Using Real-Time Detection Transformer in Park Management

Tai-Tien Chen[1*], Yao-Chung Chen[2], Chen-Yu Hao[2], Tien-Yin Chou[2], Ting-Hsuan Lai[2]

[1]Ph.D. Program for Infrastructure Planning and Engineering Feng Chia University, Taichung, Taiwan
[2]GIS research center, Feng Chia University, Taichung, Taiwan

Abstract: Object detection is one of the primary tasks in computer vision and a crucial tool for digital governance in smart cities. The YOLO (You Only Look Once) series of object detection techniques have been regarded as efficient methods. With the rise of Transformers, research in various tasks has stirred waves in the field of computer vision. RTDETR (Real Time Detection Transformer) emerges as a derivative method in the object detection domain. It is well known that each detection method has its own characteristics, and selecting an appropriate model according to task requirements is a common practice for executing the task. Therefore, this study retrained YOLOv8-x, YOLOv8-l, RTDETR-x, and RTDETR-l models separately for object categories required by our park management. The results show that in terms of detection accuracy, the YOLOv8-x and RTDETR-x models outperform the YOLOv8-l and RTDETR-l models, and the RTDETR-x demonstrates the best object detection capability.

Keywords: Object detection, You Only Look Once, Real Time Detection Transformer

I. Introduction

Smart City serves as the cornerstone of national smart governance. Achieving the goals of a smart city requires progress in various aspects such as technology, infrastructure, policies, and community participation, with the establishment of both hardware and software infrastructure being crucial. However, it is essential to emphasize the collection and effective utilization of information, particularly with the integration of artificial intelligence technology to automate and enhance urban governance efficiency. Therefore, leveraging the information collected from hardware infrastructure can propel the development of smart cities, thereby enhancing the quality of life for citizens.

Digital cameras are among the most common sensing devices in current smart cities, primarily deployed in public spaces such as parks, transportation routes, important facilities, and buildings for public safety monitoring and real-time situation awareness [1]. Parks, serving as vital venues for urban residents' leisure activities, family bonding, cultural events, and environmental conservation, can benefit from the proper installation of digital cameras and the application of artificial intelligence technology to improve park management efficiency, enhance public safety, and provide assurance for nature conservation and activity monitoring.

Smart image object detection technology [2], utilizing artificial intelligence, enables various tasks such as object detection, tracking, positioning, and measurement, with widespread applications across industries. Considering the advantages of smart image recognition in the regulatory domain and the demands for public safety and environmental monitoring in public recreational parks [3], this study focuses on the application of smart image detection technology in park governance. It explores the performance of two outstanding detection models, YOLO and RTDETR, in the application scenario of this study.

The motivation behind this research lies in addressing the demands and challenges faced in the governance of public spaces. As crucial community venues in urban areas, parks require effective management and supervision to ensure public safety, efficient resource utilization, and a good environmental quality. The application of smart image detection technology can enhance governance efficiency, strengthen public safety, and provide better public space services and management, thereby establishing more effective governance methods and system frameworks to promote better management and development of public recreational parks.

II. Related Works

Single-stage object detection is a computer vision technique aimed at directly utilizing deep neural networks to identify object positions and categories in images without the need for intermediate steps such as region extraction. The evolution of this technique has primarily focused on continual improvements in neural network architectures and training methods to enhance the accuracy and efficiency of object detection. The most representative model of single-stage object detection is the YOLO (You Only Look Once) series. The earliest model (Version 1) was proposed by Redmon et al. [4]in 2015, as an end-to-end object detection algorithm that simultaneously performs object category classification and bounding box prediction in a single neural network. A key innovation of YOLO is transforming the object detection problem into a regression problem, predicting bounding box coordinates and class probabilities, and dividing the entire image into numerous grid cells for prediction, enabling real-time object detection. One of the crucial features of YOLO is its speed and efficiency, allowing it to perform detection tasks faster compared to traditional two-stage object detection methods.

