electronics

MDPI

Article

CMCA-YOLO: A Study on a Real-Time Object Detection Model for Parking Lot Surveillance Imagery

Ning Zhao [1], Ke Wang [1], Jiaxing Yang [1], Fengkai Luan [1], Liping Yuan [2] and Hu Zhang [1,*]

[1] School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China; zhaoning@whut.edu.cn (N.Z.); 276028@whut.edu.cn (K.W.); 276140@whut.edu.cn (J.Y.); 275902@whut.edu.cn (F.L.)

[2] School of Information Engineering, Wuhan Huaxia Institute of Technology, Wuhan 430223, China; whutylp@whut.edu.cn

* Correspondence: huzhang@whut.edu.cn

Abstract: In the accelerated phase of urbanization, intelligent surveillance systems play an increasingly pivotal role in enhancing urban management efficiency, particularly in the realm of parking lot administration. The precise identification of small and overlapping targets within parking areas is of paramount importance for augmenting parking efficiency and ensuring the safety of vehicles and pedestrians. To address this challenge, this paper delves into and amalgamates cross-attention and multi-spectral channel attention mechanisms, innovatively designing the Criss-cross and Multi-spectral Channel Attention (CMCA) module and subsequently refining the CMCA-YOLO model, specifically optimized for parking lot surveillance scenarios. Through meticulous analysis of pixel-level contextual information and frequency characteristics, the CMCA-YOLO model achieves significant advancements in accuracy and speed for detecting small and overlapping targets, exhibiting exceptional performance in complex environments. Furthermore, the study validates the research on a proprietary dataset of parking lot scenes comprising 4502 images, where the CMCA-YOLO model achieves an mAP@0.5 score of 0.895, with a pedestrian detection accuracy that surpasses the baseline model by 5%. Comparative experiments and ablation studies with existing technologies thoroughly demonstrate the CMCA-YOLO model's superiority and advantages in handling complex surveillance scenarios.

Keywords: intelligent surveillance; parking lot management; deep learning; attention mechanisms; CMCA-YOLO

check for updates

Citation: Zhao, N.; Wang, K.; Yang, J.; Luan, F.; Yuan, L.; Zhang, H. CMCA-YOLO: A Study on a Real-Time Object Detection Model for Parking Lot Surveillance Imagery. *Electronics* **2024**, *13*, 1557. https://doi.org/10.3390/electronics13081557

Academic Editor: Daniel Morris

Received: 5 March 2024
Revised: 15 April 2024
Accepted: 17 April 2024
Published: 19 April 2024

1. Introduction

1.1. Research Background

In the rapidly advancing digital age, revolutionary strides in computer and artificial intelligence technologies have profoundly transformed contemporary life and work modalities. Particularly in the domain of video surveillance, the maturation of intelligent video analysis technology has evolved surveillance systems from basic video recording and playback functionalities to highly complex systems capable of intelligent identification, analysis, and early warning. Among the myriad applications in urban traffic management, parking lot management emerges as particularly critical, playing an indispensable role in enhancing parking efficiency, alleviating traffic congestion, and bolstering urban safety [1]. However, traditional parking lot management has primarily relied on manual monitoring, a method not only inefficient but also prone to errors and omissions, struggling to meet modern society's demands for high efficiency and precision [2–4].

1.2. Traditional Methods and Problems

The cornerstone of intelligent monitoring systems lies in the surveillance video analysis techniques employed. Early explorations into video analysis techniques revolved around four methods: color recognition, background subtraction, optical flow analysis, and shape feature recognition, laying the foundational framework for object identification and tracking. Color recognition technology [5], by constructing color models, extracts color information of objects to determine their motion state through changes in color across adjacent video frames. Background subtraction technology [6] creates a static background model in videos and separates moving objects by updating the background, thereby tracking and analyzing their motion. Optical flow technology [7] describes objects' motion trajectories by calculating the motion vectors of each pixel between consecutive frames, achieving dynamic tracking. Shape feature methods [8] focus on extracting objects' shape features and tracking their motion by analyzing shape changes between frames. Despite the advantages of these methods, they commonly face challenges of high costs, extensive computational requirements, and practical limitations, significantly restricting their application in actual surveillance systems and leading to a reliance on manual monitoring for video analysis.

With the swift evolution of deep learning technologies, methodologies for object detection based on this advanced framework have been extensively implemented in intelligent video surveillance systems. Since the pioneering inception of AlexNet [9], followed by relentless innovation through the R-CNN series [10–13], YOLO series [14–21], SSD series [22,23], and RetinaNet [24], the domain of object detection has witnessed a significant leap in performance capabilities. Accompanying these advancements, researchers have incessantly introduced novel modular structures, such as Residual Networks, CBAM, SE, and SiM [25–28], further enhancing detection accuracy and speed. Notably, the successful incorporation of the Transformer architecture into the field of object detection [29–31] has notably augmented the models' capacity for complex scene modeling, especially in capturing long-range contextual information, which is crucial for application areas like traffic scenarios and urban monitoring. Fundamental models and methods in the domain of object segmentation, such as SAM [32] and RSPrompter [33], particularly their applications in remote sensing detection [34], have offered invaluable insights into efforts for small object detection and segmentation, thereby fostering the continual progression of deep learning technologies. It is worth emphasizing that single-stage detection algorithms, like the YOLO series, due to their exceptional speed and efficiency, have become the preferred solution for real-time video analysis [35].

Nonetheless, the inherent complexity of parking lot surveillance scenarios still poses severe challenges to the high precision and real-time requirements of object detection. These include extreme changes in lighting conditions, frequent obstructions between vehicles and pedestrians, and a wide variety of camera angles, all of which significantly increase the difficulty of recognition. Especially in dealing with small targets like pedestrians, as well as overlapping objects such as vehicles parked side by side or on top of each other, existing intelligent surveillance systems and deep learning technologies often fail to achieve the desired effects. This limitation stems from three key challenges: the difficulty of tracking miniature targets, where limited feature information on small targets like pedestrians captured by distant cameras greatly increases the difficulty of accurate identification and localization; the complexity of recognizing overlapping targets, frequent in parking lots where vehicles and pedestrians overlap, especially during peak traffic times, making it challenging to distinguish between different overlapping objects; and the complexity and variability of scenes, from changes in lighting to the impact of weather conditions, and differences in camera angles and resolutions, all testing the precision of object detection.

1.3. Research Significance and Contribution

Faced with these challenges, the capabilities of existing intelligent parking lot surveillance systems to ensure safe operation and improve parking efficiency are limited, highlighting the urgent need for research and development of new intelligent surveillance systems. A system capable of effectively overcoming the aforementioned difficulties and enhancing the accuracy of detecting small and overlapping targets is crucial for advancing the application and development of intelligent surveillance technology. Therefore, this study innovatively introduces the CMCA module, based on cross-attention and multi-spectral channel attention mechanisms, proposing the CMCA-YOLO model, a target detection model optimized for parking lot surveillance scenarios. This model significantly enhances the recognition capabilities for small and overlapping targets in parking lot surveillance imagery. Compared to existing object detection models, CMCA-YOLO maintains high-speed detection performance while substantially improving target detection accuracy, particularly in recognizing small and overlapping targets in complex environments, demonstrating significant advantages. The contributions of this study are as follows:

(1) Conception and implementation of the CMCA-YOLO model. This study introduces the Criss-Cross Multi-Spectral Channel Attention (CMCA) module for the first time, utilizing the cross-attention mechanism to delve into pixel-level contextual information to differentiate the background and employing Discrete Cosine Transform (DCT) to analyze channel frequencies in video frames, effectively capturing and compressing key channel information. This method allows the model to more accurately focus on features of varying sizes and frequencies, significantly enhancing the recognition capabilities for small and overlapping targets. Especially in parking lot environments with frequent changes in lighting and interweaving of people and vehicles, the model demonstrates exceptional performance improvements;

(2) Creation of a dedicated parking lot scene dataset. The research team organized and constructed a parking lot scene dataset containing 4502 images, covering a variety of environmental conditions and target types. This dataset not only provides valuable resources for training and evaluating the CMCA-YOLO model but also fills the gap in existing datasets, offering a practical benchmark for future research;

(3) Comprehensive efficacy validation. Through a series of ablation experiments and comparisons with existing leading models, this study comprehensively validates the effectiveness and superiority of the CMCA-YOLO model. Experiment results show that CMCA-YOLO outperforms comparative models in multiple performance metrics, especially in handling complex surveillance scenarios, highlighting the model's tremendous potential in practical applications.

In this paper, we present the CMCA-YOLO model, a real-time object detection framework designed to address the unique challenges of parking lot surveillance, such as detecting small and overlapping targets through the novel integration of cross-attention and multi-spectral channel attention mechanisms. The structure of this paper is as follows: Section 2 reviews the evolution of intelligent surveillance technologies, from basic color recognition to advanced deep learning object detection methods. Section 3 details the CMCA module, the centerpiece of our model, explaining its innovative use of attention mechanisms to boost accuracy in complex scenes. Section 4 provides experimental validation using a custom parking lot scene dataset, showcasing the model's enhanced detection precision and speed. The conclusion in Section 5 reflects on our contributions and the advantages of CMCA-YOLO over existing models and outlines avenues for future research to broaden its practical application.

2. Related Work

2.1. The Evolution and Optimization of Intelligent Surveillance Systems

Within the realms of intelligent surveillance and automation, particularly focusing on research concerning parking lot monitoring systems, scholars have proposed numerous innovative methodologies aimed at enhancing the precision and real-time capabilities of target detection. These methods not only strive to boost algorithmic performance but also explore the potential of utilizing existing technological frameworks and emerging computational resources to optimize detection systems. With the ongoing development of Internet of Things (IoT) technology and edge computing, the design of intelligent surveillance systems is undergoing a qualitative leap, transitioning from traditional centralized computing paradigms towards distributed and edge computing architectures. In this transformational process, researchers like Ke et al. [36] have successfully enhanced the real-time detection of parking space occupancy by integrating IoT with edge computing and incorporating the SSD algorithm and real-time video stream analysis technologies. Their research demonstrates the efficiency of intelligent algorithms working in tandem with edge computing devices, although there remains room for improvement in detection accuracy under complex scenarios. Chen et al. [37], addressing maritime traffic monitoring scenarios, introduced the poly-YOLO model, which has shown substantial potential in frame-by-frame object detection, laying a technical foundation for real-time monitoring and management of vehicle trajectories in parking systems. Nguyen et al. [38] proposed a parking lot detection network based on an improved YOLOv5 architecture, YOLO5PKLot, for smart parking management systems. Their work focuses on reducing computational complexity through lightweight network design and parameter optimization while maintaining high precision in target detection. Meanwhile, Ogawa et al. [39] utilized the YOLO model to design a parking space occupancy detection system, proving the feasibility and necessity of incorporating intelligent technologies in parking environments. Wang et al. [40] introduced the Gold-YOLO model employing a Gather-and-Distribute Mechanism, not only enhancing the model's accuracy and processing speed but also improving the model's learning capability through unsupervised pre-training. These advancements demonstrate the potential advantages of considering model architecture in the design of intelligent parking monitoring systems, especially on resource-constrained edge computing devices.

