalman filter to track the position and motion of objects over time. The DeepSORT algorithm keeps track of each detected child as they move through the scene. Additionally, several parameters play a critical role in optimizing the tracking and counting process. We set these parameters to achieve optimal results. One such parameter is “n_int” which specifies the minimum number of consecutive frames a child must appear in to be considered for tracking and assigned a unique ID. We set this value to three, meaning that a child must be present in three consecutive frames to ensure its correct presence in the scene before being tracked and assigned an ID.

Another important parameter affecting tracking and counting is “max_age”. It defines the maximum number of frames a child can miss before its ID is considered lost and subsequently removed. We set max_age to seventy frames, as it provides a good balance between maintaining persistent tracks and quickly removing children who have left the scene. Finally, after the tracking process is initiated, we count number of the unique IDs of the detected children. If the total count in a given area is high, it is considered a child-crowded area.

V. RESULTS

The implementation had two phases. The first phase involved instance segmentation, which utilized the YOLOv8x-seg model.

We evaluated the model’s performance on the training and validation datasets. As a result, the model successfully identified and classified various objects in the images with high accuracy. Fig. 5 illustrates the experimental results achieved by the model using three metrics: precision, recall, and mAP@50. We also evaluated the model's performance on a new set of images outside the training set. Remarkably, our model demonstrated high accuracy, indicating that it was well-trained and capable of recognizing and differentiating between children and non-children in images. Fig. 6 represents the confusion matrix, a fundamental tool for evaluating performance. It provides a detailed representation of the model's predictions compared to the actual data. We used the confusion matrix to calculate crucial performance metrics such as precision and recall. We concluded that utilizing the YOLOv8x-seg model was the right decision, as it is designed for the instance segmentation task, enabling accurate object detection and unique label assignment, which helps with object counting.

The second phase involved tracking and counting the detected objects. Based on the results obtained from our model, we incorporated the DeepSORT tracking algorithm into our implementation. It enables us to track the children after the model detects them. The results appear as a bounding box around each child, with the track ID and confidence number. Each time a child is detected, the count increases by one. The count, along with the camera's geographical location, is then recorded. These locations near each other represent an area. If the total count in an area is high, it is considered child-crowded and thus optimal for planning child-centric sites such as schools, centers, parks, etc. This approach provides valuable insights into the density and distribution of children within a monitored area.

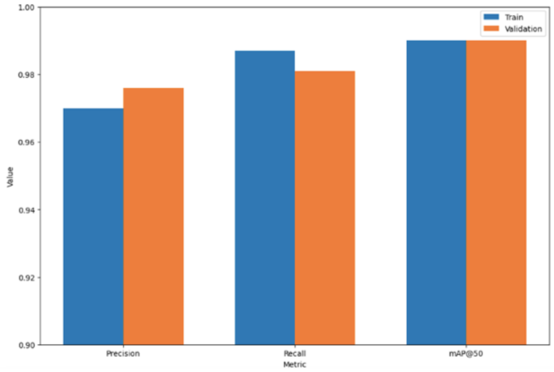


Fig. 5. Metrics comparison of training and validation.

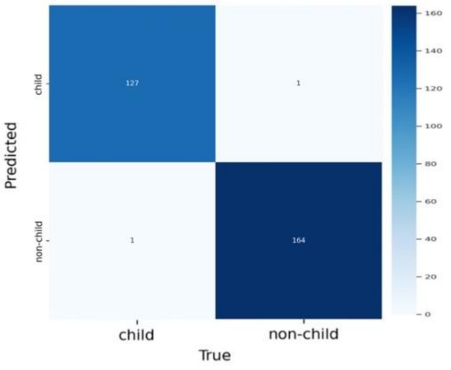


Fig. 6. Confusion matrix.

VI. Conclusion

In a time shaped by technological advancements aimed at improving the quality of life in smart cities; modern tools and techniques have become prevalent across various domains. This study targets improving site selection, assisting entrepreneurs in identifying optimal locations for their businesses while enabling governments to promote economic growth and safety. We propose integrating IoT, YOLOv8, and computer vision techniques to detect, track, and count children in crowds. This paper not only demonstrates improved site selection for entrepreneurs but also highlights broader applications, such as enhancing public safety, improving demographic data collection, and supporting smart urban planning. Our findings underscore the significant impact of our study on urban planning initiatives targeting children. Our contributions illustrate the transformative potential of IoT and AI in addressing societal challenges and fostering economic growth by providing actionable insights for child-focused site optimization. Consequently, we propose an IoT architecture that collects frames from surveillance cameras and processes them using deep learning and computer vision techniques.

Our primary contribution revolves around the IoT processing layer, implemented in two phases: the YOLOv8x-seg model and the DeepSORT tracking algorithm. Our implementation accurately identifies and classifies children in images with a precision rate of 98%. Utilizing emergent techniques such as IoT and deep learning can yield valuable insights that assist businesses and governments in making informed decisions. However, it is crucial to consider the ethical implications, particularly when dealing with sensitive data related to child tracking and monitoring. Although we propose a model for predicting optimal areas for child-centric sites, its deployment must operate within a regulated governance framework that ensures data privacy. Ultimately, the AI-driven system for defining optimal areas for governments and businesses represents a powerful tool that aligns technology utilization with community needs.