YOLOv1 consists of 24 convolutional layers followed by 2 fully connected layers, used for predicting bounding box coordinates and probabilities. Except for the last layer that employs a linear activation function, all layers use leaky rectified linear unit activations. YOLOv2, introduced by Redmon et al. [5] in 2017, adopts the Darknet-19 architecture and improves the original YOLO by incorporating batch normalization, anchor boxes, and dimension clustering, resulting in a better-performing model that maintains the same speed while being more robust and capable of detecting 9000 categories. YOLOv3, proposed by Redmon et al. [6] in 2018, utilizes the Darknet-53 backbone architecture and further enhances the model's performance with more efficient multiple anchor points and spatial pyramid pooling. YOLOv4, introduced by Bochkovskiy et al. [7] in 2020, employs the CSPDarknet53 backbone architecture and introduces techniques such as Mosaic data augmentation, new anchor-free detection heads, and a new loss function. YOLOv5, proposed by Glen Jocher, founder, and CEO of Ultralytics [8], in 2020, modifies the CSPDarknet53 backbone network, improves aspects of YOLOv4, and is developed in PyTorch instead of Darknet, while also adopting the Ultralytics algorithm called AutoAnchor. YOLOv6, presented by Meituan [9] in 2022, incorporates backbone blocks such as RepVGG or CSPStackRep and utilizes post-training quantization and channel-level distillation for enhanced quantization techniques, resulting in faster and more accurate detectors. YOLOv7, proposed by Wang et al. [10] in 2022, employs the Extended Efficient Layer Aggregation Network (E-ELAN) and RepVGG, enhancing the network's learning ability through shuffling and merging cardinalities without disrupting the original gradient paths. YOLOv8, released by Ultralytics [11] in 2023, utilizes the modified CSPDarknet53 backbone network and provides five scaled versions: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. YOLOv8 supports various visual tasks, including object detection, segmentation, pose estimation, tracking, and classification.

Vision Transformer (VIT) is a deep learning model based on the Transformer architecture, first introduced by Dosovitskiy et al. [12] in 2020 in the paper "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." This research explores the application of the Transformer architecture to image classification tasks, proposing a novel approach of slicing images into fixed-size patches and feeding them into a Transformer model for processing, without using traditional convolutional neural networks (CNNs). The study demonstrates that Transformer architecture can achieve comparable performance to traditional CNNs in image classification. This paper lays the foundation for the development and subsequent applications of the ViT model in the field of image processing.

DETR (Detection Transformer), proposed by Carion et al. [13] in 2020, is a model that adopts the Transformer architecture for end-to-end object detection. The DETR model utilizes attention mechanisms, eliminating the need for traditional components like anchor boxes and non-maximum suppression (NMS) in conventional object detection algorithms, directly encoding and predicting object positions using Transformers. Deformable DETR, introduced by Zhu et al. [14] in 2020, integrates deformable mechanisms into the Transformer model, allowing the model to flexibly adjust receptive fields to adapt to objects of different scales and shapes, enabling better feature acquisition and understanding based on the actual shapes and positions of objects in the image.Meng et al. [15] proposed Conditional DETR in 2021, which learns conditional spatial queries from decoder embeddings for decoder multi-head cross-attention. This approach allows each cross-attention head to focus on bands containing different regions, such as extreme points of a target or areas inside the target box. The advantage of this approach is reducing the spatial range for localizing different regions for target classification and box regression, thereby alleviating the dependence on content embeddings and simplifying the training process.

RT-DETR, introduced by Lv et al. [16] in 2023, designs an efficient hybrid encoder to process multi-scale features efficiently by decoupling intra-scale interaction and cross-scale fusion. Additionally, it proposes IoU-aware query selection to improve the initialization of object queries. Furthermore, the proposed detector supports flexible adjustment of inference speed by using different decoder layers without the need for retraining, facilitating the practical application of real-time object detectors.

