remote sensing

MDPI

Article

Hybrid Cross-Feature Interaction Attention Module for Object Detection in Intelligent Mobile Scenes

Di Tian [1,2], Yi Han [2,*], Yongtao Liu [2], Jiabo Li [1], Ping Zhang [2] and Ming Liu [3]

[1] Mechanical Engineering College, Xi'an Shiyou University, Xi'an 710065, China
[2] School of Automobile, Chang'an University, Xi'an 710064, China
[3] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
* Correspondence: hyi@chd.edu.cn

Abstract: Object detection is one of the fundamental tasks in computer vision, holding immense significance in the realm of intelligent mobile scenes. This paper proposes a hybrid cross-feature interaction (HCFI) attention module for object detection in intelligent mobile scenes. Firstly, the paper introduces multiple kernel (MK) spatial pyramid pooling (SPP) based on SPP and improves the channel attention using its structure. This results in a hybrid cross-channel interaction (HCCI) attention module with better cross-channel interaction performance. Additionally, we bolster spatial attention by incorporating dilated convolutions, leading to the creation of the cross-spatial interaction (CSI) attention module with superior cross-spatial interaction performance. By seamlessly combining the above two modules, we achieve an improved HCFI attention module without resorting to computationally expensive operations. Through a series of experiments involving various detectors and datasets, our proposed method consistently demonstrates superior performance. This results in a performance improvement of 1.53% for YOLOX on COCO and a performance boost of 2.05% for YOLOv5 on BDD100K. Furthermore, we propose a solution that combines HCCI and HCFI to address the challenge of extremely small output feature layers in detectors, such as SSD. The experimental results indicate that the proposed method significantly improves the attention capability of object detection in intelligent mobile scenes.

Keywords: intelligent mobile scenes; deep learning; computer vision; object detection; attention mechanism

Citation: Tian, D.; Han, Y.; Liu, Y.; Li, J.; Zhang, P.; Liu, M. Hybrid Cross-Feature Interaction Attention Module for Object Detection in Intelligent Mobile Scenes. *Remote Sens.* **2023**, *15*, 4991. https://doi.org/10.3390/rs15204991

Academic Editors: Claudio Piciarelli, Melanie Vanderhoof, Jong-Eun Ha, Hyoseok Hwang and Ronghui Zhan

Received: 6 August 2023
Revised: 24 September 2023
Accepted: 13 October 2023
Published: 17 October 2023

1. Introduction

The significant advancements in deep learning have greatly propelled computer vision technology and improved its ability to represent environmental features [1–3]. Object detection [4–7], as one of the fundamental tasks in computer vision, aims to identify specific objects in an image and label their bounding boxes. The progress in deep learning has provided a strong foundation for its application in complex scenes. Currently, object detection holds significant potential for diverse fields such as intelligent driving [8–10], remote sensing [11–13], and medical healthcare [14–16].

While computer vision has advanced significantly, there are still some differences between its approach to environmental perception and that of humans. Computers typically conduct direct feature extraction across the entire image [17,18], whereas humans utilize attention mechanisms to focus on areas of interest. Due to these differences in information processing, researchers have been exploring the use of human-like attention mechanisms to enhance computer vision performance. In recent years, extensive research has been conducted to leverage attention mechanisms for improving computer vision, resulting in a series of representative attention algorithms. These methods usually allocate weights in the channel or spatial dimensions, effectively driving network architecture improvements and algorithm performance enhancements. Nonetheless, within this research, we have

identified limitations in the commonly employed attention modules for object detection in intelligent mobile scenes, resulting in insufficient correlation between different features. Therefore, this paper primarily focuses on investigating whether improving the attention modules can lead to better information-focusing ability, consequently enhancing the performance of object detection in intelligent mobile scenes.

Building upon the aforementioned insights, this paper conducts an in-depth analysis of the commonly used attention modules. It becomes evident that the attention modules commonly used in computer vision are generally developed for general computer vision tasks. In the context of object detection tasks, spatial-domain attention mechanisms are closely associated with the precision of object localization, while channel-domain attention mechanisms are closely linked to capturing object features. Consequently, the application of commonly used single-domain attention mechanisms in object detection tasks may have certain limitations.

Secondly, in order to enhance the performance of attention mechanisms, researchers have proposed various effective attention modules that have been widely applied in many algorithms. However, these attention mechanisms often achieve performance improvements at the cost of increased computational complexity. In recent years, researchers have proposed numerous deep self-attention networks, demonstrating the significant potential of attention mechanisms. For instance, PCT (point cloud transformer) [19] enables networks to autonomously learn important relationships among different elements without the need for manual rule design. This aids in capturing global features more effectively than just focusing on local information. T2T-ViT (tokens-to-token vision transformer) [20] introduces a novel model architecture that helps models better capture relationships among different parts within images, providing new insights for research in the visual domain. Additionally, CvT (convolutions to vision transformers) [21] presents a novel hybrid model, allowing CvT to establish connections between different regions in images, capture local information, and thus achieve better performance in image processing tasks. However, the core of such methods is always related to self-attention mechanisms, which require computing correlations between each input position, significantly increasing the computational burden. As computational resources are often limited in mobile scenes, the most commonly used attention mechanisms in such contexts are still methods like ECA (efficient channel attention) [22], which can effectively enhance model performance at a lower computational cost. Therefore, highly complex attention modules currently have certain limitations when applied to intelligent mobile scenes.

Furthermore, in commonly used attention modules like [23] and CBAM (convolutional block attention module) [24] in SE (squeeze and excitation) intelligent mobile scenes, dimensionality reduction operations with side effects are often employed to capture feature relationships and reduce model complexity. To mitigate the side effects caused by dimensionality reduction, ECA utilizes 1D convolution to achieve cross-channel interaction capabilities. However, due to the inherent complexity of current mainstream object detection algorithms, which often have a large number of channels and feature maps with varying sizes, the interaction capabilities between different features using the commonly used attention modules remain relatively limited. As a result, there are certain limitations in the application of these attention modules to object detection tasks.

Based on the preceding analysis, this paper aims to improve the performance of object detection using attention mechanisms. To achieve this, we introduce the HCFI attention module, which consists of two parts: the channel attention module and the spatial attention module. Firstly, we improved the structure of SPP [25] by increasing the pooling kernels, yielding MKSPP. Next, we utilized the structure of MKSPP to enhance the cross-channel interaction capability of the channel attention. This improvement effectively addressed the issues of insufficient synchronous channel interaction range and poor adaptive interaction range adjustment. Consequently, we derived the HCCI attention module, which exhibits superior cross-channel interaction performance. Then, we combine dilated convolutions to improve the cross-spatial interaction capability of the spatial attention, effectively ad-

dressing the issue of insufficient spatial interaction range. This improvement results in the CSI attention module having better cross-spatial interaction performance. Building upon these enhancements, we sequentially combine the two modules to propose the HCFI attention module, which significantly enhances the model's ability to focus on important information in object detection tasks. Ultimately, experimental results have validated the effectiveness of the proposed method for object detection tasks in intelligent mobile scenes. The primary contributions of this paper are summarized as follows:

(1) Building upon the structure of SPP, we introduced the MKSPP module. Subsequently, by leveraging the MKSPP structure, we improved the channel attention, resulting in the HCCI module having better cross-channel interaction performance.
(2) We further enhanced spatial attention by incorporating dilated convolutions, resulting in the CSI module with improved cross-spatial interaction capabilities.
(3) Based on the aforementioned improvements in the channel attention module and spatial attention module, this paper introduces the HCFI attention module. Additionally, for cases where certain detectors have extremely small output feature layers, we propose a solution that combines HCCI with HCFI. We conducted experimental evaluations on various detectors and datasets, and the results validate the effectiveness of the proposed method in object detection tasks.