References

[1] Tahir, A, Khalid, S.K.A, Fadzil, L.M. "Child Detection Model Using YOLOv5". J. Soft. Compute. Data Min. 2023, 4, 72–81.

[2] Basaran, C, Yoon, H.J, Ra, H.K, Son, S.H, Park, T, Ko, J. "Classifying children with 3D depth cameras for enabling children's safety applications". In proc. of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13–17 September 2014; pp. 343–347.

[3] Balci, B, Alkan, B, Elihos, A, Artan, Y. "Front seat child occupancy detection using road surveillance camera images". In proc. of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 1927–1931.

[4] Kaiser, M.S, Lwin, K.T, Mahmud, M, et al. "Advances in Crowd Analysis for Urban Applications Through Urban Event Detection". IEEE Trans. Intell. Transp. Syst. 2018, 19, 2518–2531. https://doi.org/10.1109/TITS.2017.2771746.

[5] Weda, H, Barbieri, M. "Automatic children detection in digital images". In proc. of the 2007 IEEE International Conference on Multimedia and Expo, Beijing, China, 2–5 July 2007; pp. 1687–1690.

[6] Xu, M, Wang, T, Wu, Z, Zhou, J, Li, J, Wu, H. "Demand Driven Store Site Selection Via Multiple Spatial-Temporal Data". In proc. of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Burlingame, CA, USA, 31 October–3 November 2016; pp. 1–10. https://doi.org/10.1145/2996913.2996996.

[7] Mouha, R.A. “Internet of Things (IoT)”. J. Data Anal. Inf. Process. 2021, 9, 77–101. https://doi.org/10.4236/jdaip.2021.92006.

[8] Shambour, M.K, Gutub, A. “Progress of IoT Research Technologies and Applications Serving Hajj and Umrah”. Arab. J. Sci. Eng. 2022, 47, 1253–1273.

[9] Krishnamurthi, R, Kumar, A, Gopinathan, D, Nayyar, A, Qureshi, B. “An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques”. Sensors 2020, 20, 6076. https://doi.org/10.3390/s20216076

[10] Alam, T. “The Relation of Artificial Intelligence With Internet Of Things: A survey”. ResearchGate 2020, 1, 30–34.

[11] McCarthy, J. “What Is Artificial Intelligence?” Stanford University: Stanford, CA, USA, 2007. Available online: http://jmc.stanford.edu/articles/whatisai.html (accessed on 12 June 2024).

[12] Varghese, E.B, Thampi, S.M. “Application of Cognitive Computing for Smart Crowd Management”. IT Prof. 2020, 22, 43–50.

[13] IBM. “What Is Computer Vision?” Available online: https://www.ibm.com/topics/computer-vision (accessed on 3 May 2024).

[14] Chan, A.B, Vasconcelos, N. “Crowd Analysis Using Computer Vision Techniques”. IEEE Signal Process. Mag. 2010, 27, 66–77. https://doi.org/10.1109/MSP.2010.937394.

[15] Solmaz, G, Wu, F, Cirillo, F, et al. “Toward Understanding Crowd Mobility in Smart Cities through the Internet of Things”. IEEE Signal Process. Mag. 2019, 57, 40–46. https://doi.org/10.1109/MCOM.2019.1800611.

[16] Altwayan, L, AlMuhayfith, M, Alshahrani, et al. “An Intelligent IoT Approach for Analyzing and Managing Crowds”. IEEE Access 2021, 9, 104874–104886. https://doi.org/10.1109/ACCESS.2021.3099531

[17] Li, L. “A Crowd Density Detection Algorithm for Tourist Attractions Based on Monitoring Video Dynamic Information Analysis”. Complex. J. 2020, 2020, 14. https://doi.org/10.1155/2020/6635446.

[18] Wong, V.W.H, Law, K.H. “Fusion of CCTV Video and Spatial Information for Automated Crowd Congestion Monitoring in Public Urban Spaces”. Algorithms 2023, 16, 154. https://doi.org/10.3390/a16030154.

[19] Zhou, Q, Gu, J, Lu, X, Zhuang, F, Zhao, Y, Wang, Q, Zhang, X. “Modeling Heterogeneous Relations across Multiple Modes for Potential Crowd Flow Prediction”. AAAI Conf. Artif. Intell. 2021, 35, 4723–4731. https://doi.org/10.1609/aaai.v35i5.16603.

[20] Chua, S.N.D, Lim, S.F, Lai, S.N, Chang, T.K. “Development of a Child Detection System with Artificial Intelligence Using Object Detection Method”. J. Electr. Eng. Technol. 2019, 13, 2523–2529.