2.2. Research on Lightweight Strategies for Object Detection Models

When addressing the complexity and efficiency challenges of object detection models, researchers face multiple hurdles. These challenges involve achieving high-accuracy detection while reducing dependence on computational resources, ensuring the model's capability for real-time video stream processing to meet low latency and rapid response requirements, and enhancing the model's generalization ability across different environmental conditions to ensure stable operation in diverse monitoring scenarios. To tackle these challenges, researchers have proposed a series of solutions. Lightweight network structures such as MobileNet [41] and ShuffleNet [42] have been widely adopted to reduce model complexity. Knowledge distillation and model pruning techniques are also employed to decrease the number of model parameters. Additionally, employing multi-scale training and data augmentation to enhance the model's generalization ability has become a common practice. These strategies aim to develop object detection models that not only meet the demands for real-time performance and high accuracy but are also adaptable to resource-constrained environments. On this foundation, Zhao et al. [43] made lightweight improvements to the YOLOv5 model, successfully reducing model complexity and enhancing detection efficiency. However, this method primarily focuses on optimizing model structure and parameters, and the model's adaptability and robustness in extreme monitoring scenarios may still need reinforcement. Moreover, Zhang et al.'s CDNet [44] network demonstrated practicality and efficiency in pedestrian crosswalk detection on Jetson Nano devices. However, the network's reliance on edge computing devices may limit its deployment flexibility in broader application scenarios. Additionally, while synthetic fog enhancement algorithms can adapt to foggy conditions, their performance assurance in other adverse weather conditions remains unclear. Song et al. [45] improved the inference speed and accuracy of the YOLOv5-MS network for pedestrian detection through video stream multi-threading capture and module optimization. Yet, multi-threading capture and module optimization might increase the system's overall complexity, posing a challenge in balancing resource usage and algorithm performance in computationally limited environments. Liu [46] proposed a lightweight improvement to the YOLOv5 model by introducing MobileNetv2 to achieve a lightweight YOLOv5 backbone and incorporating attention mechanisms and improved loss functions to enhance the object detection model's robustness and generalization ability. However, the model's adaptability to different types of camera-captured image qualities and performance stability under varied lighting conditions still require further verification and optimization. Xuedong Dong et al. [47] introduced an improved lightweight YOLOv5 method for vehicle detection, specifically incorporating C3Ghost and Ghost modules into the YOLOv5 neck network and introducing convolutional block attention modules into the YOLOv5 backbone. This method boosted vehicle detection performance, but when facing large-scale datasets or complex multi-object detection tasks, the model's accuracy and robustness may be limited by the design of the introduced lightweight modules and attention mechanisms. In summary, these studies have made progress in enhancing the efficiency and accuracy of object detection models. However, in a wide range of practical application scenarios, especially under conditions of environmental complexity and variability, further optimization of these models to improve their robustness, adaptability, and generalizability remains a focal point for future research.

2.3. Optimization Strategies for Object Detection Models in Complex Scenarios

Simultaneously, in the research and application of modern monitoring systems, drone technology has garnered attention for its unique advantages, particularly its high flexibility and adaptability in executing complex monitoring tasks. However, drone monitoring technology and parking lot monitoring systems share numerous challenges in practice, involving extreme changes in lighting, complex dynamic backgrounds, and the impact of variable monitoring environments on image quality [48,49]. These challenges include fluctuations in natural lighting conditions and visual disturbances under adverse weather conditions (such as rain, fog, snow, etc.), as well as complex interference caused by dynamic backgrounds and moving targets. Thus, monitoring models must possess extremely high adaptability and flexibility to accurately identify and track targets of various sizes and postures. Especially in the field of parking lot monitoring, the aforementioned issues are more pronounced, with the uniqueness of parking lot environments, such as extreme changes in lighting conditions and potential obstructions between vehicles, demanding higher accuracy and real-time performance from monitoring systems. Hence, researchers like Li et al. [48] and Zhu et al. [49], by optimizing the YOLOv5 model and employing data augmentation techniques, significantly improved the performance of object detection, offering valuable references. Although these methods have made some progress in enhancing the model's generalizability and dealing with complex monitoring environments, specific challenges unique to parking lot monitoring, such as extreme changes in light intensity and high-density target occlusion, still need to be addressed. Therefore, to further improve the recognition accuracy and real-time performance of parking lot monitoring systems, some studies have begun exploring new avenues, such as [50] introducing continuous image sequences and frame-to-frame optical flow processing methods to simulate human visual mechanisms and [42,51] aiming to enhance the detection capability for small moving targets by improving model structures and loss functions. These innovative methods have significantly improved the performance of monitoring models under specific conditions, but their universality and robustness in actual parking lot monitoring applications, especially in dealing with multi-target occlusion and extreme weather conditions in image capture, remain key issues for current research to explore in depth. Therefore, in exploring the development of parking lot monitoring systems, researchers face challenges not limited to reduced image quality but more broadly encompass maintaining algorithmic real-time performance and accuracy under adverse weather and low-light conditions. For instance, Mahaur et al. [52], by improving the YOLOv5 model, successfully enhanced the model's detection accuracy and speed under low light and adverse weather conditions, showcasing the potential of deep learning models to adapt to extreme environmental conditions. Their research not only addressed specific image processing challenges but also improved input image quality through image preprocessing and enhancement techniques, increasing the model's sensitivity to small objects. However, performing multi-object detection in high-density parking environments, especially when there is occlusion between targets, remains a significant challenge in effectively balancing detection speed and accuracy. This is mainly because, in complex scenes, the model needs to process a large amount of information in a very short time, while occlusion and similar targets may lead to misidentification and missed detections. Further, Qu et al. [53] improved the YOLOv5 by implementing cross-layer fusion of multi-scale features, effectively addressing the detection of large objects, especially in cases of significant target size variation. This improvement increased the model's adaptability to targets of different sizes, but optimizing model performance and resource allocation strategies in high-density scenes featuring both small and large objects still requires further research. Omar et al. [54], with their license plate detection and recognition cascade algorithm, and Lou et al. [55], with the DC-YOLOv8 model, focused on detecting small-sized objects, showed superior performance in specific application scenarios, but these methods face challenges in universality and adaptability in broader parking lot monitoring applications. Especially with significant differences in image quality captured by different types of cameras and under variable environmental conditions, ensuring the stability and accuracy of detection algorithms becomes a major issue. Thus, while these studies have made progress in specific areas, exploring further optimization of models to improve their robustness and generalizability in complex real-world application scenarios remains a focal point for future research.

2.4. Comparison with Existing Models

In summary, addressing the aforementioned issues, this study proposes the CMCA-YOLO model, which, by leveraging cross-attention mechanisms and multi-spectral channel attention mechanisms, not only aggregates contextual information and extracts key frequency components but also significantly enhances the identification capability for overlapping targets in monitoring scenarios, effectively augmenting the model's analytical ability in complex scenes. The research outcomes of the CMCA-YOLO model have made significant contributions in terms of detection accuracy and real-time performance and have provided more effective solutions tailored to the unique needs of parking lot monitoring scenarios, highlighting the complementarity and application value of cross-disciplinary research outcomes when facing similar challenges. These explorations not only bring new technological breakthroughs to the field of parking lot monitoring but also offer important directions for future research in enhancing model robustness and generalizability in complex real-world application scenarios.

3. Methodology

3.1. CMCA-YOLO Model

The architectural design of the CMCA-YOLO model, as illustrated in Figure 1, essentially follows and optimizes the framework of YOLOv5 to meet the specific requirements of parking lot surveillance systems. The key component of this architecture is the CMCA module, which employs both Criss-Cross Attention (CCA) and Multi-Spectral Channel Attention (MSCA) mechanisms. Through meticulous aggregation of contextual information, it achieves efficient recognition of overlapping targets. Moreover, it enhances the expression of inter-channel information through frequency domain analysis, enabling more precise feature representation. The cross-attention mechanism generates attention maps by calculating the relationship between each pixel in the feature map and all other pixels, allowing each pixel to extract information from the entire image. This iterative operation improves the model's capability to recognize overlapping targets through the aggregation of contextual information both horizontally and vertically. The multi-spectral channel attention mechanism, on the other hand, analyzes channel information in the frequency domain through DCT, enabling the model to focus on a broader range of frequency components and thus enhancing channel representation capabilities. The detailed design of the CMCA module is further elaborated in Section 3.2.

Within the CMCA-YOLO architecture, feature extraction from images is the initial step. The backbone network processes images through a series of Convolution-Batch Normalization-Activation (CBA) modules for feature extraction and nonlinear transformation, generating primary feature maps. These feature maps are then fed into the

CMCA module for deeper processing through CCA and MSCA mechanisms, enhancing the model's analytical power in complex environments.