III. Methods and results

The main purpose of this study is to evaluate the performance of YOLO and RTDETR models in object detection tasks for park management and explore their effectiveness in real-world applications. We selected seven categories including people, cars, motorcycles, bicycles, cats, dogs, and golf carts to perform real-time statistics on their occurrences in the park. This information is crucial for park managers to understand park usage patterns and take appropriate measures to improve management and service quality. Furthermore, we extended the models to detect fallen individuals to further enhance park safety. Once a person falls, the model will promptly detect it and send a notification to the park management staff, enabling them to take swift action, provide assistance, and prevent potential accidents. This real-time alert system can significantly improve park safety and allow people to engage in park activities with greater peace of mind.

To achieve this goal, we designed a comprehensive research framework. Firstly, we collected and preprocessed a large amount of training data, including images and labels of various objects. Then, we used this data to train the YOLO and RTDETR models to accurately identify and classify different objects. Subsequently, we tested the trained models to evaluate their performance in park scenarios. Finally, we deployed these models into actual park management systems and validated their effectiveness in real-world applications.The entire research process was developed using Python, providing us with a rich set of tools and libraries to achieve our objectives. We utilized the OpenCV library for image data processing, while the YOLO and RTDETR models were implemented using deep learning frameworks. In the training and testing phases of the models, we employed substantial computational resources to accelerate the experimental process and ensure that the models achieve good generalization performance in various scenarios. The overall framework is illustrated in Figure 1.

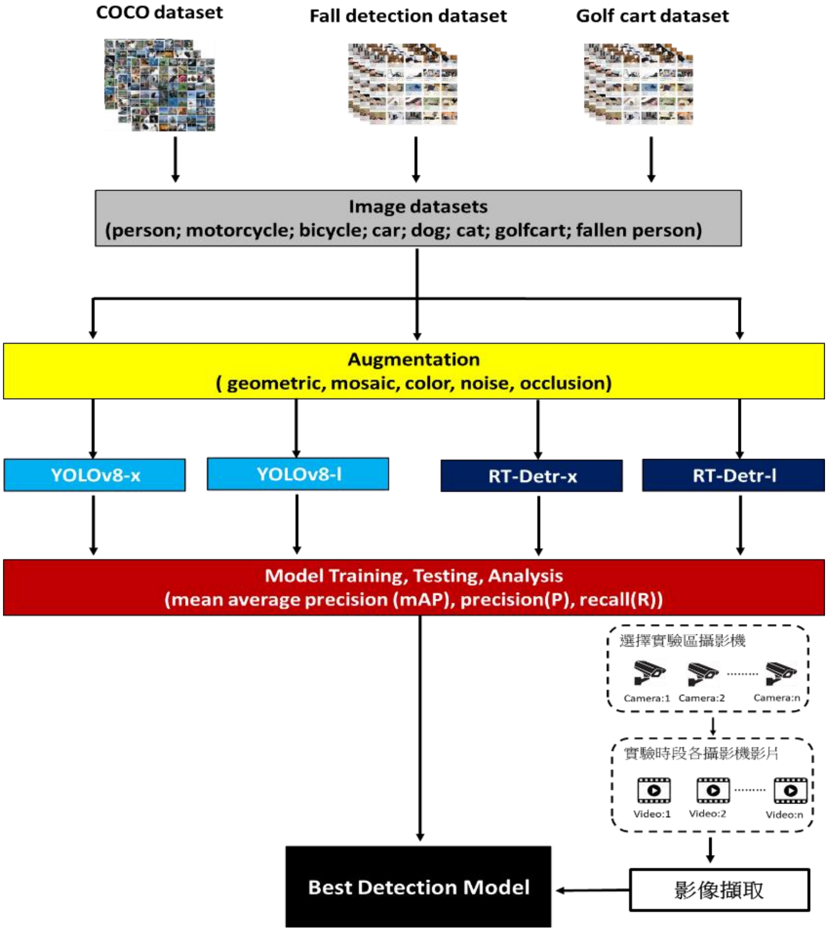


Figure 1. The proposed research framework

3.1 Data Collection and Preprocessing

In this study, a total of 5,322 original photos were collected, all of which were sourced from open-access datasets and were already annotated with information such as object categories and positions within the images. However, to enhance the diversity of the training data and improve the model's ability to generalize, we employed data augmentation techniques on these original photos.Data augmentation is a commonly used technique that involves applying various transformations and manipulations to the original data to generate new data, thereby expanding the training dataset. In this study, we utilized several data augmentation methods (Figure 2), including:

- **Geometric transformations:** The original photos were subjected to rotations, affine transformations, and scaling to introduce samples from different angles and scales.
- **Color adjustments:** Parameters such as hue, brightness, and contrast were adjusted in the original photos to simulate different lighting conditions.
- **Occlusion:** Objects such as text, sunshades, shadows, and sunflares were added to the original photos to simulate occlusion scenarios commonly encountered in real-world settings.
- **Mosaic:** Multiple original photos were stitched together, followed by cropping and rearranging of the stitched image to generate new images, thereby increasing the diversity of the data.

Through the application of these data augmentation techniques, we generated a total of 16,003 augmented photos, enriching the variety of samples in the training dataset. Therefore, our training dataset comprises both original and augmented photos, totaling 21,325 images.Such data augmentation processes not only contribute to enhancing the model's generalization ability but also enable the model to better adapt to various scenes and conditions, thereby improving its performance in real-world applications.

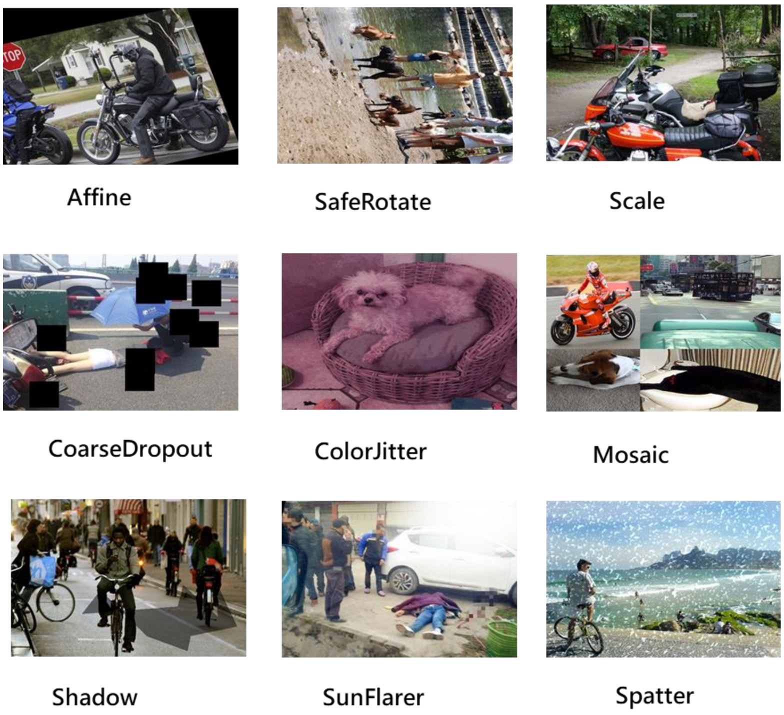


Figure 2. Augmentation Methods

3.2 Model Training

In this study, we selected two advanced one stage methods, YOLOv8 and RT-Detr, for object detection tasks. These two methods are highly regarded models in the field of computer vision, known for their outstanding performance and effectiveness in object detection tasks.

YOLOv8 is an improved version of YOLOv5, which utilizes deep convolutional neural networks (CNN) to directly predict bounding boxes and class probabilities of objects in images. Compared to YOLOv5, YOLOv8 further optimizes network structure, data augmentation, and training techniques to achieve better detection performance. This method offers faster speed and higher accuracy, making it suitable for real-time detection scenarios.On the other hand, RT-Detr is a hybrid detection model that combines Transformer and CNN. Inspired by the success of Transformer in natural language processing, RT-Detr transforms object detection problems into attention mechanism problems. By leveraging self-attention mechanisms, RT-Detr effectively models relationships among objects in images, thereby improving detection accuracy and robustness.