The rest of the paper includes the following: In the second part, a description of the previous related works is provided. The third part introduces the proposed methodology and explains the improvements in detail. The fourth part presents experimental validation and result analysis. Finally, the fifth part provides a comprehensive conclusion to the paper.

2. Related Work

2.1. Object Detection

Object detection is one of the most important tasks in computer vision. In the early stages, researchers employed manually crafted features in conjunction with classifiers [26,27]. However, the accuracy of these methods was limited by the feature representation approach. In 2014, researchers proposed the deep-learning-based RCNN (Region-Convolutional Neural Network) [28], which brought object detection technology into the realm of deep learning. Subsequently, object detection techniques gained widespread attention, and researchers further developed Fast RCNN [29] and Faster RCNN [30] based on the foundation of RCNN. These methods typically generate candidate bounding boxes and then perform predictions, known as two-stage algorithms. However, two-stage algorithms consume substantial computational resources and have relatively slow processing speeds. To address this, researchers proposed single-stage algorithms like SSD (single-shot multiBox detector) [31], YOLO (you only look once) [32–36], etc. These methods significantly improved processing speed and expanded the applicability of object detection. In recent years, some researchers have argued that anchor-based designs, to some extent, constrain the performance limits of detection algorithms. Therefore, they further proposed anchor-free algorithms such as CornerNet [37] and YOLOX [38], which have effectively enhanced the convenience of object detection algorithms.

2.2. Attention Mechanism

The attention mechanism is one of the pivotal aspects of human vision, enabling individuals to selectively concentrate on regions that are more likely to contain objects. Due to the importance of attention mechanisms in human vision, researchers have conducted extensive studies recently, attempting to leverage attention mechanisms to augment the capabilities of computer vision [39–41]. Based on how networks handle different types of information during modeling, attention mechanisms in current computer vision can be broadly categorized into three types: spatial domain, channel domain, and hybrid domain. Their working principles and classifications are shown in Figure 1.

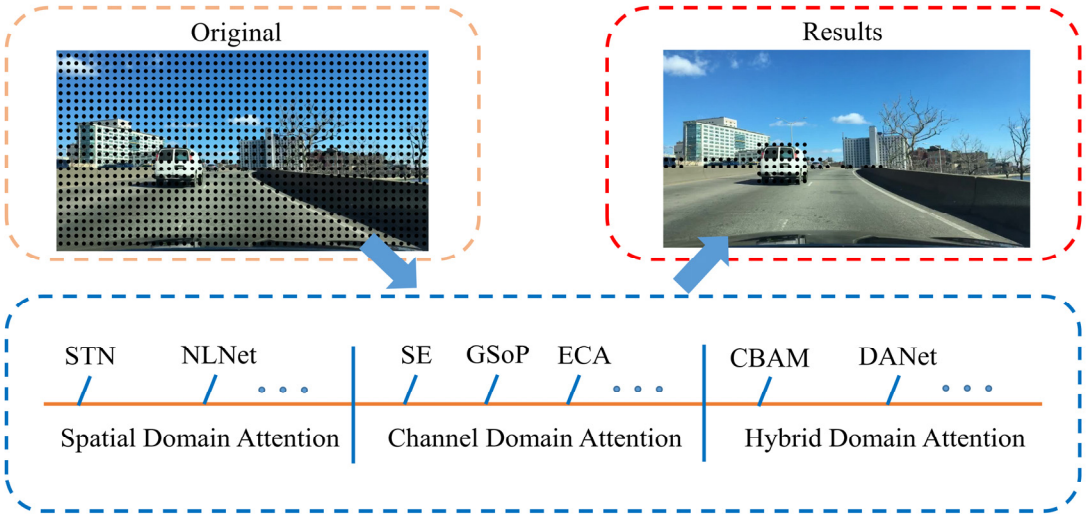

Figure 1. Classification of Attention Mechanisms.

The essence of the spatial domain attention mechanism is a spatial selection mechanism for images. In the early days, researchers introduced the STN (spatial transformer networks) [42] to mitigate the significant information loss caused by pooling. STN is indeed a typical spatial attention mechanism. It can perform arbitrary and complex spatial transformations on input data, thereby enhancing the network's robustness to image deformations, rotations, and other geometric transformations. Additionally, it can be conveniently integrated into general convolutional modules, improving the model's ability to focus on important features in the spatial dimension. However, due to the fact that STN primarily deals with tasks related to spatial geometric transformations, its applicability may be limited when facing other types of tasks. Additionally, STN was initially designed primarily for affine transformations, which might constrain its performance when confronted with more complex geometric transformations. In response to the limitation where each feature point on the feature map can only perceive information from the surrounding corresponding points on the input feature map, researchers have introduced the NLNet (non-local neural networks) module. This module effectively enhances the network's understanding of the global perspective by capturing global information using an attention matrix [43]. However, NLNet's relatively high complexity results in a significant demand for computational resources, limiting the applicability of such methods in intelligent mobile scenes.

The essence of the channel domain attention mechanism is a selection mechanism for images in the channel dimension. The SE module enhances the response of useful features and suppresses irrelevant ones by learning weights between feature channels, thereby improving the quality of feature representation. However, in order to capture dependencies between different channels, it employs dimensionality reduction operations along the channel dimension, which may introduce certain side effects to the model's optimization process. Considering that second-order information can provide richer insights compared to first-order information, researchers have further proposed the GSoP (global second-order pooling) channel attention module [44]. The GSoP module not only takes into account global information but also captures high-order statistical information from the input data. This enables the model to better understand the relationships between features. However, it is precisely because the GSoP module simultaneously considers information at different levels that it significantly increases computational complexity, leading to substantial limitations in its applicability in intelligent mobile scenes. To reduce model complexity and efficiently obtain channel attention, researchers introduced the ECA module, which uses 1D convolution to replace channel dimensionality reduction. This not only reduces model complexity but also effectively avoids the side effects of dimensionality reduction. Compared to common attention mechanisms, ECA offers lower computational costs and can be easily integrated into various CNN architectures without significantly increasing

the computational burden. However, ECA attention modules have limited capabilities in capturing global information, and they do not consider spatial relationships, which may lead to certain limitations in specific situations.

Different tasks have varying demands on attention mechanisms during model training. Object detection tasks require both object recognition ("what") and precise localization ("where"), necessitating the use of both spatial and channel attention. The hybrid domain attention mechanism involves simultaneously constructing spatial attention modules and channel attention modules. To address the issue of incomplete attention mechanisms in both spatial and channel domains, researchers have introduced the CBAM attention module. This module establishes corresponding attention mechanisms separately for both channel and spatial dimensions and then combines them, enabling the model to simultaneously consider feature information in both channel and spatial dimensions. However, CBAM may not necessarily be suitable for all types of tasks, and its channel attention module simultaneously employs two potentially detrimental dimensionality reduction operations. Similarly to the aforementioned methods, researchers have proposed the DANet (dual attention network) module [45] to comprehensively consider channel attention and spatial attention during model training. DANet achieves this by employing two parallel attention modules to independently obtain spatial attention and channel attention. It adaptively integrates the relevance of local and global features, further improving feature representation. However, this approach introduces attention matrices into different attention modules to enhance feature representation, which increases the computational burden of the model, limiting its practical applicability.

Although researchers have developed many attention modules, the current commonly used attention modules for object detection in intelligent mobile scenes still have certain limitations, as indicated by the previous analysis. In contrast to general object detection tasks, intelligent mobile scenes usually involve a small number of core objects, such as pedestrians and vehicles. Algorithms should focus more on these core objects, making attention mechanisms in computer vision highly valuable. There is still significant room for improvement in current research achievements. For this purpose, this paper proposes the HCFI module for object detection, which effectively enhances the ability of object detection in intelligent mobile scenes to focus on important objects.