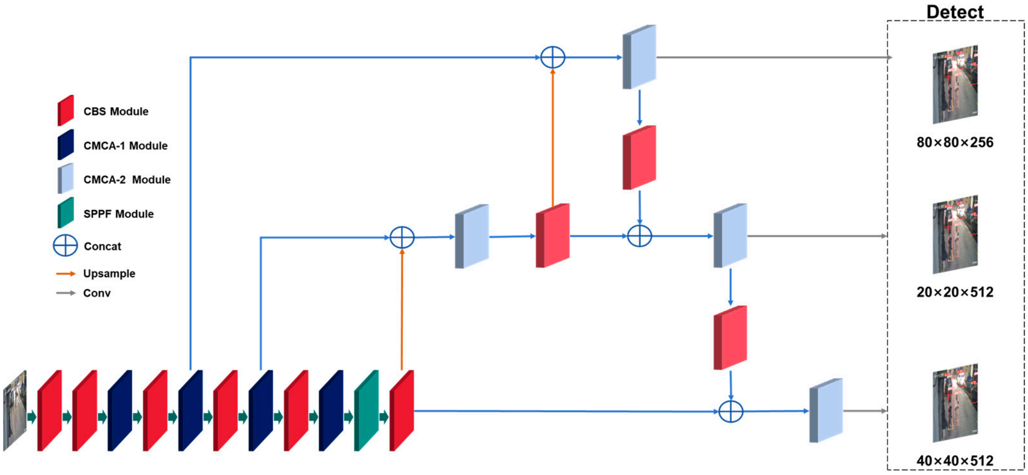


Figure 1. CMCA-YOLO structure, where the network initially extracts features from the input image through the backbone network, paying extra attention to overlapping and small targets via the CMCA module. Subsequently, the neck network performs feature fusion, producing multi-scale outputs, and finally, the detection network effectively identifies objects of various sizes.

At the end of the backbone network, feature maps are further processed through the SPPF module. The SPPF module employs a series of maximum pooling operations to aggregate contextual information from different scales without changing the spatial dimensions of the feature maps. The structure of the SPPF module is shown in Figure 2.

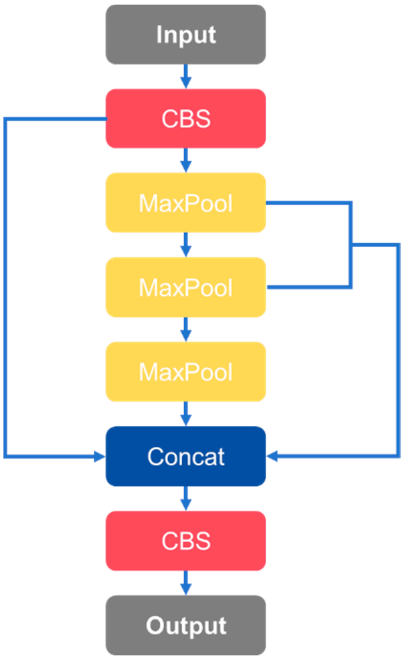


Figure 2. SPPF module, where the feature maps obtained from each pooling operation are aggregated, and features are extracted through a CBA module to produce the output.

Subsequently, feature maps are input into the neck network for feature fusion across different levels. The neck network utilizes the structure of the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) to merge features from various layers, thereby enhancing the model's detection performance for objects of all sizes. FPN integrates feature maps from different levels through upsampling and downsampling operations, creating a multi-scale feature pyramid to address targets of varying sizes and positions within the image. PAN, as part of the neck module, strengthens the connection between high-level semantic features and low-level detail features through top-down and bottom-up feature fusion, optimizing the model's ability to detect small or overlapping targets.

The feature maps output from the neck network is then passed to the detection network, which is responsible for performing object classification and localization tasks. During this process, the model employs anchor mechanisms and binary cross-entropy loss functions to optimize feature maps, achieving precise regression of target positions and high accuracy in classification. After completing object classification and localization, the detection network refines the output using Non-Maximum Suppression (NMS) technology, eliminating overlapping detection boxes to ensure each target's unique positioning.

The CMCA-YOLO model is a finely structured, highly adaptable object detection framework. It not only maintains the rapid detection capabilities of YOLOv5 but also significantly enhances target detection performance in complex environments through the innovative application of the CMCA module.

3.2. CMCA Module

This study delves into the CCA and MSCA mechanisms on which basis the CMCA module is proposed. This module is designed to deeply analyze the contextual information of image features and their complex inter-channel associations, thereby achieving efficient extraction and optimized representation of image features. The specific architecture and implementation details of the module are depicted in Figure 3.

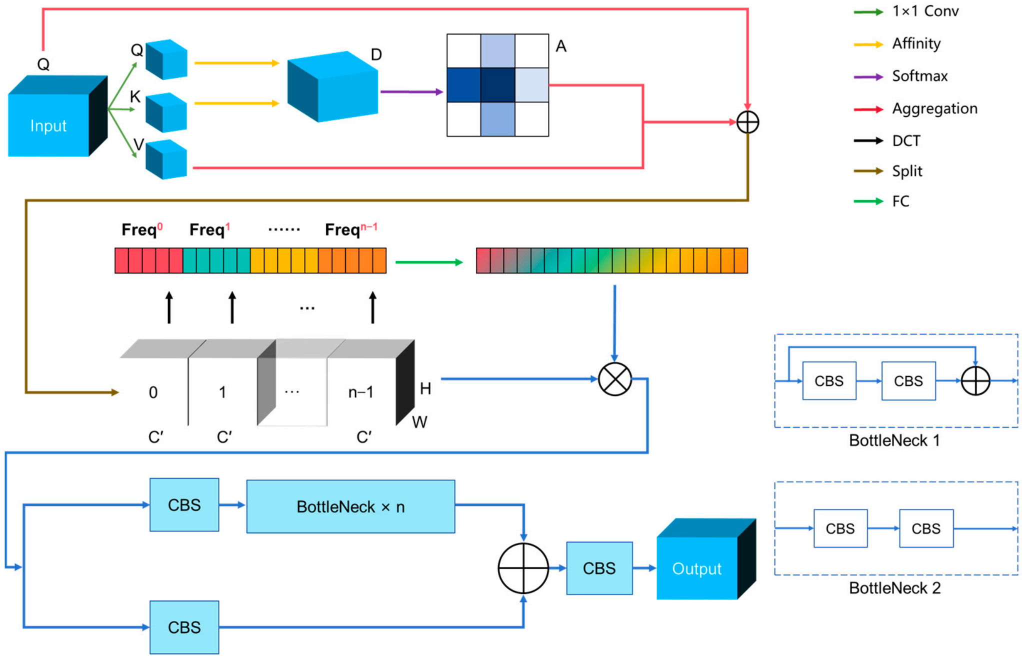


Figure 3. The structure of the CMCA module. BottleNeck 2 contains two consecutive CBS blocks, refining features through layer-by-layer processing while maintaining a balance between network depth and complexity. The difference between BottleNeck 1 and BottleNeck 2 lies in the addition of a residual connection after the two CBS modules in BottleNeck 1, increasing network depth and mitigating the problem of gradient vanishing. The BottleNeck part in CMCA-1 is BottleNeck 1, while in CMCA-2, it is BottleNeck 2.

During the operation of the CMCA module, the CCA enhances the understanding of target boundaries and shapes by aggregating information in both horizontal and vertical directions for each pixel point in the feature map. This mechanism not only improves the model's perception of target details but also provides clearer differentiation for overlapping targets in the scene. The MSCA finely selects and reorganizes channel features in the frequency domain, effectively compressing and enhancing feature representation. By selecting key frequency components, MSCA enhances the network's sensitivity to specific frequency information, which is crucial for capturing minor variations and subtle features. The execution process of MSCA includes the DCT transformation of input feature maps, selection of frequency components, and calculation of attention weights through fully connected layers. Ultimately, by multiplying the weighted and original input feature maps, an enhanced feature representation is formed. The module concludes with a CBS unit and a BottleNeck module, ensuring the effectiveness of feature processing and the stability of network training. Through the Concat operation, the CMCA module integrates features from different levels, ensuring that the module can not only capture key features but also maintain processing efficiency and training stability, ultimately outputting rich and distinctive feature maps.

3.2.1. CCA Mechanism

The CCA mechanism [56] significantly enhances the ability to recognize overlapping targets in surveillance scenes by aggregating contextual information through its unique paths. Specifically, the CCA module collects contextual information for each pixel point along its horizontal and vertical paths. This mechanism generates dimension-reduced feature maps Q and K through a 1 × 1 convolution layer, as well as an adaptive feature map V. Through the Affinity operation, CCA calculates the degree of correlation between positions, producing an attention map A, which is then normalized by a SoftMax layer. The mechanism is detailed in Figure 4.

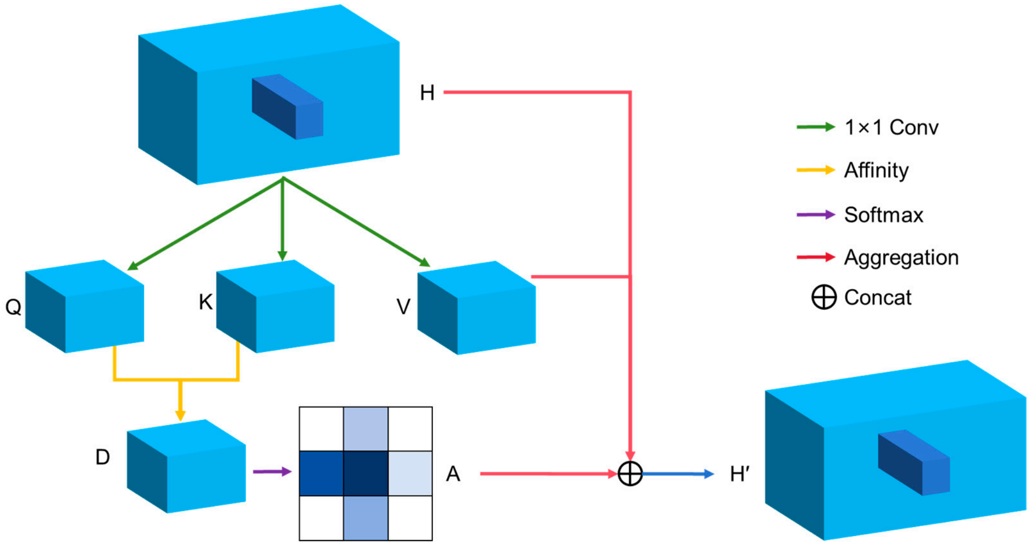


Figure 4. CCA Mechanism.