During the training process, we utilized a total of 21,325 images as training dataset, including 5,322 original images and 16,003 images processed through data augmentation. These images were randomly split into 80% for training and 20% for validation to ensure the models learn and generalize effectively.The main training parameters are as follows:

- **Epochs:** Total number of training epochs set to 100.
- **Batch size:** Number of images per training batch set to 16.
- **Image size:** Size of images in pixels set to 640x640.
- **Optimizer:** AdamW optimizer, which combines fast convergence speed with regularization effects of weight decay, aiding in improving the models' generalization performance.

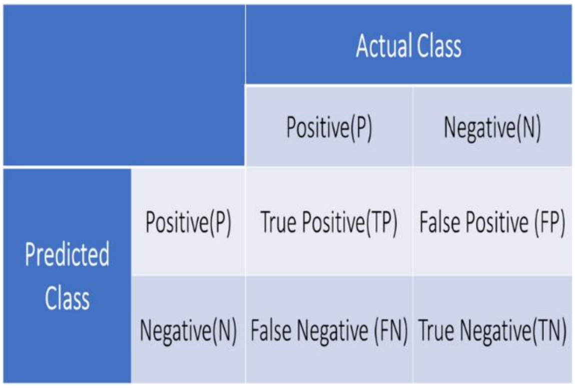
3.3 Evaluation Metrics

This study adopts four commonly used evaluation metrics for image classification and object detection as the basis for model evaluation:

- **Precision:** Precision measures the probability of correctly predicting positive cases among all predicted positive cases (TP/(TP+FP)). It indicates how many detections are correct.
- **Recall:** Recall measures the probability of correctly identifying positive cases among all actual positive cases (TP/(TP+FN)). It assesses the model's ability to recognize all instances of objects in images.
- **mAP50:** Mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5, which considers the precision of "easy" detections only.
- **mAP50-95:** Mean Average Precision (mAP) over different IoU thresholds ranging from 0.5 to 0.95 with an interval of 0.05. It comprehensively reflects the model's performance under different detection difficulty levels.

Confusion matrix as shown in Figure 3:

- **TP (True Positive):** True positives indicate instances where the model correctly classifies positive samples as positive. It means the predicted result matches the ground truth.
- **TN (True Negative):** True negatives indicate instances where the model correctly classifies negative samples as negative. It means the predicted result matches the ground truth.
- **FP (False Positive):** False positives indicate instances where the model incorrectly classifies negative samples as positive. It means the predicted result contradicts the ground truth.
- **FN (False Negative):** False negatives indicate instances where the model incorrectly classifies positive samples as negative. It means the predicted result contradicts the ground truth.



		Actual Class	
		Positive(P)	Negative(N)
Predicted Class	Positive(P)	True Positive(TP)	False Positive (FP)
	Negative(N)	False Negative (FN)	True Negative(TN)

Figure 3. confusion matrix

3.4 Results Analysis

YOLOv8 and RT-DETR are two different structured detection models trained in this study. However, there are currently five variations of the YOLOv8 model publicly available, with model parameters ranging from small to large, namely YOLOv8-n, YOLOv8-s, YOLOv8-m, YOLOv8-l, and YOLOv8-x. For RT-DETR, there are two publicly available models, namely RT-DETR-l and RT-DETR-x. These two models, RT-DETR-l and RT-DETR-x, correspond to the large models in their respective detection models, YOLOv8-l and YOLOv8-x. Therefore, we compared these four models as training models under as similar conditions as possible to ensure the reliability of the research results.