3. Method

3.1. Multiple Kernel SPP Block

The initially proposed SPPNet aimed to address the issue of varying input image sizes. This network can transform images of different sizes into a fixed size for output. Building upon SPPNet, researchers further introduced the SPP block in subsequent algorithms. This structure achieves feature fusion on different scales, significantly enhancing the performance of object detection. Currently, the SPP block is widely applied in various object detection algorithms. The structure of the SPP block is illustrated in Figure 2a.

While SPP exhibits good feature fusion performance, its computation density is relatively sparse due to the limitations of the pooling pathway. Therefore, we have made improvements to it, and the modified structure of MKSPP in this paper is illustrated in Figure 2b. Building upon the SPP block, we have enlarged the receptive field range of the main features by adding additional pooling paths. This approach further enhances the model's feature fusion capability. Since MKSPP includes more sizes of pooling kernels, it can consider more important features from various interaction ranges.

3.2. Hybrid Cross-Channel Interaction Attention Module

Inspired by the typical channel attention module SE, channel attention modules typically employ dimensionality reduction operations to obtain cross-channel interaction capabilities while reducing model complexity. However, in the context of ECA, its developers held the view that dimensionality reduction could have certain side effects on the model, and using fully connected layers to capture channel dependencies was unnecessary

and inefficient. Therefore, ECA utilizes 1D convolution to attain cross-channel interaction capabilities, effectively improving performance while maintaining low model complexity. To obtain different cross-channel interaction ranges, the ECA attention module incorporates an adaptive adjustment method for the convolution kernel size, as shown in Equation (1).

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \quad (1)$$

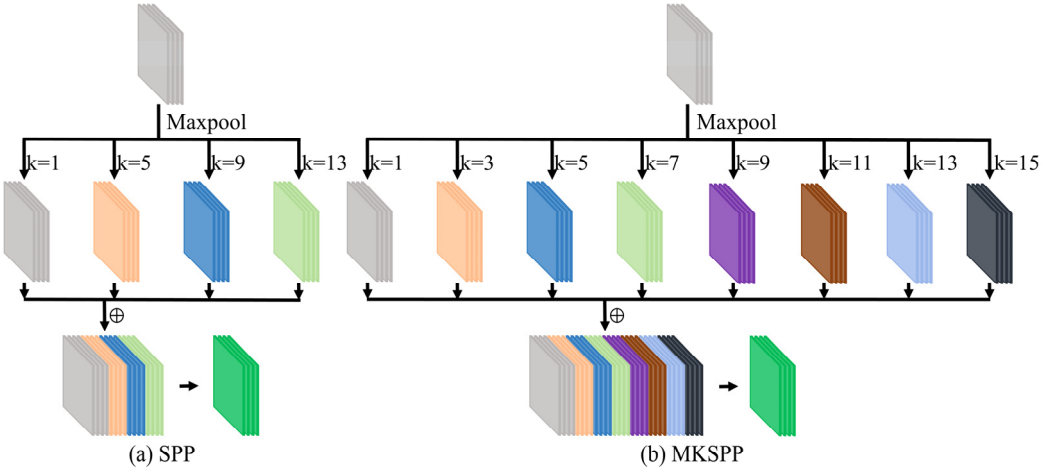

Figure 2. Structure diagrams of SPP and MKSPP.

In Equation (1), C represents the channel dimension, k denotes the convolution kernel size, and $|t|_{odd}$ indicates the nearest odd number to t. In the original paper, γ and b are set to 2 and 1, respectively.

While Equation (1) theoretically allows for adjusting the convolution kernel size based on different channel numbers, this study has revealed a significant limitation when applied to object detection tasks. Common object detection algorithms typically have output channel numbers ranging from 128 to 1024. For instance, SSD's six output feature layers have channel numbers of 512, 1024, 512, 256, 256, and 256; YOLOv4's three output feature layers have channel numbers of 128, 256, and 512; and YOLOX's three output feature layers have channel numbers of 256, 512, and 1024. Based on Equation (1) and the code provided in the original paper, when the output channel dimension falls within the range of 128 to 1024, the 1D convolution kernel size in ECA is consistently set to 5. This situation leads to a problem where the designed adaptive adjustment method cannot effectively modify the kernel size.

To ensure that the network possesses adequate feature extraction capabilities and mitigate the problem of information loss due to excessive dimensionality reduction, common object detection networks typically maintain channel dimensions ranging from 64 to 2048. Most modules operate with channel dimensions primarily between 128 and 1024. However, owing to the severely constrained adaptability of ECA's strategy, the model consistently extracts channel dependencies at a fixed scale. Therefore, for contemporary object detection networks, which typically feature complex architectures, the one-dimensional convolution kernel size in ECA remains almost unchanged, leading to a weakness in its cross-channel interaction capability. To further enhance the focus on important targets in complex intelligent mobile scenes, we propose the HCCI attention module without using high-complexity operations. The structure of the HCCI module is shown in Figure 3.

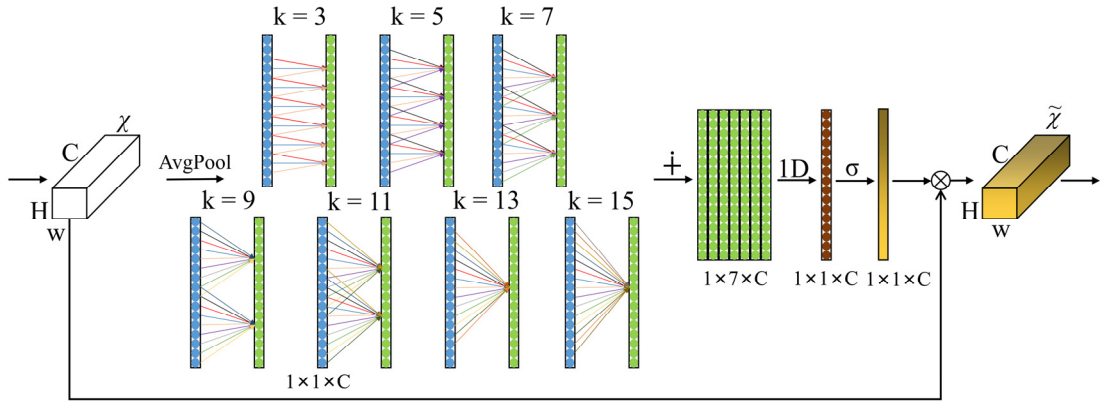

Figure 3. Structure diagram of HCCI.

As shown in Figure 3, we utilize the structure of MKSPP to improve the cross-channel interaction of the ECA module, resulting in the HCCI module having better cross-channel interaction capability. It also utilizes one-dimensional convolution to achieve cross-channel interaction. To avoid the issue of limited adaptive adjustment range or even loss of adjustment effectiveness in object detection, we incorporate multiple convolution kernels of different sizes in HCCI to simultaneously capture dependencies between different channels. The computation method is described in Equation (2).

$$\omega = \sigma\left[\mathrm{C1D}_1\left(\sum_{k}^{K}\mathrm{C1D}_k(y)\right)\right],\ K \in [3,\ 5,\ 7,\ 9,\ 11,\ 13,\ 15] \tag{2}$$

In Equation (2), C1D represents one-dimensional convolution, k denotes the size of the one-dimensional convolution kernel, y represents the result of global average pooling along the channel dimension, σ represents the sigmoid function, and ω represents the final attention weight.