Given a feature map H $\in R^{C\times W\times H}$, two convolutional layers with 1 × 1 kernels first generate feature maps Q and K, where $\{Q,K\} \in R^{C'\times H\times W}$, and C' is the number of channels, satisfying $C' < C$. After obtaining feature maps Q and K, an attention map A $\in R^{(H+W-1)\times W\times H}$ is generated through the Affinity operation. At each position u in the spatial dimension of feature map Q, a vector $Q_u \in R^{C'}$ is obtained, and then a set $\Omega_u \in R^{(H+W-1)\times C'}$ is derived by extracting feature vectors from K that are in the same row or column as position u, where $\Omega_{u,i} \in R^{C'}$ is the ith element of Ω_u. The definition of Affinity is as shown in Equation (1):

$$d_{i,u} = Q_u \Omega_{i,u}^T \tag{1}$$

where $d_{i,u} \in D$ represents the degree of correlation between feature Q_u and $\Omega_{i,u}$, with $i = [1, \ldots, |\Omega_u|]$, and $D \in R^{(H+W-1)\times W\times H}$. A SoftMax layer is then applied to D along the channel dimension to calculate the attention map A.

Another 1 × 1 filter convolutional layer is applied to H to generate $V \in R^{C\times W\times H}$ for feature adaptation. At each position u in the spatial dimension of feature map V, a vector $V_u \in R^C$ and a set $\Phi_u \in R^{(H+W-1)\times W\times H}$ are obtained. The set Φ_u is a collection of feature vectors from V that are in the same row or column as position u. Contextual information is acquired through the Aggregation operation, as shown in Equation (2):

$$H'_u = \sum_{i\in|\Phi_u|} A_{i,u}\Phi_{i,u} + H_u \tag{2}$$

where H'_u is the feature vector at position u in the output feature map $H' \in R^{C\times W\times H}$. $A_{i,u}$ is the scalar value of channel i and position u in A. Contextual information is added to local features to enhance local representations and integrate context and global information selectively based on spatial attention maps. This provides a broad contextual view and significantly improves the model's ability to accurately classify and locate overlapping pedestrians and vehicles in complex environments like parking lots.

The implementation algorithm for the CCA mechanism, Algorithm 1, transforms the input feature map into query, key, and value feature representations through three parallel 1 × 1 convolutional layers. The features for the query and key are dimensionally reduced, while the value features maintain their original dimensions. The reconstructed query and key feature vectors are used to compute the association of each pixel position in horizontal and vertical directions through batch matrix multiplication, excluding self-association, to avoid feature redundancy. The attention scores obtained by processing these associations with a SoftMax function are then used to weight the value feature vectors, further adjusted by a learnable parameter γ, and added to the original input feature map to produce the final enhanced feature map.

Algorithm 1. Criss-Cross Attention

```
Input: X $\in R^{B\times C\times W\times H}$
Output: Y $\in R^{B\times C\times W\times H}$
Initialize γ as a learnable scale parameter set to 0
Define Conv1×1_query, Conv1×1_key, Conv1×1_value as 1 × 1 convolution layers with input dimension
Define INF as a large negative value (e.g., $-\infty$) for masking
for each x ∈ X do
    Q ← Conv1×1_query(x)
    K ← Conv1×1_key(x)
    V ← Conv1×1_value(x)
    $Q_H$, $Q_W$ ← reshape($Q$) for horizontal and vertical attention
    $K_H$, $K_W$ ← reshape($K$) for horizontal and vertical attention
    $A_H$ ← softmax(bmm($Q_H$ , $K_H$) + repeat(INF, B × W, H, H))
    $A_W$ ← softmax(bmm($Q_W$ , $K_W$))
    $V_H$, $V_W$ ← reshape(V) for horizontal and vertical attention
    $O_H$ ← bmm($A_H$ , $V_H$)
    $O_W$ ← bmm($A_W$ , $V_W$)
    Y ← Y + γ ∗ ($O_H$ +$O_W$)
end for
return Y
```

The CCA module enables each pixel to extract information from all other pixels in the image through a recursive operation. In this way, the feature map gains richer contextual information with each iteration. This mechanism offers a wide view and selectively integrates context and global information based on spatial attention maps when dealing with small or overlapping targets, greatly enhancing the model's feature representation capabilities.

3.2.2. MSCA Mechanism

The MSCA Mechanism [57] views channel representation as a compression process using frequency analysis, employing DCT to compress channel information in the frequency domain. By analyzing multiple frequency components, it achieves richer channel representations. The typical two-dimensional DCT basis function is shown in Equation (3), and the two-dimensional DCT can be expressed in Equation (4).

$$B_{h,w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right) \tag{3}$$

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} B_{h,w}^{i,j} \tag{4}$$

where $h \in \{0,1,\cdots,H-1\}, w \in \{0,1,\cdots,W-1\}$ is the two-dimensional DCT spectrum, $x^{2d} \in R^{H\times W}$ is the input, H is the height of x^{2d}, and W is the width of x^{2d}.

The specific details of the Multi-Spectral Attention Mechanism are illustrated in Figure 5.

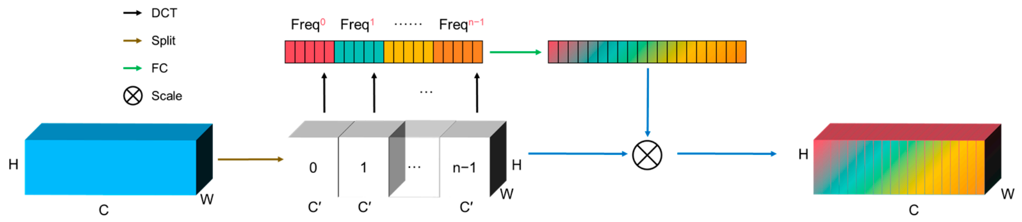


Figure 5. Schematic of the Multi-Spectral Attention Mechanism, which captures spectral features within channels by selecting specific frequency components, thereby obtaining richer and more effective channel representations.

The Multi-Spectral Attention Mechanism first divides the input X along the channel dimension into multiple parts, denoted as $[X^0, X^1, \cdots, X^{n-1}]$, where $X^i \in R^{C'\times H\times W}$, $i \in \{0,1,\cdots,n-1\}$ and $C' = \frac{C}{n}$, with C divisible by n. For each part, corresponding two-dimensional DCT frequency components are allocated. The result of the two-dimensional DCT can serve as the compressed result of channel attention, thus leading to Equation (5).

$$\mathrm{Freq}^{\mathrm{i}} = \mathrm{2DDCT}^{\mathrm{u_i,v_i}}\left(\mathrm{X^i}\right) = \sum_{\mathrm{h=0}}^{\mathrm{H-1}}\sum_{\mathrm{w=0}}^{\mathrm{H-1}} \mathrm{X}^{\mathrm{i}}_{:,\mathrm{h,w}}\mathrm{B}_{\mathrm{h,w}}^{\mathrm{u_i,v_i}} \tag{5}$$

where $i \in \{0,1,\cdots,n-1\}$, $[u_i, v_i]$ are the two-dimensional frequency component indices corresponding to X^i, $\mathrm{Freq}^i \in R^{C'}$ is the compressed C'-dimensional vector. The entire compressed vector can be obtained by concatenation, leading to Equation (6).

$$\mathrm{Freq} = \mathrm{compress}(\mathrm{X}) = \mathrm{cat}\left(\left[\mathrm{Freq}^0,\ \mathrm{Freq}^1,\cdots,\mathrm{Freq}^{\mathrm{n}-1}\right]\right) \tag{6}$$

where $\mathrm{Freq} \in R^C$ is the resulting multi-spectral vector. The entire Multi-Spectral Channel Attention framework can be expressed as in Equation (7).

$$MSA = \sigma(\mathrm{fc}(\mathrm{Freq})) \tag{7}$$

This process enriches the model's ability to distinguish relevant channel information, utilizing both local and global spectral features, thereby improving model performance in various visual tasks. This advanced representation captures complex inter-channel relationships and spectral features, significantly enhancing the network's ability to process and utilize channel information effectively. The implementation algorithm for the Multi-Spectral Attention Mechanism, as shown in Algorithm 2, constructs a custom filter set based on the DCT basis functions corresponding to selected frequency indices. This layer multiplies the input feature map with these filters element-wise, then sums them in the spatial dimension to extract corresponding frequency domain features. Finally, these features are used to generate attention weights for each channel, which are expanded and applied to the input feature map, enhancing the model's sensitivity and discrimination in specific frequency bands.

Algorithm 2. Multi-Spectral Channel Attention

```
Input: X $\in R^{(N\times C\times H\times W)}$, channel, $dct_h$, $dct_w$, reduction
Output: Y $\in R^{(N\times C\times H\times W)}$
$mapper_x,\ mapper_y \leftarrow$ GetFreqIndices(FreqSelMethod)
$mapper_x,\ mapper_y \leftarrow$ scale_indices($mapper_x\ ,\ mapper_y,\ dct_h/7,\ dct_w/7$)
$DCT_{filter} \leftarrow$ GetDctFilter(channel, $dct_h\ ,\ dct_w,\ mapper_x,\ mapper_y$)
FC ← Sequential(Linear(channel, channel/reduction), ReLU, Linear(channel/reduction, channel), Sigmoid)
Initialize Y as a tensor of zeros with the same shape as X
for each feature map x in X do
    if h ≠ dct_h or w ≠ dct_w then
        $x_{resized} \leftarrow$ AdaptiveAvgPool2d(x, $dct_h\ ,\ dct_w$)
    else
        $x_{resized} \leftarrow$ x
    y_dct ← apply DCT_filters to x_resized and sum over the last two dimensions
    attention_weights ← FC(y_dct).reshape(N, C, 1, 1)
    Y ← Y + (x ∗ AttentionWeights)
end for
return Y
```

The Multi-Spectral Attention Mechanism offers a novel form of channel attention by filtering and emphasizing features in the frequency domain, significantly improving deep learning model performance, especially in recognizing targets with complex spectral characteristics. By incorporating the Multi-Spectral Attention Mechanism, the CMCA-YOLO model effectively enhances feature extraction capabilities for pedestrian targets and overlapping vehicle targets in parking lot environments, thereby improving model performance.

4. Experiments

4.1. Dataset

To support and validate the proposed CMCA-YOLO model, this study has constructed a comprehensive dataset of parking lot surveillance images. This dataset is designed to provide a broad testing platform for the development and evaluation of real-time object detection algorithms. It contains 4502 high-resolution images, manually annotated by professional annotators, to ensure high-quality authenticity and precision. The dataset includes 15,818 instances of motor vehicles and 3944 pedestrian instances, collected from various times of the day and under diverse weather conditions to reflect the diversity and uncertainty of real-world environments. The image collection originates from fixed surveillance cameras, covering different types of urban parking lot scenes to reflect the natural distribution of vehicles and pedestrians in daily life. This includes variations in vehicle models, colors, and sizes, as well as parking methods, pedestrian traffic, and vehicle dynamics.