Table 1 presents the training results of four models. According to the findings, RTDETR-l performs best in Precision, reaching 0.8423. This indicates that in all cases predicted as positive samples by the model, 84.23% were correctly predicted. This demonstrates that RTDETR-l has higher accuracy in identifying positive samples. RTDETR-x achieves the best Recall at 0.7252, indicating that the model can identify actual positive samples with a probability of 72.52%. This suggests that RTDETR-x has higher capability in recognizing all object instances in images. RTDETR-x also achieves the best mAP50 at 0.7896, indicating that its average precision is 78.96% when IoU (Intersection over Union) is greater than 0.5. This indicates that RTDETR-x has overall better performance in detecting objects, especially in easier detection scenarios. Yolo-v8x performs best in mAP50-95 at 0.5315, indicating an average precision of 53.15% across all IoU values between 0.5 and 0.95. This suggests that Yolo-v8x demonstrates more balanced performance across various detection difficulties.

Overall, the results show that larger models, such as RTDETR-x and Yolo-v8x, generally outperform smaller models, such as RTDETR-l and Yolo-v8l, in terms of precision and performance. Additionally, models that combine CNN and Transformer, such as RTDETR, exhibit higher detection accuracy. RTDETR-x outperforms other models by approximately 3-6% in Recall and mAP50, while being slightly lower than RTDETR-l by 0.34% and YOLOv8-x by 0.15% in Precision and mAP50-90, respectively. Therefore, considering all factors, RTDETR-x demonstrates superior model training accuracy. At the same time, it is evident that the model training results of Yolov8-l are relatively inferior.

Table1. Training results

	Precision	**Recall**	**mAP50**	**mAP50-95**
YOLOv8-l	0.7942	0.6651	0.7432	0.5072
YOLOv8-x	0.8007	0.6955	0.7649	0.5315
RTDETR-l	0.8423	0.6962	0.7405	0.5171
RTDETR-x	0.8389	0.7252	0.7896	0.5300

To assess the practical application performance of the trained RTDETR-x and YOLOv8-x models, this study conducted tests on object detection in park CCTV images, as shown in Figure 4. The results indicate that in the same scenario, RTDETR-x outperforms YOLOv8-x in terms of detection efficiency. Firstly, RTDETR-x is better at detecting occluded objects, such as vehicles in parking lots and pedestrians obscured by roadside trees in the images. Additionally, for identical objects present in the scene, RTDETR-x provides higher score estimates for object detection. This aspect is equally crucial, as in certain cases, to avoid erroneous detection results, a threshold is usually set to filter detected objects. A higher score indicates greater confidence and reliability in detecting objects, reducing the likelihood of false positives being filtered out.In the second row of Figure 4, RTDETR-x correctly detects the two persons on the motorcycle, while YOLOv8-x, at the same time, categorizes the two persons as one. However, RTDETR-x is not without flaws, as it incorrectly identifies a group of motorcycles as cars in the image. Although YOLOv8-x does not make this particular misjudgment, it can be observed that it also generates false positives for bicycles. Undoubtedly, RTDETR-x exhibits remarkable detection performance, whether in the model training phase for data fitting capability or in predicting new application scenarios.