To improve the interaction performance between different elements of the model, many methods utilize attention matrices to calculate attention weights between each input element and all other input elements. This requires the model to store a significant amount of relational information, which significantly increases the complexity of the model. To avoid the computational burden associated with excessive complexity, the proposed HCCI in this paper effectively extracts the most important relational information around input elements using multi-scale interaction kernels without performing extensive computations. In the HCCI attention module, we first use global average pooling to consolidate input features along the channel dimension. Subsequently, we employ multiple one-dimensional convolution operations at various scales, inspired by the concept of the MKSPP structure, which handles different ranges. This approach ensures that the model can extract channel dependencies from varying interaction ranges under any circumstances. It enhances the information foundation for feature selection during the training process. The interaction information between features within different ranges makes it more likely for the model to emphasize crucial information. Based on the above, adjusting the number of convolution kernels can impact the model's performance. More convolution kernels lead to better cross-channel interaction performance, but they also require more computational resources. Finally, the obtained feature set is integrated again using one-dimensional convolution. The integrated features thoroughly consider channel dependencies within different ranges, effectively addressing the issue of limited feature interaction range. This enhancement boosts the cross-channel interaction performance of the object detection model.

3.3. Cross-Space Interaction Attention Module

Due to its lower model complexity and wide applicability, CBAM is one of the most commonly used attention modules at present. It consists of two parts: the channel domain

and the spatial domain. In the spatial attention module, it leverages the spatial relationships of features to generate a spatial attention map. Convolutional operations are used to extract dependencies between features, and ultimately, attention weights are generated using the sigmoid function. The computation method for attention allocation in the CBAM spatial attention module is described in Equation (3).

$$M_s(F) = \sigma\left(f^{7\times7}([F^{s}_{avg}; F^{s}_{max}])\right) \quad (3)$$

In Equation (3), F^{s}_{avg} and F^{s}_{max} represent two feature maps generated using pooling, σ represents the sigmoid function, and $f^{7\times7}$ represents the convolutional kernel size of 7×7.

Although the CBAM spatial attention module employs convolutional operations to achieve some level of cross-spatial information interaction, research has encountered challenges in determining the optimal convolutional kernel size. Smaller kernel sizes have a lower computational cost and are less likely to lose important features, but they have a limited interaction range and may struggle to obtain sufficient cross-spatial information interaction. On the other hand, larger kernel sizes can achieve stronger cross-spatial information interaction performance, yet they are more susceptible to feature loss and result in heightened computational parameters. To strike a balance between these factors, the CBAM spatial attention module employs a convolutional kernel size of 7×7.

However, in current mainstream object detection algorithms, to achieve effective multi-scale detection results, different scale output feature layers are often used simultaneously. When the convolutional interaction range on larger-scale output feature layers remains the same as that on smaller-scale feature layers, it leads to a weakened cross-spatial interaction capability.

However, in current mainstream object detection algorithms, to achieve effective multi-scale detection results, different scale output feature layers are often used simultaneously. When the convolutional interaction range on larger-scale output feature layers remains the same as that on smaller-scale feature layers, it leads to a weakened cross-spatial interaction capability. Furthermore, in object detection networks, there are many feature maps with scales larger than the output layers. If the application scope of the CBAM spatial attention module is extended to the entire network, the problem of weaker cross-spatial interaction capability will be further emphasized. Similar to the previous discussion, improving the spatial attention module using the structure of MKSPP can effectively enhance cross-spatial interaction performance. The improved structure is shown in Figure 4.

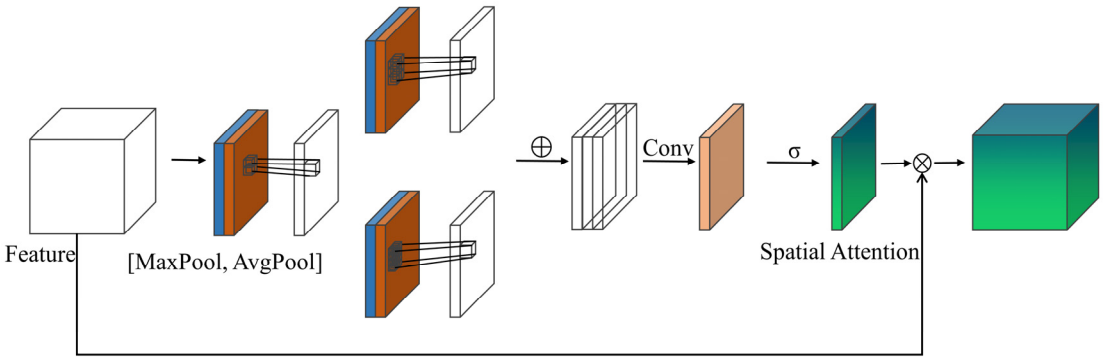

Figure 4. Structure diagram of the improved spatial attention module.

Similar to the previous section, the approach presented in Figure 4 effectively extracts the most critical contextual information around input elements using multi-scale interaction kernels without extensive computations. However, considering the limited computing resources and the high real-time requirements for algorithms in intelligent mobile scenes, in order to reduce computational complexity as much as possible, this paper further improves

the spatial attention module. Without using any additional computations, the CSI attention module is proposed, as shown in Figure 5.

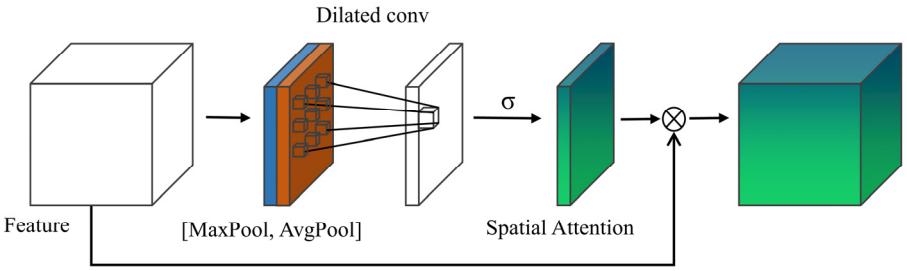

Figure 5. Structure diagram of CSI.

As shown in Figure 5, we replaced regular convolutions with dilated convolutions, effectively increasing the computation range of feature dependencies and obtaining a CSI module with better cross-space interaction capabilities. In our study, we used dilated convolutions with a dilation rate of 2 instead of standard convolutions, and the computation is performed according to Equation (4).

$$M_s(F) = \sigma\left(f_2^{7\times7}([F^{\,s}_{avg};\, F^{s}_{max}])\right) \tag{4}$$

In Equation (4), $f_2^{7\times7}$ represents a 7×7 dilated convolution kernel with a dilation rate of 2, and the remaining terms are consistent with the previous text.

In the CSI attention module, we first integrate input features in the spatial dimension using global max-pooling and global average-pooling operations. Subsequently, we replace the standard convolutions used to capture spatial dependencies in the spatial attention module with dilated convolutions. This approach effectively extends the range of feature selection in the spatial attention module, enhancing the information foundation for feature selection during model training. As a result, it becomes easier to highlight crucial information when dealing with large feature maps in object detection tasks. Additionally, compared to the approach of increasing the convolution kernel size to enhance the interaction range, the CSI module is less likely to lose important features, allowing it to retain more valuable information. Furthermore, adjusting the dilation rate of dilated convolutions can impact the model's performance. A larger dilation rate results in a broader consideration of the interaction range. Considering the typical feature map sizes in object detection models, this study selects a dilation rate of 2. In response to the high real-time requirements of intelligent mobile scenes, our CSI module achieves a larger spatial interaction range without introducing additional computational parameters. This low-complexity improvement is beneficial for the practical application of object detection in intelligent mobile environments.

3.4. Hybrid Cross-Feature Interaction Attention Module

The requirements of the object detection task, which involve object recognition and precise localization, necessitate the incorporation of both spatial attention and channel attention simultaneously. Furthermore, there are higher constraints on the complexity of models in practical intelligent mobile scenes. Based on these considerations, to enhance the focus of the object detection model on important objects in intelligent mobile scenes, we propose the HCFI attention module by combining the aforementioned HCCI attention module and CSI attention module. Its structure is illustrated in Figure 6.