[21] Ince, O.F, Ince İ, F, Park, J.S. “Using biometric features on long-distance videos for accurate pedestrian age classification”. Bilişim Teknol. Derg. 2017, 10, 123–128. https://doi.org/10.17671/gazibtd.309264.

[22] Ince, O.F, Ince, I.F, Park, J.S, Song, J.K, Yoon, B.W. “Child and adult classification using biometric features based on video analytics”. ICIC Express Lett. Part B Appl. 2017, 8, 819–825.

[23] Agarwal, M, Jain, S. “Image Classification for Underage Detection in Restricted Public Zones”. In proc. of the 8th International Advance Computing Conference (IACC), Greater Noida, India, 14–15 December 2018; pp. 355–359.

[24] Bewley, S, Ge, Z, Ott, L, Ramos, F, Upcroft, B. “Simple Online and Realtime Tracking with a Deep Association Metric”. In proc. of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 3464–3468. https://doi.org/10.1109/ICIP.2016.7533003.

[25] Wu, J, Wang, X, Huang, L, Wang, Z, Wan, D, Li, P. “Parameterized Site Selection Approach of Park Entrance Based on Crowd Simulation and Design Requirement”. Appl. Sci. J. 2023, 13, 6280, . https://doi.org/10.3390/app13106280.

[26] Alossta, O.E, Badi, I. “Resolving a Location Selection Problem By Means of an Integrated AHP-RAFSI Approach”. ResearchGate 2021, 2, 135–142. https://doi.org/10.31181/rme200102135a.

[27] Alhothali, B.A, Faisal, K, Alshammari, et al. “Location-Allocation Model to Improve the Distribution of COVID-19 Vaccine Centers in Jeddah City, Saudi Arabia”. Int. J. Environ. Res. Public Health 2022, 19, 8755. https://doi.org/10.3390/ijerph19148755.

[28] Akbari, V, Rajabi, M.A, Shams, R, Chavoshi, S.H. “Landfill Site Selection by Combining GIS and Fuzzy Multi Criteria Decision Analysis, Case Study: Bandar Abbas, Iran”. World Appl. Sci. J. 2008, 3, 39-47.

[29] Snieska, V, Zykiene, I, Burksaitiene, D. “Evaluation of Location’s Attractiveness for Business Growth in Smart Development”. Econ. Res.-Ekon. Istraživanja 2019, 32, 925–946. https://doi.org/10.1080/1331677X.2019.1590217.

[30] Akpan, S.J, Uford, I. “Retail Outlet Location and Business Performance in Uyo Metropolis”. Int. J. Adv. Manag. Econ. 2023, 12, 9–18.

[31] Yakovlev, S, Kiseleva, O, Chumachenko, D, Podzeha, D. “Maximum Service Coverage in Business Site Selection Using Computer Geometry Software”. Electronics 2023, 12, 2329. https://doi.org/10.3390/electronics12102329.

[32] Cheruvu, S, Kumar, A, Smith, N, Wheeler, D.M. “Demystifying Internet of Things Security: Successful IoT Device/Edge and Platform Security Deployment”, 1st ed, Apress: Berkeley, CA, USA, 2020.

[33] Farchione, D. Children vs Adults Classification; Kaggle: San Francisco, CA, USA, 2023.

[34] YouTube Channel. The Real Samui Webcam. Available online: https://www.youtube.com/@TheRealSamuiWebcam (accessed on 13 March 2024).

[35] Redmon, J, Divvala, S, Girshick, R, Farhadi, A. “You Only Look Once: Unified, Real-Time Object Detection”. arXiv, 2015, arXiv.1506.02640. https://doi.org/10.48550/arXiv.1506.02640.

[36] Lee, L.J, Desa, H, Azizan, M.A, Hussain, A.T, Tanveer, M.H. “Object Detection and Instance Segmentation with YOLOV8: Progress and Limitations”. Proc. Int. Conf. Artif. Life Robot. 2024, 29, 724–728. https://doi.org/10.5954/icarob.2024.os23-5.

[37] Saidani, T. “Deep Learning Approach: YOLOv5-based Custom Object Detection”. Eng. Technol. Appl. Sci. Res. 2023, 13, 12158–12163.

[38] Ultralytics. Segment Documentation. Available online: https://docs.ultralytics.com/tasks/segment/ (accessed on 3 March 2024).

[39] Lin, J.-M, Lin, W.-L, Fan, C.-P. “Age Group Classifier of Adults and Children with YOLO-based Deep Learning Pre-Processing Scheme for Embedded Platforms”. In proc. of the 12th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, 2–6 September 2022; pp. 1–5. https://doi.org/10.1109/ICCE-BERLIN56473.2022.9937129.

[40] Sharma, N, Baral, S, Paing, M.P, Chawuthai, R. “Parking Time Violation Tracking Using YOLOv8 and Tracking Algorithms”. Sensors 2023, 23, 5843. https://doi.org/10.3390/s23135843.