Particularly, the dataset also focuses on special cases that pose challenges to object detection algorithms, such as complex occlusions, changes in lighting and shadows, and target distortions caused by camera angles. The purpose of this approach is to ensure the adaptability and robustness of the developed models, enabling them to perform efficiently and accurately in practical applications. Furthermore, the dataset is designed to evaluate the model's ability to detect parking statuses, monitor illegal occupancy, and track pedestrian behavior in specific tasks. Through extensive coverage of these complex scenes, the dataset not only serves as a basis for model performance evaluation but also as a foundation for algorithm improvement and optimization.

Figure 6 displays sample images from the dataset, revealing typical features and challenges of surveillance scenes and providing researchers with a comprehensive perspective to understand and address issues in real-time object detection.

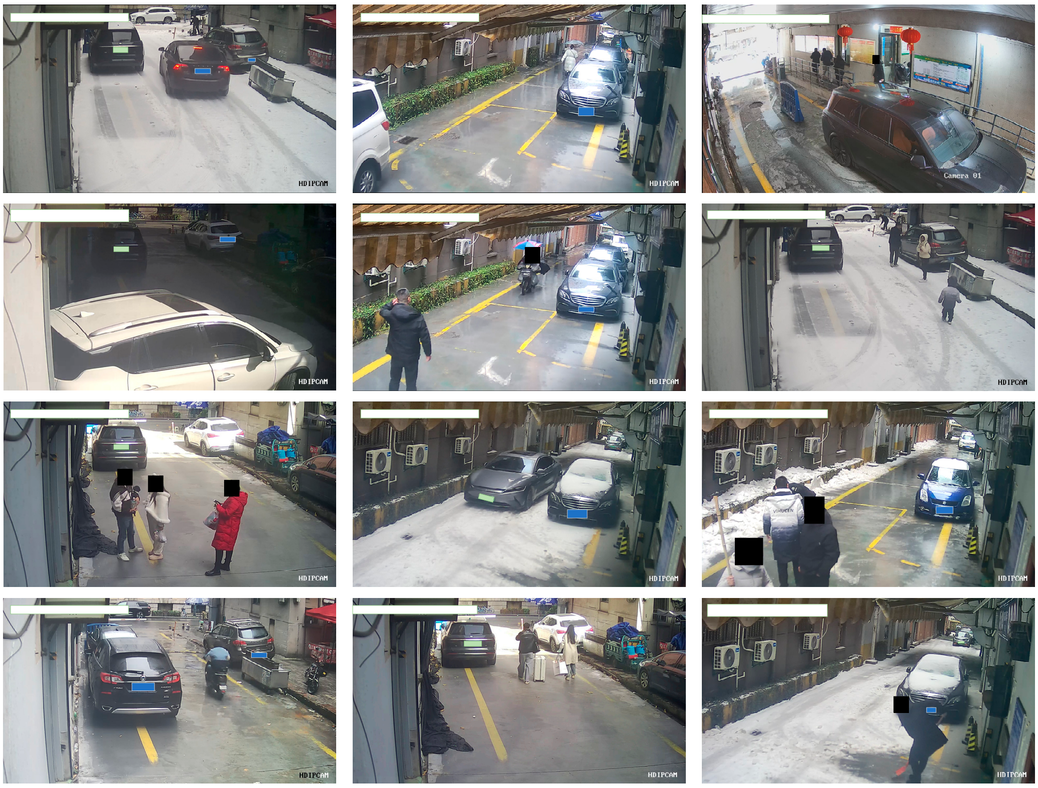


Figure 6. Dataset Samples.

As shown, the dataset samples showcase parking lot scenes captured by multiple cameras, including vehicles and pedestrians under various time points and lighting conditions. These image samples provide the model with rich learning resources to achieve efficient and accurate object detection under real-world conditions.

4.2. Experiments and Testing

This study conducted experiments and tests on a custom dataset with the train, validation, and test set ratio set to 7:2:1. Experimental configurations included the use of an NVIDIA GeForce RTX 3060 Laptop GPU, a 12th Gen Intel(R) Core(TM) i9-12900H 2.50 GHz CPU, DDR4 16G 4800HZ memory, Windows 11 operating system, and software environment of Python 3.9 paired with Pytorch 1.9.0, CUDA11.1, and cudnn8.0. Model training parameters are listed in Table 1.

Table 1 presents key parameters for model training, including epochs, batch size, input image size, and choice of optimizer. These parameters were finely tuned to maximize model learning efficiency and performance.

The model's performance evaluation employed key metrics such as Precision, Recall, Average Precision (AP), and Mean Average Precision (mAP). Precision measures the model's ability to correctly identify positive predictions among all positive class predictions, while Recall reflects the model's efficiency in recognizing actual positive class samples.

AP is the mean value of the area under the precision–recall curve across various decision thresholds, assessing the model's accuracy in predicting positive classes and its comprehensive performance regarding recall levels. mAP, as an evaluation standard across multiple categories, provides a comprehensive perspective to measure the overall performance of the model.

Table 1. Model training parameters.

Parameters	Value
Epoch	120
Batch size	8
Input image Size	640 × 640
weight decay	0.0005
Initial learning rate	0.01
Momentum	0.937
Optimizer	SGD
Workers	8
Data Augmentation	Mosaic

Precision and Recall are calculated as per Equations (8) and (9).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

where TP (True Positives) denotes the number of correctly predicted positive class samples, FP (False Positives) indicates the number of negative class samples incorrectly predicted as positive, and FN (False Negatives) refers to the number of positive class samples incorrectly predicted as negative.

AP, calculated as per Equation (10), measures model performance by averaging precision over multiple threshold settings.

$$\text{AP} = \int_0^1 \text{p}(\text{r})\, \text{dr} \tag{10}$$

mAP is calculated based on AP values across all categories, averaged to obtain the final value as per Equation (11).

$$\text{mAP} = \frac{1}{\text{N}} \sum_{i=1}^{\text{N}} AP_i \tag{11}$$

where N is the number of categories, and AP_i is the Average Precision for the ith category.

FPS (Frames Per Second), an essential metric for evaluating object detection model performance, indicates the model or system's capability to process video streams. A higher FPS value suggests a stronger video stream processing capability.

In the course of rigorous experimental investigations, the CMCA-YOLO model underwent stringent testing, the results of which, as illustrated in Figure 7, underscore the model's superior performance in multi-class object detection tasks. These evaluations encompassed a diverse array of automotive dimensions, from sedans to SUVs, as well as pedestrian scenarios ranging from solitary individuals to groups, demonstrating the model's robust adaptability in varied environments. Particularly under challenging conditions such as partial occlusions, significant variations in lighting, and complex backgrounds, the CMCA-YOLO model exhibited its exceptional capability in maintaining high recognition rates and precision in localization.

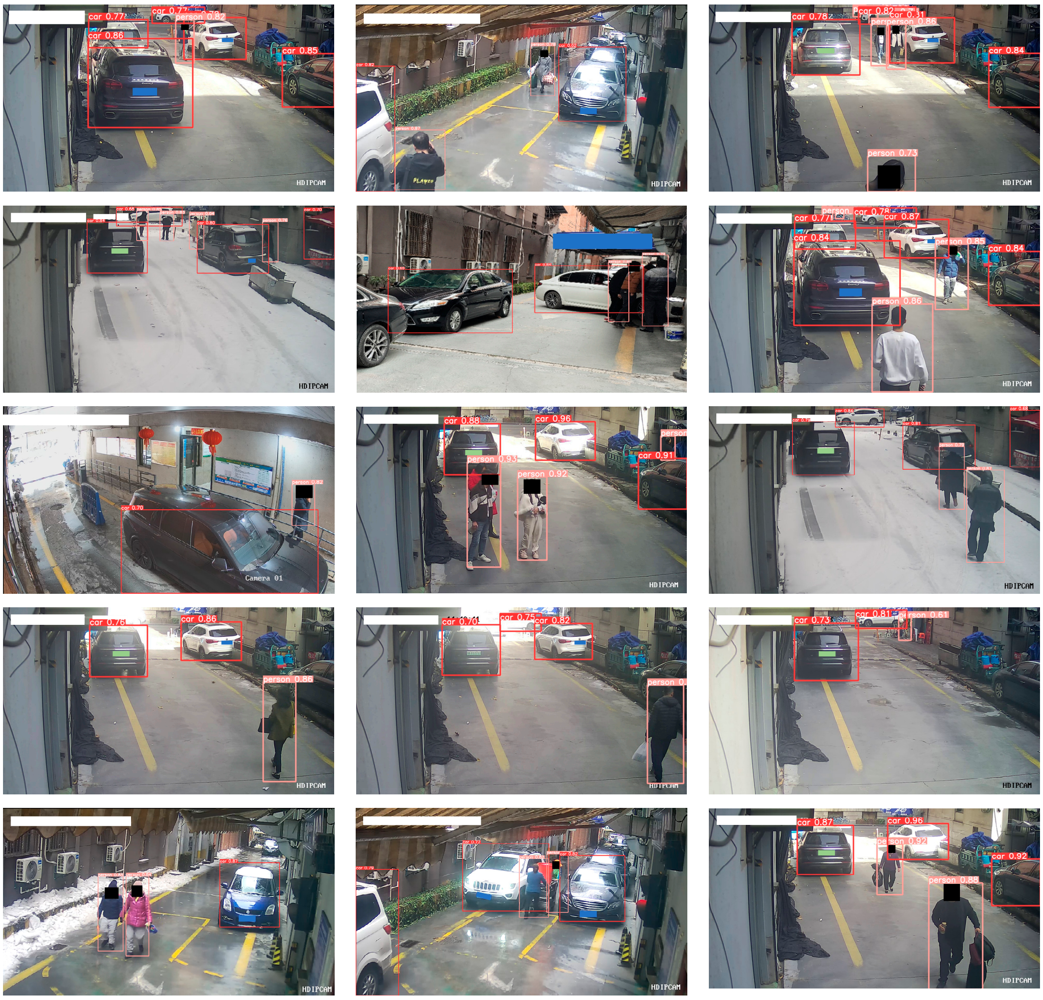


Figure 7. Model testing diagram.