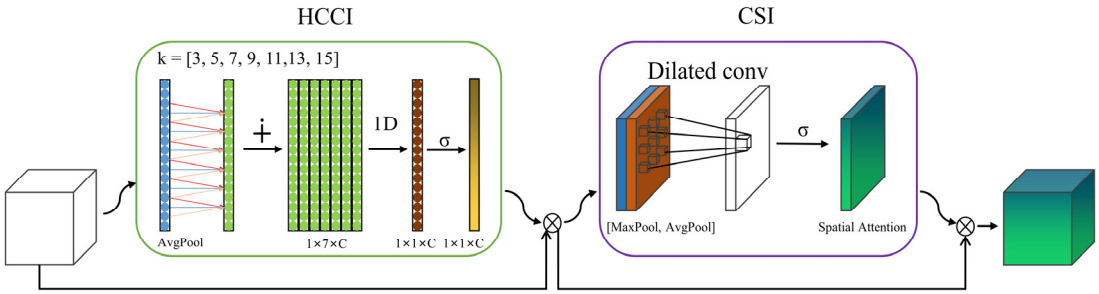

Figure 6. Structure diagram of HCFI.

Figure 6 illustrates the process: once the feature map is fed into the HCFI attention module, it initially goes through the HCCI channel attention module. After integrating the input feature map through global average pooling, it employs a multi-scale interactive kernel to capture channel dependencies across various interaction ranges, followed by feature integration. This approach considers both important information from neighboring channels and the relationships between distant channels, exhibiting a multi-scale cross-channel interactive effect. Following the HCCI channel attention module, the feature map further enters the CSI spatial attention module. After integrating the input features along the channel dimension using two pooling methods, it then employs dilated convolutions to obtain a larger spatial interaction range. This method effectively enhances the interaction range without increasing the model's complexity. The proposed method in this paper improves the focus of object detection on important objects by enhancing the cross-feature interaction capability in both channel and spatial dimensions. The algorithm workflow is shown in Algorithm 1.

Algorithm 1: HCFI Module

Input: Input features with shape $[N_1, C_1, H_1, W_1]$
Output: Output features with shape $[N_3, C_3, H_3, W_3]$
1. Calculate cross-channel interaction features for different interaction ranges:
C3, C5, C7, C9, C11, C13, C15
2. Calculate the integrated channel features: F = Conv1d(C3 + C5 + C7 + C9 + C11 + C13 + C15)
3. Calculate the output features of the HCCI module:
$[N_2, C_2, H_2, W_2] = [N_1, C_1, H_1, W_1] \times$ Sigmoid(F)
4. Calculate the integrated spatial features: F^S_{avg}, F^S_{max}
5. Calculate cross-space interaction features using dilated convolution: $M = f_2^{7\times7}[F^S_{avg}; F^S_{max}]$
6. Calculate the output features of the CSI module:
$[N_3, C_3, H_3, W_3] = [N_2, C_2, H_2, W_2] \times$ Sigmoid(M)

In summary, the proposed HCFI attention module in this paper enhances the performance of object detection by seamlessly combining both channel and spatial attention mechanisms. Building upon the foundations of the HCCI and CSI modules, the HCFI module demonstrated better cross-feature interaction capability without the need for high-complexity operations. This effectively improves the attention capacity of object detection for important information in intelligent mobile scenes.

4. Experiment and Discussion

4.1. Dataset and Detection Algorithm

In this section, we evaluate our proposed method through experiments conducted on three datasets: VOC [46], COCO [47], and BDD100K [48]. The VOC and COCO datasets are widely recognized in the field of object detection, covering diverse object classes frequently encountered in mobile scenes, such as people, bicycles, motorcycles, cars, buses, trucks, trains, traffic lights, and traffic signs. By choosing VOC and COCO datasets for our experiments, we can verify the effectiveness of the proposed method in intelligent

mobile scene object detection techniques and explore its potential applications in other scenes. Additionally, BDD100K is one of the largest and most diverse complex intelligent transportation datasets, and it is among the most authoritative datasets used for object detection tasks in mobile scenes. Choosing BDD100K as an experimental dataset allows us to further validate the effectiveness of the proposed method in intelligent mobile scenes.

To comprehensively evaluate the performance of our proposed algorithm, we employ three well-established object detection algorithms: YOLOX, YOLOv5, and SSD, as evaluation benchmarks. All algorithms in the experiments are implemented using PyTorch 1.7.1. To fully validate the effectiveness of the proposed algorithm, we embed the HCFI attention module into different object detection frameworks for comparative experiments. Considering that different object detection frameworks may have structural differences leading to variations in details, to ensure the fairness of comparative experiments, all other influencing factors except for the attention module are kept consistent within the same object detection framework.

4.2. Experiments on YOLOX

In this section, we conduct experiments on the YOLOX detector. YOLOX is a recently released anchor-free detector that has shown excellent performance on multiple benchmarks. First, we conduct experiments using YOLOX on the COCO dataset. Due to the extensive size of the complete COCO dataset and considering the limited performance of our hardware platform, we use the COCO val-2017 subset as the experimental dataset. This subset comprises 5000 images and 36,781 objects, and it is divided into training, validation, and test sets in an 8:1:1 ratio. To accommodate various application scenes, YOLOX has released multiple versions. In this paper, we used YOLOX-S for the experiments. YOLOX has three output feature layers, and we added attention modules to each output feature layer for optimization. The structure of YOLOX and the added attention modules are illustrated in Figure 7.

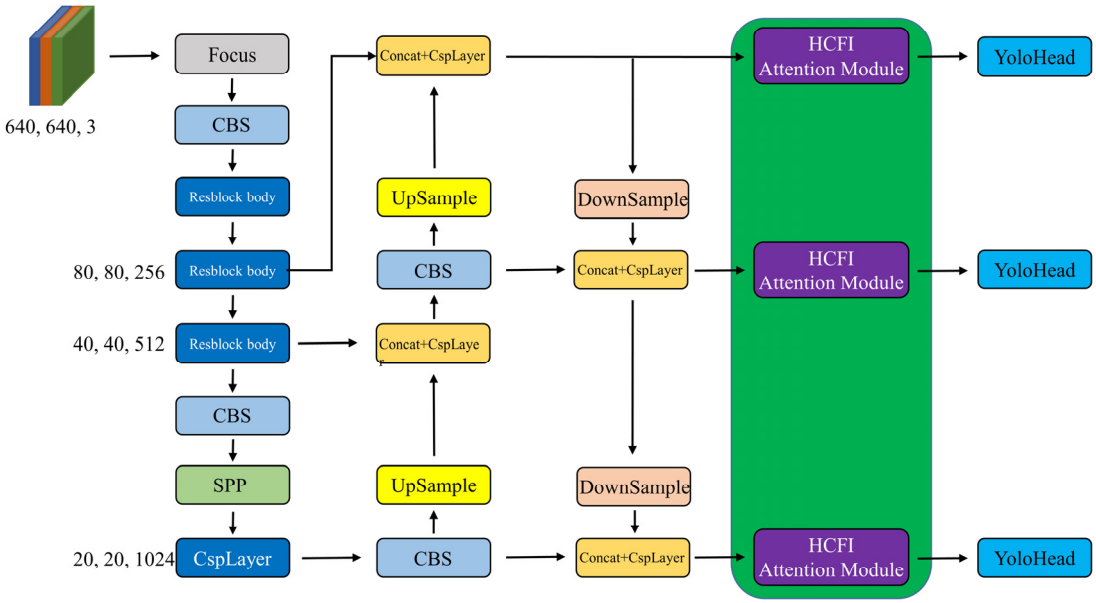

Figure 7. The structure of YOLOX, where the attention modules are highlighted with a green background.