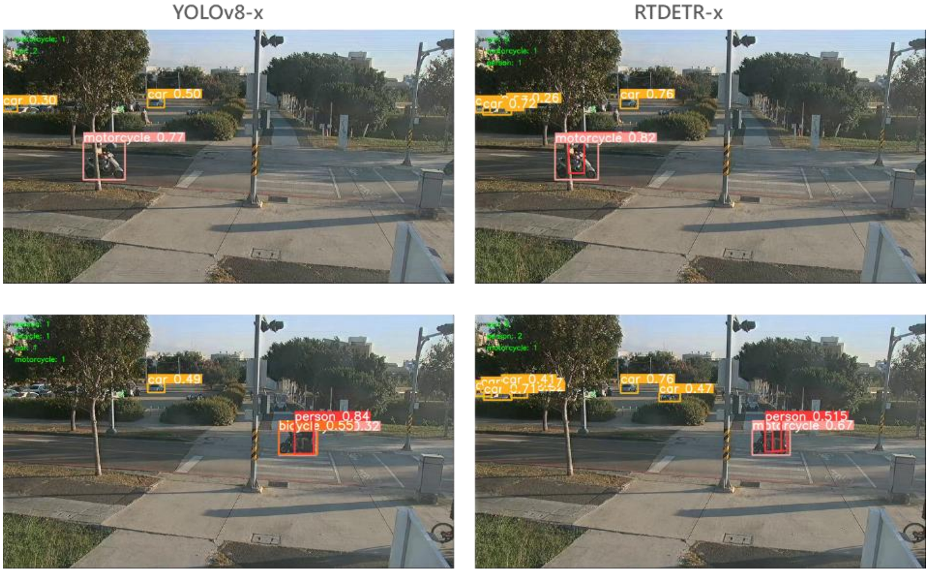


Figure 4. Detection Results Comparison

IV. Discussion and Conclusion

A fine object detection model should achieve comparative validation results during the training process and demonstrate excellent generalization ability on unseen images, enabling its application to various downstream detection tasks. In this study, we utilized state-of-the-art object detection models, YOLO and RTDETR, and trained them on the object classes required for park management. Various image augmentation techniques were employed to enhance data diversity during the training process.The results indicate that RTDETR exhibits superior detection accuracy in terms of Precision and Recall compared to YOLO for overall object detection. This may be attributed to RTDETR's utilization of Transformer encoder, decode, which provide better understanding of global visual features in images. However, when examining the mAP50 and mAP50-95 metrics, the best results are obtained from RTDETR-x and YOLOv8-x, respectively. This implies that these two models excel in high-precision or reliable object detection capabilities. In other words, when higher precision or reliability of detected objects is required, employing these two models can meet the application needs, with RTDETR-x overall achieving the best trained model.When applying RTDETR-x and YOLOv8-x models in real park scenarios, it was observed that RTDETR-x outperforms YOLOv8-x in detecting partially occluded or smaller objects. Therefore, RTDETR-x demonstrates remarkable advantages in detection accuracy, showcasing a promising trend in combining CNN and Transformer methods in the field of object detection, becoming a new trend and research focus.

References

[1]. Laufs, J.; Borrion, H.; Bradford, B. Security and the Smart City: A Systematic Review. Sustain. Cities Soc. 2020, 55, 102023.

[2]. Ahmad, H.M.; Rahimi, A. Deep Learning Methods for Object Detection in Smart Manufacturing: A Survey. J. Manuf. Syst. 2022, 64, 181–196.

[3]. Jemmali, M.; Melhim, L.K.B.; Alharbi, M.T.; Bajahzar, A.; Omri, M.N. Smart-Parking Management Algorithms in Smart City. Sci. Rep. 2022, 12, 6533.

[4]. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; pp. 779–788.

[5]. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; pp. 7263–7271.

[6]. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. ArXiv Prepr. ArXiv180402767 2018.

[7]. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal Speed and Accuracy of Object Detection. ArXiv Prepr. ArXiv200410934 2020.

[8]. Jocher, G. YOLOv5 by Ultralytics. 2020. Available Online: Https://Github.Com/Ultralytics/Yolov5 (Accessed on March 2024).

[9]. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. ArXiv Prepr. ArXiv220902976 2022.

[10]. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023; pp. 7464–7475.

[11]. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics. 2023. Available Online: Https://Github.Com/Ultralytics/Ultralytics (Accessed on March 2024).

[12]. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv Prepr. ArXiv201011929 2020.

[13]. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European conference on computer vision; Springer, 2020; pp. 213–229.

[14]. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable Detr: Deformable Transformers for End-to-End Object Detection. ArXiv Prepr. ArXiv201004159 2020.

[15]. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional Detr for Fast Training Convergence. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 3651–3660.

[16]. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. Detrs Beat Yolos on Real-Time Object Detection. ArXiv Prepr. ArXiv230408069 2023.