The CMCA-YOLO model adeptly identified and located vehicles and pedestrians of various sizes across multiple surveillance video frames. Notably, even under conditions of partial occlusion and fluctuating lighting, the model displayed outstanding detection performance. This not only attests to the model's high degree of robustness but also confirms its adaptability to complex backdrops and dynamic settings.

In the comparative experimental section of this study, to comprehensively assess the performance of the CMCA-YOLO model, a detailed selection of current leading object detection models was meticulously compared. These models, including SSD, RetinaNet, Faster-RCNN, Cascade R-CNN, ATSS, FSAF, and YOLOv8s, are widely regarded as benchmarks of excellence within the domain of object detection. This paper employed two pivotal performance metrics, namely, model mAP@0.5 and FPS, which respectively measure the model's detection accuracy and processing speed, serving as crucial indicators for evaluating the efficacy of real-time object detection systems in practical applications.

The data displayed in Table 2 elucidate the performance of various models with respect to the metrics of mAP@0.5 and FPS. It is evident from the data that the CMCA-YOLO model proposed in this study surpasses all comparative models with a mAP@0.5 score of 0.895, not only evidencing the significant advantage of the proposed model in detection accuracy but also highlighting its exceptional capability in processing small and overlapping targets in parking lot surveillance imagery. In terms of frame rate, the CMCA-YOLO model achieves a result of 143.7 FPS, on par with YOLOv8s, which is currently recognized as the benchmark for speed. Moreover, despite SSD exhibiting a faster processing speed in terms of frame rate, its performance on mAP@0.5 is significantly lower than that of the CMCA-YOLO model. This disparity underscores the CMCA-YOLO model's ability to maintain high detection accuracy while still achieving rapid detection speeds, which is crucial for real-time responses in practical application scenarios.

Table 2. Comparative Performance of Classic Models and the Proposed Model.

Method	mAP@0.5	FPS
SSD [22]	0.783	120.7
RetinaNet [24]	0.853	73.2
Faster-RCNN [12]	0.817	85.3
Cascade R-CNN [58]	0.832	63.6
ATSS [59]	0.836	110.7
FSAF [60]	0.824	115.4
YOLOv8s [21]	0.872	153.7
Our Model	0.895	143.7

To dissect the contribution of each component of the CMCA-YOLO model, this study conducted a series of ablation experiments aimed at individually and collectively assessing the impact of the CCA Mechanism (M1) and the MSCA Mechanism (M2) on the overall performance of the model. This approach revealed the specific contributions of each module to enhancing detection accuracy and evaluated their potential impact on the model's speed.

Data in Table 3 meticulously demonstrates the performance of models with different component integrations. When the Cross-modal Attention Mechanism (M1) was integrated alone, the model's mAP@0.5 increased from 87.0% with the original YOLOv5s model to 88.1% and further to 88.3% upon the introduction of the Multi-spectral Attention Mechanism (M2), indicating that each independent mechanism positively contributes to performance enhancement. Combining both mechanisms to form the CMCA-YOLO model significantly improved performance, maintaining vehicle detection accuracy while significantly increasing pedestrian detection accuracy from 74.8% to 79.8%, with only a slight decrease in frame rate from 149.6 FPS with YOLOv5s to 143.7 FPS. These results indicate that the proposed CMCA module significantly enhances the model's ability to recognize small and overlapping targets while maintaining a high frame rate, which is vital in dynamic and complex surveillance scenarios.

Table 3. Ablation Study Results.

Method	Class		mAP@0.5	FPS
	Car (AP)	Person (AP)		
YOLOv5s	0.992	0.748	0.870	149.6
YOLOv5 + M1	0.992	0.771	0.881	144.8
YOLOv5 + M2	0.992	0.775	0.883	146.5
Our Model	0.992	0.798	0.895	143.7

In M2, each channel is represented through a set of selected frequency components obtained via DCT computations. We have introduced three criteria for the selection of frequency components: Low Frequency-based selection (LF), Two-Step selection (TS), and Neural Architecture Search selection (NAS). The LF criterion is predicated on the assumption that lower frequency components typically contain more significant information; the TS criterion selects the best-performing components by evaluating the independent impact of each frequency component; and the NAS criterion employs Neural Architecture Search technology to automatically identify the optimal combination of frequency components. The NAS approach utilizes Neural Architecture Search technology to autonomously discover the most suitable frequency component combination for a given task and dataset. Through exploration and optimization during the training process, NAS is capable of identifying which frequency components are most appropriate for the task at hand. The objective of this method is to uncover optimized frequency combinations that may be overlooked by manual methodologies through an automated search process. However, the NAS method typically requires substantial computational resources and time. Compared to manual selection or simple selection strategies, NAS demands a more complex implementation mechanism, including search strategies, performance evaluation, and the determination of final selections. In practical testing, we found no significant performance disparity among the LF, TS, and NAS methods; thus, the LF method was adopted for this paper.

To further evaluate the performance of the proposed CMCA-YOLO model, this study utilized Gradient-weighted Class Activation Mapping (Grad-CAM) technology [61] for visual analysis of the attention mechanisms of CMCA-YOLO compared to YOLOv5s, as shown in Figure 8. In this visual analysis, areas highlighted in red within the heatmaps signify the model's focal points, showcasing the concentration of attention during object detection tasks. This intuitive display allows for the observation that the CMCA-YOLO model exhibits a higher degree of focus in object detection compared to YOLOv5s. Specifically, when dealing with smaller or partially occluded targets, the CMCA-YOLO model demonstrates more concentrated attention areas. This finding suggests that the CMCA-YOLO model is capable of effectively extracting key features of targets, enabling precise identification and localization. In contrast, the YOLOv5s model often shows more dispersed attention areas under these circumstances, which could adversely affect the accuracy of target detection.

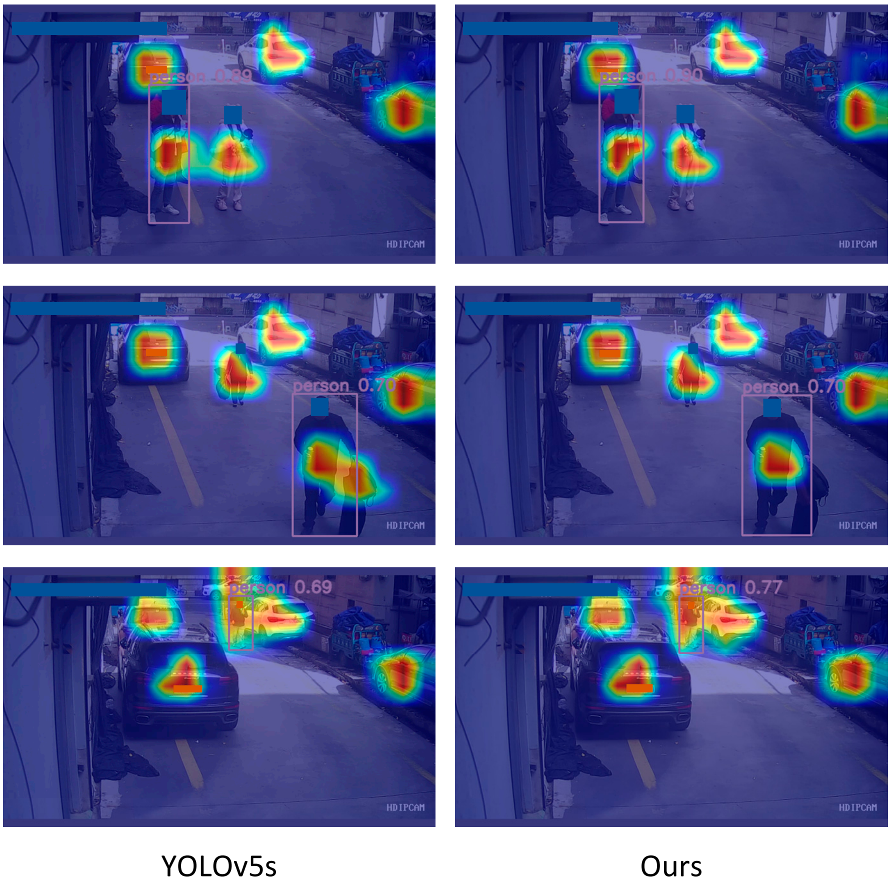


Figure 8. Comparative Heatmaps, with the model proposed in this study on the left and the baseline model on the right. It is evident that the optimized model can more accurately focus on pedestrian targets located in overlapping positions and complex backgrounds.

Through the visualization analysis employing Grad-CAM technology, this study not only intuitively showcases the advantages of the CMCA-YOLO model in object detection tasks but also scientifically confirms its efficiency and accuracy in handling small or occluded targets in complex scenes. These results further elucidate the pivotal role of the CMCA module in enhancing the model's capability to recognize target features.

To comprehensively evaluate the superiority and advancement of the proposed CMCA module in object detection tasks, this study conducted exhaustive comparative experiments covering variants of the YOLOv5 model integrated with different attention mechanism modules. These variants included models integrated with modules such as CBAM, SE, and SimAM, which are considered significant strategies for enhancing model performance in the object detection domain. The comparison was conducted using two key performance indicators, mAP@0.5 and FPS, aiming to unveil the unique advantages of the CMCA module in improving detection accuracy while maintaining processing speed.

The data in Table 4 reveal the significant performance enhancement of the CMCA module compared to other attention mechanism modules. In the vehicle detection category, although all models achieved a nearly perfect recognition rate of up to 99.2%, in the pedestrian detection category, the CMCA-YOLO model outperformed all comparison models with a mAP of 79.8%, significantly better than the YOLOv5 model variants integrated with CBAM, SE, and SimAM modules. Moreover, while maintaining high accuracy, the processing speed of the CMCA-YOLO model (143.7 FPS) was almost equivalent to the YOLOv5s baseline model. This balance demonstrates the CMCA module's significant improvement in the model's ability to recognize small and overlapping targets in complex scenes without sacrificing detection speed.

Table 4. Comparison of the CMCA Module with Other Modules.