As shown in Figure 7, the added attention modules are highlighted with a green background. In this section, we first validate the stability of the proposed MKSPP structure in YOLOX. Figure 8 provides a visual representation of the experimental outcomes pertaining to the fluctuation in loss attributed to the MKSPP module.

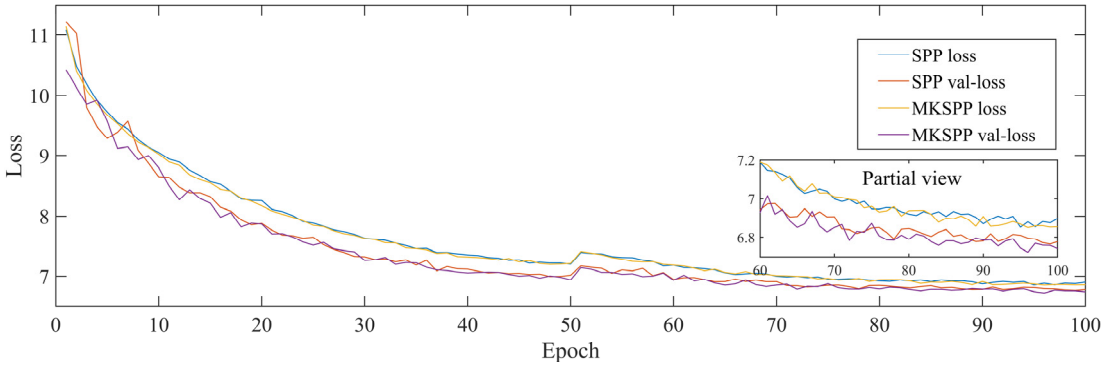

Figure 8. The experimental results regarding the stability of the MKSPP structure.

Upon observing Figure 8, it becomes apparent that the MKSPP structure induces a relatively minor variation in loss values overall. When we zoom in on the latter stages of training, we can observe that the validation loss associated with MKSPP is comparatively lower, effectively substantiating the stability of the MKSPP module. Since the proposed HCFI attention module is composed of two parts: the HCCI channel attention module and the CSI spatial attention module, we conducted ablation experiments to evaluate the effectiveness of each part separately. To ensure the stability of model training, we employed pre-trained weights during the experiments, and the results are shown in Table 1.

Table 1. The ablation experimental results of the proposed method in YOLOX.

Method	HCCI	CSI	HCFI	mAP (%)
YOLOX	×	×	×	59.81
	√	×	×	60.28
	×	√ (+CBAM channel)	×	60.37
	×	×	√	60.52

From Table 1, it becomes evident that all attention modules contribute to improving object detection performance to some extent. Within the YOLOX algorithm, the proposed HCCI and CSI attention modules effectively enhance the model's attention in the channel and spatial dimensions, respectively. Moreover, by combining the HCCI and CSI modules to create the HCFI attention module, the best performance is achieved, resulting in a performance improvement of 1.19% compared to the original version. This effectively demonstrates the ability of the proposed method to enhance the focus of object detection. Building on these insights, to further validate the stability of the proposed HCFI attention module's performance, this section evaluates the performance of different methods using multiple metrics such as mAP50, mAP75, and mAP50:95. The experimental results are shown in Table 2.

Table 2. The performance evaluation results on the YOLOX detector.

Method/ Evaluation	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	FPS
YOLOX (base)	59.81	44.03	40.42	121.33
Relative improv. %				
base + ECA	60.11	44.19	40.80	115.64
Relative improv. %	0.50%	0.36%	0.94%	
base + CBAM	60.28	43.98	40.63	108.28
Relative improv. %	0.79%	−0.11%	0.52%	
base + HCFI	60.52	45.51	41.04	107.46
Relative improv. %	1.19%	3.36%	1.53%	

The results in Table 2 reveal that, on the COCO dataset, the proposed HCFI attention module consistently outperforms commonly used attention modules across three distinct evaluation metrics (mAP50, mAP75, and mAP50:95). Compared to the original YOLOX model, it achieves a noteworthy performance improvement of 1.53% in the comprehensive mAP50:95 metric. Furthermore, we conducted an analysis of the computational speed when incorporating different attention modules into the model. The results indicate that the proposed HCFI approach in this paper demonstrates excellent computational efficiency, thereby validating that our method can enhance focus on important objects while maintaining model computational efficiency.

Building upon the experiments mentioned earlier, in order to thoroughly validate the effectiveness and robustness of our approach, we extended our research by conducting experiments using YOLOX on the BDD100K dataset. Similar to the COCO dataset, and taking into account the extensive size of the BDD100K dataset and the limitations of our experimental platform, we conducted experiments using a subset of 5000 images from its validation set. We partitioned these images into training, validation, and test sets in an 8:1:1 ratio. Because the BDD100K dataset exclusively comprises object instances from traffic scenes and has a substantial number of such instances, it has led to a significant imbalance in the dataset. In this section, we first conducted a statistical analysis of the number of objects in different categories within the experimental dataset. The results of this analysis are presented in Table 3.

Table 3. The number of object instances for different categories in the BDD100K dataset.

Classes	Bicycle	Train	Pedestrian	Truck	Traffic Sign	Car	Bus	Motorcycle	Rider	Traffic Light
Number	494	8	5681	1907	15,710	46,499	728	189	286	11,886

As shown in Table 3, within the dataset used for the experiments in this section, the category with the highest quantity is "car", which includes 46,499 object instances, while the category with the fewest instances is "train", which contains only 8 object instances. There is an immense disparity in sample quantities among different categories. To mitigate the issues caused by a severe sample imbalance during model training, we opted to exclude categories with fewer than 500 instances in our experiments.

Building on the experiments conducted earlier in this section, we initially assessed the performance of various methods on the BDD100K dataset across multiple metrics such as mAP50, mAP75, and mAP50:95. Furthermore, we further validated the performance of our approach using metrics like mAP_{small}, mAP_{medium}, and mAP_{large}. The experimental results are presented in Table 4.

Table 4. Performance evaluation of the YOLOX Detector on the BDD100K dataset.

Method/Evaluation	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	mAP_{small} (%)	mAP_{medium} (%)	mAP_{large} (%)	FPS
YOLOX (base)	35.93	18.48	19.82	3.70	25.45	52.35	125.83
Relative improv. %							
base + ECA	36.08	19.05	20.05	4.54	25.42	52.03	119.62
Relative improv. %	0.42%	3.08%	1.16%	22.70%	−0.12%	−0.61%	
base + CBAM	36.15	18.84	19.99	3.86	25.53	52.21	111.35
Relative improv. %	0.61%	1.95%	0.86%	4.32%	0.31%	−0.27%	
base + HCFI	36.60	19.26	20.39	4.70	26.60	51.00	110.94
Relative improv. %	1.86%	4.22%	2.88%	27.03%	4.52%	−2.58	

The results in Table 4 demonstrate that, on the BDD100K dataset, the proposed HCFI attention module consistently outperforms common attention modules across three distinct evaluation metrics (mAP50, mAP75, and mAP50:95). Compared to the original YOLOX model, it achieves a remarkable performance improvement of 2.88% in the comprehensive

mAP50:95 metric. Furthermore, in the performance testing for objects of different sizes, our approach still excels in two metrics, validating the effectiveness and robustness of our method. Finally, we conducted a computational speed analysis when incorporating different attention modules into the model. The results indicate that the proposed HCFI approach in this paper exhibits excellent computational efficiency. Compared to commonly used attention modules in the context of intelligent mobile scenes, it achieves competitive computational speed, providing strong evidence that our method can enhance focus on important objects while maintaining model computational efficiency.