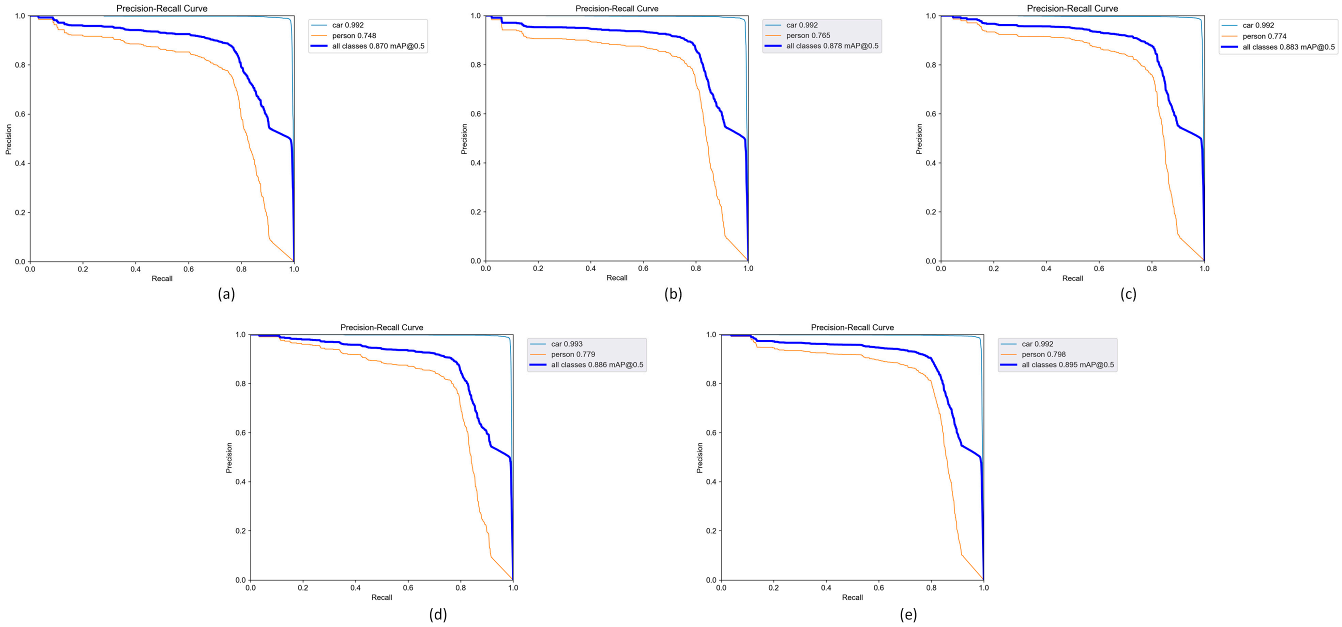
Method	Class		mAP@0.5	FPS
	Car (AP)	Person (AP)		
YOLOv5s	0.992	0.748	0.870	149.6
YOLOv5-CBAM [26]	0.992	0.765	0.878	145.4
YOLOv5-SE [27]	0.992	0.774	0.883	147.8
YOLOv5-SimAM [28]	0.993	0.779	0.886	146.7
Our Model	0.992	0.798	0.895	143.7

The advantage of the CMCA module is particularly prominent in pedestrian detection tasks, which is crucial in dynamic monitoring environments such as parking lots. Pedestrian targets, due to their small size, ease of blending into the surrounding environment, and frequent overlap with other targets, make detection a challenging task. The CMCA module proposed in this study, through the design of cross-modal and multi-spectral channel attention, effectively addresses these issues and exhibits excellent performance in real-world application scenarios.

4.3. Discussion of Experimental Results

In the detailed discussion of this study, we present a comparative analysis of the performance between the baseline model, the CMCA-YOLO model, and other models with modified modules. This comparison is visualized through precision-recall curves (PR curves) and performance evolution graphs during the training process.

As demonstrated by the PR curve comparisons in Figure 9, the CMCA-YOLO model exhibits higher precision and recall rates compared to other model variants. These results not only reflect the statistical superiority of the model but also emphasize its applicability in real-world monitoring scenarios for complex object detection tasks. Notably, the CMCA-YOLO model maintains high precision at higher recall levels, indicating that its detection accuracy and reliability are both precise and dependable.



(a) (b) (c)

(d) (e)

Figure 9. The PR curves for the YOLOv5 baseline model and its improved variants, where (**a**) represents the baseline model curve, (**b**) the YOLOv5-CBAM curve, (**c**) the YOLOv5-SE curve, (**d**) the YOLOv5-SimAM curve, and (**e**) the CMCA-YOLO curve.

The training epoch performance evolution graph in Figure 10, processed through a data smoothing algorithm for the mAP@0.5 performance metric, offers a more intuitive comparison of different models' performances throughout the training process. In this study, a moving average method, as illustrated in Algorithm 3, was employed for data smoothing—a common approach to reducing random variations in data through convolution operations. The smoothed performance curves clearly depict the learning progress and stability of the CMCA-YOLO model, reflecting its rapid adaptation to training data and robust performance during training. The smoothed data indicate that the CMCA-YOLO model quickly reached a high-performance level early in training and maintained this advantage throughout the process. Compared to the baseline YOLOv5s model, the CMCA-YOLO model demonstrated faster learning and more stable performance, which is particularly important for high-performance demands in real-time monitoring applications. The advantages in convergence speed and stability of the model proposed in this paper are attributed to the effective design of the CMCA module, which, through the incorporation of cross-modal and multi-spectral channel attention mechanisms, enhances the model's recognition capabilities in the presence of occlusions and complex backgrounds.

Algorithm 3. Data Smooth

```
Input: y, $box_{pts}$
Output : $y_{smooth}$
y ← np.array(y)
reflection ← y
box ← np.ones(box_pts)/box_pts
if $box_{pts} > 1$ then
    reflection ← concatenate(2∗[0]−y$[box_{pts}$:1:−1], y, 2∗y[−1] − y$[-box_{pts}$:−1:−1])
$y_{smooth-full}$ ← np.convolve(reflection, box, mode='same')
start ←$box_{pts}$ if $box_{pts} > 1$ else 0
end ← $-box_{pts}$ if $box_{pts} > 1$ else None
$y_{smooth} \leftarrow y_{smooth-full}$[start:end]
return $y_{smooth}$
```

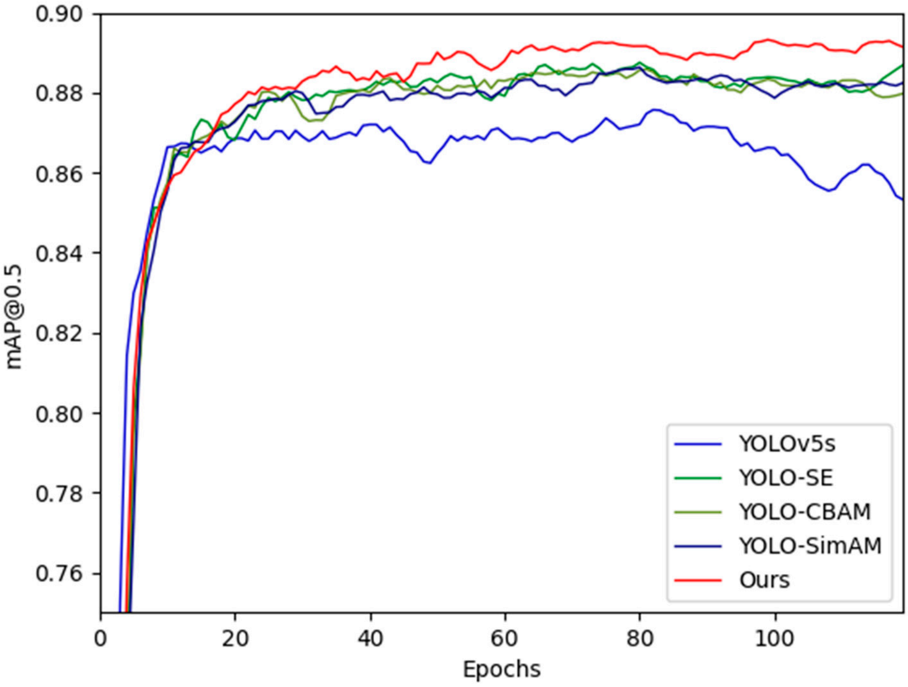


Figure 10. The performance evolution over training epochs revealing that, upon reaching 120 epochs, the models could achieve optimal performance.

The algorithm begins by converting input data into a NumPy array format for numerical operations. Subsequently, when the smoothing window size exceeds one, it addresses boundary issues by appending mirrored reflections of the data at both ends, thus avoiding smoothing biases caused by edge effects. Next, a mean filter is defined and applied to the boundary-treated data through convolution operations, achieving data smoothing. Finally, the portion of the convolution result matching the original data length is extracted as the final smoothed output. This algorithm effectively reduces random fluctuations or noise in the data, particularly demonstrating excellent performance in smoothing at data boundaries.

In practical parking lot surveillance scenarios, the overlap of pedestrians and vehicles is a common challenge. The superior performance of the CMCA-YOLO model in handling such overlapping targets, especially in maintaining high precision at high recall rates, indicates its effective capability in capturing key features and robustness in dynamic environments. This robustness is a direct result of the design of the model, incorporating cross-modal and multi-spectral channel attention mechanisms that enhance the model's perception of key features and significantly improve recognition accuracy without sacrificing processing speed.

The performance of the CMCA-YOLO model exhibits clear advantages across various evaluation metrics. Through comprehensive analysis of PR curves and training epoch performance evolution, this study not only validates the effectiveness of the CMCA module but also demonstrates the model's potential application in real-time monitoring scenarios.

5. Conclusions

This study has designed and successfully developed the CMCA-YOLO model, which, through the integration of the CMCA module, exhibits outstanding performance in object detection within parking lot surveillance images, particularly in the identification of small and overlapping targets. The significant performance enhancement brought by the CMCA module has been validated through a series of comprehensive ablation experiments. Especially noteworthy is its performance on the key metric of mAP@0.5, where it demonstrated broad adaptability and exceptional detection accuracy in diverse monitoring environments.

By merging cross-modal and multi-spectral channel attention mechanisms, the CMCA module not only maintains a high frame rate for real-time processing capabilities but also excels in capturing complex scene details. The module's design thoroughly considers the challenges of monitoring environments, especially in scenarios characterized by unpredictable lighting and frequent intermingling of pedestrians and vehicles, efficiently distinguishing targets from the background to ensure the accuracy and reliability of object detection.

When compared with current leading models, including YOLOv8s and other variants of the YOLO series, the CMCA-YOLO model shows significant advantages across multiple performance metrics, particularly in detection accuracy, which highlights the importance of the CMCA module and the model's high applicability in real-world scenarios. Despite the significant achievements of this study, we recognize that there is still room for improvement in the model's performance under extreme conditions, such as strong glare or complete occlusions. Moreover, specific environmental conditions, such as extreme rain, snow, fog, and the cycle of day and night, remain challenges. Future work will focus on optimizing and refining the model to further explore the application potential of the CMCA module in a wider range of surveillance scenarios, including but not limited to various types and layouts of parking lots, as well as the impact of different weather conditions and diurnal cycles on model performance. Through these explorations, we anticipate that the CMCA-YOLO model will better adapt to complex and variable real-world application scenarios. Additionally, the lightweight modification of the model represents another direction for future research. We will explore advanced model lightweight techniques, including but not limited to network pruning, model quantization, and knowledge distillation, to reduce the computational complexity and enhance the operational efficiency of the model. Considering the trend towards edge computing, we will also investigate integration solutions for the model with edge computing devices, facilitating the model's deployment and application in resource-constrained environments.

Given the remarkable proficiency of large language models in handling complex semantic information and inference tasks, their application in the field of intelligent surveillance, particularly in analyzing video content and understanding scene dynamics, could significantly enhance the intelligence level of monitoring systems [62]. By integrating the advanced capabilities of large language models with the precise target detection offered by the CMCA-YOLO model, we can further improve the surveillance system's understanding of scenes, such as automatically annotating events within surveillance footage, providing richer and more accurate contextual information, thereby facilitating more intelligent event prediction and security monitoring [63]. The introduction of large language models also endows surveillance systems with the ability to process natural language queries, enabling users to search for specific events or objects through simple linguistic commands, thus markedly enhancing the user experience and practicality of the surveillance system [64,65]. Given the rapid advancements in open-vocabulary object detection technology, the innovative applications of the YOLO-World [66] model in this field serve as a valuable reference. Particularly noteworthy are its strategies for handling open vocabularies and enhancing vision-language interactions. Through these comprehensive optimization measures, we aim to propel the widespread application and practice of the CMCA-YOLO model in the realm of intelligent surveillance.

Author Contributions: Conceptualization, N.Z. and H.Z.; Methodology, K.W. and J.Y.; Software, K.W.; Validation, J.Y., F.L. and L.Y.; Formal analysis, F.L.; Investigation, H.Z.; Resources, N.Z.; Data curation, L.Y.; Writing—original draft, K.W. and F.L.; Writing—review & editing, N.Z. and J.Y.; Project administration, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Wuhan University of Technology and Hubei Hongxun Parking Technology Co., Ltd. through their joint research and development project (Project No. 20211h0014).

Data Availability Statement: The dataset utilized in the study is a custom dataset.

Acknowledgments: The authors would like to express their gratitude to Hubei Hongxun Parking Technology Co., Ltd. and Shuqin Zeng for their support and contributions to this research. These supports and contributions were crucial to the advancement of the paper.

Conflicts of Interest: The authors declare that this study received funding from Wuhan University of Technology and Hubei Hongxun Parking Technology Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

References

1. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103–116. [CrossRef]
2. Gowsikhaa, D.; Abirami, S.; Baskaran, R. Automated human behavior analysis from surveillance videos: A survey. *Artif. Intell. Rev.* **2014**, *42*, 747–765. [CrossRef]
3. Verma, K.K.; Singh, B.M.; Dixit, A. A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. *Int. J. Inf. Technol.* **2022**, *14*, 397–410. [CrossRef]
4. Kumar, M.; Ray, S.; Yadav, D.K. Moving human detection and tracking from thermal video through intelligent surveillance system for smart applications. *Multimed. Tools Appl.* **2023**, *82*, 39551–39570. [CrossRef]
5. Swain, M.J.; Ballard, D.H. Color indexing. *Int. J. Comput. Vis.* **1991**, *7*, 11–32. [CrossRef]
6. Gupte, S.; Masoud, O.; Martin, R.F.K.; Papanikolopoulos, N.P. Detection and classification of vehicles. *IEEE Trans. Intell. Transp. Syst.* **2002**, *3*, 37–47. [CrossRef]
7. Stein, G.P. System and Method for Detecting Obstacles to Vehicle Motion and Determining Time to Contact Therewith Using Sequences of Images. U.S. Patent 7,113,867, 26 September 2006.
8. Sun, D.; Roth, S.; Black, M.J. Secrets of optical flow estimation and their principles. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 2432–2439.
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef]
11. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
13. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
16. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
17. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
18. Jocher, G. yolov5. 2021. Available online: https://github.com/ultralytics/yolov5 (accessed on 19 June 2023).
19. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
20. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
21. Jocher, G. yolov8. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 25 June 2023).
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
23. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
24. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
28. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; PMLR, pp. 11863–11874.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
30. 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, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.
33. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef]
34. Chen, K.; Zou, Z.; Shi, Z. Building extraction from remote sensing images with sparse token transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
35. Vijayakumar, A.; Vairavasundaram, S. YOLO-based Object Detection Models: A Review and its Applications. *Multimed. Tools Appl.* **2024**, 1–40. [CrossRef]
36. Ke, R.; Zhuang, Y.; Pu, Z.; Wang, Y. A smart, efficient, and reliable parking surveillance system with edge artificial intelligence on IoT devices. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 4962–4974. [CrossRef]
37. Chen, X.; Wang, M.; Ling, J.; Wu, H.; Wu, B.; Li, C. Ship imaging trajectory extraction via an aggregated you only look once (YOLO) model. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107742. [CrossRef]
38. Nguyen, D.L.; Vo, X.T.; Priadana, A.; Jo, K.H. YOLO5PKLot: A Parking Lot Detection Network Based on Improved YOLOv5 for Smart Parking Management System. In *International Workshop on Frontiers of Computer Vision*; Springer Nature: Singapore, 2023; pp. 95–106.
39. Ogawa, M.; Arnon, T.; Gruber, E. Identifying Parking Lot Occupancy with YOLOv5. *J. Stud. Res.* **2023**, *12*. [CrossRef]
40. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 51094–51112.
41. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
42. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
43. Zhao, Q.; Ma, W.; Zheng, C.; Li, L. Exploration of Vehicle Target Detection Method Based on Lightweight YOLOv5 Fusion Background Modeling. *Appl. Sci.* **2023**, *13*, 4088. [CrossRef]
44. Zhang, Z.D.; Tan, M.L.; Lan, Z.C.; Liu, H.C.; Pei, L.; Yu, W.X. CDNet: A real-time and robust crosswalk detection network on Jetson nano based on YOLOv5. *Neural Comput. Appl.* **2022**, *34*, 10719–10730. [CrossRef]
45. Song, F.; Li, P. YOLOv5-MS: Real-time multi-surveillance pedestrian target detection model for smart cities. *Biomimetics* **2023**, *8*, 480. [CrossRef] [PubMed]
46. Liu, L.; Liang, J.; Wang, J.; Hu, P.; Wan, L.; Zheng, Q. An improved YOLOv5-based approach to soybean phenotype information perception. *Comput. Electr. Eng.* **2023**, *106*, 108582. [CrossRef]
47. Dong, X.; Yan, S.; Duan, C. A lightweight vehicles detection network model based on YOLOv5. *Eng. Appl. Artif. Intell.* **2022**, *113*, 104914. [CrossRef]
48. Li, S.; Yang, X.; Lin, X.; Zhang, Y.; Wu, J. Real-Time Vehicle Detection from UAV Aerial Images Based on Improved YOLOv5. *Sensors* **2023**, *23*, 5634. [CrossRef]
49. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 2778–2788.
50. Sun, Y.; Zhi, X.; Han, H.; Jiang, S.; Shi, T.; Gong, J.; Zhang, W. Enhancing UAV Detection in Surveillance Camera Videos through Spatiotemporal Information and Optical Flow. *Sensors* **2023**, *23*, 6037. [CrossRef]
51. Li, Y.; Fan, Q.; Huang, H.; Han, Z.; Gu, Q. A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. *Drones* **2023**, *7*, 304. [CrossRef]
52. Mahaur, B.; Mishra, K.K. Small-object detection based on YOLOv5 in autonomous driving systems. *Pattern Recognit. Lett.* **2023**, *168*, 115–122. [CrossRef]
53. Qu, Z.; Gao, L.; Wang, S.; Yin, H.; Yi, T. An improved YOLOv5 method for large objects detection with multi-scale feature cross-layer fusion network. *Image Vis. Comput.* **2022**, *125*, 104518. [CrossRef]

54. Omar, N.; Sengur, A.; Al-Ali, S.G.S. Cascaded deep learning-based efficient approach for license plate detection and recognition. *Expert Syst. Appl.* **2020**, *149*, 113280. [CrossRef]
55. Lou, H.; Duan, X.; Guo, J.; Liu, H.; Gu, J.; Bi, L.; Chen, H. DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics* **2023**, *12*, 2323. [CrossRef]
56. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
57. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 783–792.
58. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
59. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9759–9768.
60. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.
61. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
62. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv* **2023**, arXiv:2305.18290. Available online: https://arxiv.org/abs/2305.18290 (accessed on 4 April 2024).
63. de Zarzà, I.; de Curtò, J.; Roig, G.; Calafate, C.T. LLM Multimodal Traffic Accident Forecasting. *Sensors* **2023**, *23*, 9225. [CrossRef] [PubMed]
64. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv* **2023**, arXiv:2301.12597. Available online: https://arxiv.org/abs/2301.12597 (accessed on 4 April 2024).
65. Teterwak, P.; Sun, X.; Plummer, B.A.; Saenko, K.; Lim, S.-N. CLAMP: Contrastive Language Model Prompt-tuning. *arXiv* **2023**, arXiv:2312.01629. Available online: https://arxiv.org/abs/2312.01629 (accessed on 4 April 2024).
66. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-Time Open-Vocabulary Object Detection. *arXiv* **2024**, arXiv:2401.17270.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.