4.3. Experiments on YOLOv5

Similar to the experiments in the previous section, in order to further validate the effectiveness of the method proposed in this paper, we conducted experiments using the YOLOv5 detector on both the BDD100K dataset and the COCO dataset. The data used in these experiments are the same as in the previous section. To accommodate the varying requirements for detection speed and accuracy in different scenes, YOLOv5 has released multiple versions. To balance detection speed and accuracy, we use the YOLOv5-m version for experimentation in this section. Moreover, YOLOv5 is one of the most widely used object detectors, and it also has three output feature layers. Similar to the previous section, we added attention modules to each output feature layer for optimization. First, we conducted experiments using YOLOv5 on the BDD100K dataset. Similar to the experiments described earlier, in this section, we utilized ablation experiments to validate each component of the proposed module. Pre-trained weights were applied during the experiments, and the results are shown in Table 5.

Table 5. The ablation experimental results of the proposed method in YOLOv5.

Method	HCCI	CSI	HCFI	mAP (%)
YOLOv5	×	×	×	45.39
	√	×	×	45.85
	×	√ (+CBAM channel)	×	45.81
	×	×	√	45.98

From Table 5, it can be observed that the proposed HCCI channel attention module and CSI spatial attention module effectively enhance the model's attention in the channel and spatial dimensions, respectively. Furthermore, the HCFI attention module achieves the best performance, providing strong evidence for the effectiveness of the proposed method. Building upon these findings, to more intuitively demonstrate the performance of the proposed method, we conducted a comparative analysis of the model's performance with the addition of different attention modules. The results are shown in Figure 9.

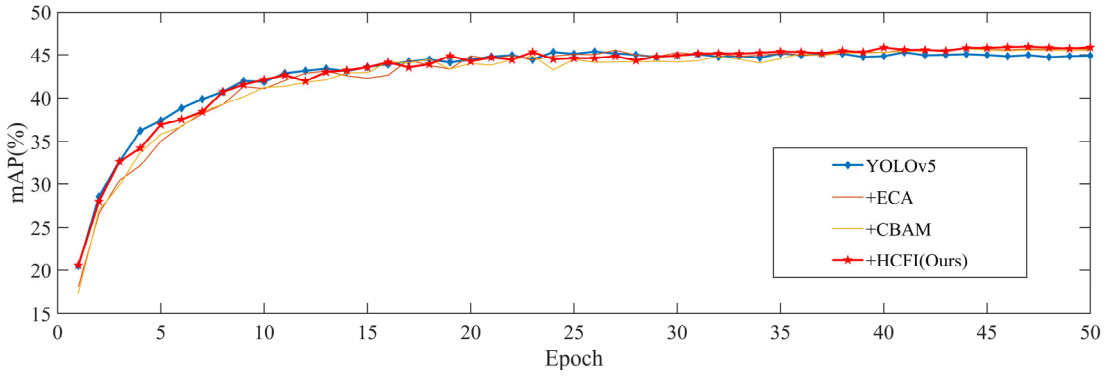

Figure 9. The performance of the model under different attention modules.

In Figure 9, the blue curve marked with the diamond represents the experimental results without adding any attention module, while the red curve marked with the asterisk represents the experimental results with the addition of the proposed HCFI attention module from this paper. The experimental results clearly illustrate that the proposed method outperforms ECA and CBAM, effectively improving the algorithm's focus on important objects. To further compare the overall performance enhancement effect of the proposed method, we evaluated the detection results under various metrics, including mAP50, mAP75, and mAP50:95. The experimental results are shown in Table 6.

Table 6. Performance evaluation of YOLOv5 on the BDD100K dataset.

Method/ Evaluation	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	FPS
YOLOX (base)	45.39	24.55	25.83	97.71
Relative improv. %				
base + ECA	45.73	25.26	25.91	95.07
Relative improv. %	0.75%	2.89%	0.31%	
base + CBAM	45.75	24.61	26.10	92.45
Relative improv. %	0.79%	0.24%	1.05%	
base + HCFI	45.98	25.26	26.36	89.73
Relative improv. %	1.30%	2.89%	2.05%	

As shown in Table 6, in the experiments on the BDD100K mobile scene dataset, the proposed HCFI attention module outperforms the currently widely used attention modules across different evaluation metrics, effectively improving the model's performance. Compared to the original YOLOv5, under the more comprehensive mAP50:95 metric, HCFI improves the accuracy to 26.36% mAP, resulting in a 2.05% enhancement in model detection performance. In addition, we conducted a computational speed analysis when adding various attention modules to the model. The results indicate that the proposed HCFI approach presented in this paper exhibits excellent computational efficiency, affirming that our method can enhance object detection focus while maintaining model computational efficiency.

Based on the experiments mentioned above, in order to thoroughly validate the effectiveness and robustness of our approach, we further conducted experiments using YOLOv5 on the COCO dataset. Following the experiments described earlier in this section, we initially assessed the performance of various methods on the COCO dataset across multiple metrics such as mAP50, mAP75, and mAP50:95. Furthermore, we further validated the performance of our approach using metrics like mAP_{small}, mAP_{medium}, and mAP_{large}. The experimental results are presented in Table 7.

Table 7. Performance evaluation of the YOLOv5 detector on the COCO dataset.

Method/Evaluation	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	mAP_{small} (%)	mAP_{medium} (%)	mAP_{large} (%)	FPS
YOLOX (base)	63.51	47.10	43.00	22.56	43.70	55.59	96.27
Relative improv. %							
base + ECA	63.66	48.33	44.42	24.07	45.43	56.89	93.81
Relative improv. %	0.24%	2.61%	3.30%	6.69%	3.96%	2.34%	
base + CBAM	63.64	48.33	44.44	23.26	46.26	56.95	88.28
Relative improv. %	0.20%	2.61%	3.35%	3.10%	5.86%	2.45%	
base + HCFI	64.14	48.49	44.91	23.69	46.68	57.39	86.37
Relative improv. %	0.99%	2.95%	4.44%	5.01%	6.82%	3.24%	

The results in Table 7 demonstrate that, on the COCO dataset, the proposed HCFI attention module consistently outperforms common attention modules across three distinct evaluation metrics (mAP50, mAP75, and mAP50:95). Compared to the original YOLOv5 model, it achieves a remarkable performance improvement of 4.44% in the comprehensive

mAP50:95 metric. Furthermore, in the performance testing for objects of different sizes, our approach still excels in two metrics, validating the effectiveness and robustness of our method. Finally, we conducted a computational speed analysis when incorporating different attention modules into the model. The results indicate that the proposed HCFI approach presented in this paper exhibits excellent computational efficiency. Compared to commonly used attention modules in the context of intelligent mobile scenes, it achieves competitive computational speed, providing strong evidence that our method can enhance focus on important objects while ensuring model computational efficiency.

4.4. Experiments on SSD

Following a similar methodology as the previous experiments, this section further substantiates the effectiveness of the HCFI attention module across different detectors. We conducted experiments using the SSD detector on the VOC2007 dataset, which includes 9963 images with 20 object categories. The SSD detector is a classic framework in the field of object detection, and in this paper's experiments, we used VGG16 as the backbone network for the detector.

Since the SSD detector is different from the two detectors mentioned earlier, it has a total of six output layers, with the last four output layers being very small, especially the final output layer having only one pixel size. In our research, we found that this configuration is not conducive to the application of spatial attention modules. Consequently, we applied the proposed HCFI attention module only to the first two larger feature layers, while applying the proposed HCCI channel attention module to the subsequent four smaller feature layers. The structure of SSD and the added attention modules are illustrated in Figure 10.

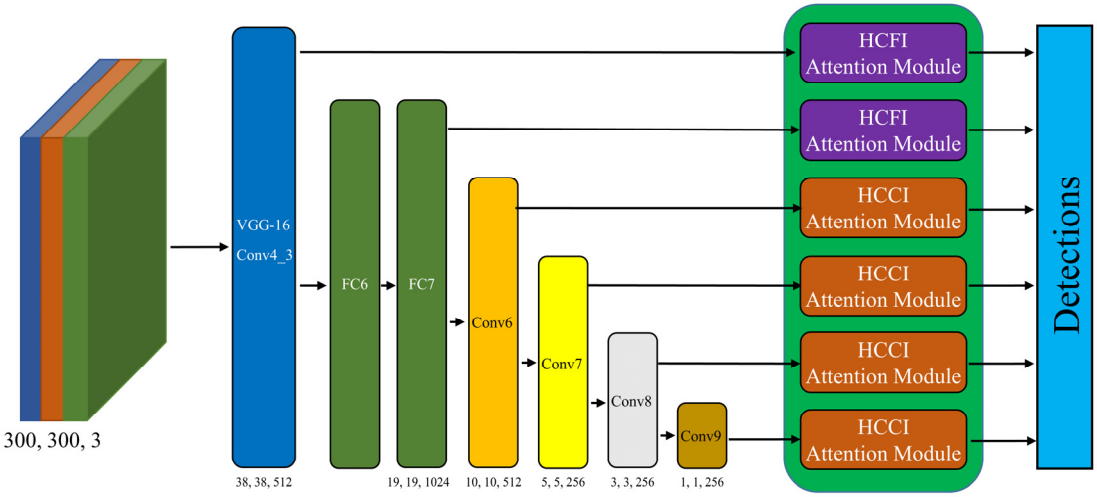

Figure 10. The structure of SSD, where the attention modules are highlighted with a green background.

Building on the insights gained from the earlier ablation experiments, in order to comprehensively compare the overall performance enhancement effect of the proposed method, we evaluated the detection results under various metrics, including mAP50, mAP75, and mAP50:95. The experimental results are shown in Table 8.

The experimental results presented in Table 8 demonstrate that the proposed attention mechanism outperforms the currently widely used attention modules in all three evaluation metrics, effectively improving the model's performance. Compared to the original SSD algorithm, the proposed method achieved a 0.74% improvement in model detection performance under the more comprehensive mAP50:95 metric, providing strong evidence for the effectiveness of the proposed method in enhancing the object detection attention capability. In the experiments, CBAM achieved lower results in both mAP75 and mAP50:95

metrics compared to the original detection results, once again highlighting the limitations of spatial attention modules in object detection algorithms. However, the proposed method obtained the best results, showcasing that the combination of HCFI and HCCI proposed in this paper is an excellent solution when dealing with smaller output feature layers.

Table 8. The performance evaluation results on the SSD detector.

Method/Evaluation	mAP50 (%)	mAP75 (%)	mAP50:95 (%)
SSD (base)	86.13	61.70	55.56
base + ECA	86.20	61.64	55.64
Relative improv. %	0.08%	−0.10%	0.14%
base + CBAM	86.20	61.38	55.31
Relative improv. %	0.08%	−0.52%	−0.45%
base + HCFI	86.67	62.54	55.97
Relative improv. %	0.63%	1.36%	0.74%

4.5. Discussion

In the preceding experimental section, we systematically assessed the efficacy of the proposed HCFI attention module across different detectors and datasets. The experiments with various detectors and datasets consistently showed that the proposed method achieved the best results, demonstrating the performance of the proposed approach. Building upon the objective analysis mentioned above, we further conducted visual experiments on the dataset to provide an intuitive discussion. Partial visual results are shown in Figure 11.

Figure 11. Partial visual experimental results.

As shown in Figure 11, each image pair features the original detection results on the left and the results obtained with the method proposed in this paper on the right. In the first group of visual experiment results, it can be observed that our method additionally pays attention to the traffic lights and the car on the left. In the second group of visual experiment results, our method also captures the car on the left, which was previously obscured in the ground truth. In the third group of visual experiment results, our method additionally identifies two people overlapping with the car instance in the lower right corner. In addition to the presented detection results, to further investigate the variations in the model's performance, we further utilize heatmap visualization on the mentioned images to demonstrate the differences in the model's detections. The heatmap visualization results are shown in Figure 12, where the leftmost image in each group is the original image, the middle image is the heatmap of the original detector, and the right image is the heatmap of our model in this paper.

Figure 12. Heatmap visualization results.

Analyzing the heatmap visualization results in Figure 12 alongside the detection results in Figure 11, it can be observed that the heatmap corresponding to the performance improvement of our proposed method is more distinct. For instance, in the first group of heatmaps, the region corresponding to the missed traffic light is much more prominent, indicating that our model has a stronger attention capability during the detection process. Similar to the aforementioned visual experiments, to further validate the effectiveness of our proposed method in real local mobile scenes, we collected real transportation scene data in Xi'an City for visualization verification, as shown in Figure 13.

Figure 13 showcases the visualization results of our proposed method applied to real local transportation scene data. In each set of images, the left side represents the original detection results, and the right side shows the detection results of the method proposed in this paper. In the first group of visual experiment results, it can be observed that our method additionally pays attention to the traffic signs in the middle of the images. In the second group of visual experiment results, it can be observed that our method not only captures the cars in the middle of the images but also avoids misidentifying three electric motorcycles as cars. In the third group of visual experiment results, our method additionally identifies the car on the left side and the pedestrians on the right side of the images. These diverse visualized detection results demonstrate that the proposed HCFI

attention module successfully pays attention to more objects, providing strong evidence for the effectiveness and wide applicability of our proposed method in this paper.

Figure 13. Visual experiment results in real local transportation scenes.

5. Conclusions

The objective of this paper is to enhance the applicability of object detection technology in intelligent mobile scenes by improving the model's attention to important objects. Diverging from recent developments like PCT and T2T-ViT, which propose deep self-attention networks, our primary approach is to minimize the additional computational burden on the model, making it suitable for mobile scenes. To achieve this, the paper conducted an in-depth study of the limitations of commonly used attention mechanisms when applying them to intelligent mobile scenes. Building upon this research, the paper first proposes the MKSPP structure and improves the channel attention to obtain the HCCI attention module with better cross-channel interaction performance. Furthermore, the paper improved spatial attention by incorporating dilated convolutions, resulting in the CSI attention module having better cross-space interaction performance. By sequentially combining these two modules, the paper achieves the improved HCFI attention module without using high-complexity operations. Finally, experimental results demonstrate that the proposed HCFI method increases the performance of YOLOX on the COCO dataset by 1.53% in terms of the comprehensive mAP50:95 metric and improves the performance of YOLOv5 on the BDD100K dataset by 2.05%. Additionally, the paper proposes a solution that combines HCCI and HCFI for detectors like SSD that have extremely small output feature layers. The experimental results validate the effectiveness of the proposed method in enhancing the model's attention to important objects. This research significantly advances the field and effectively enhances object detection's practicality in intelligent mobile scenes.

Author Contributions: Conceptualization, D.T.; methodology, D.T.; software, Y.H.; validation, D.T. and P.Z.; formal analysis, P.Z.; writing—original draft preparation, D.T.; writing—review and editing, Y.H., Y.L. and J.L.; supervision, M.L.; funding acquisition, Y.L. and P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China (grant number 2020YFB1713300), in part by the National Key Research and Development Program of China (grant number 2021YFB2601000), in part by the Natural Science Foundation of Shaanxi Province (grant number 2023-JC-QN-0664), and in part by the Key Research and Development Program of Shaanxi Province (grant number 2023JBGS-13).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
3. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
4. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2019**, *111*, 257–276. [CrossRef]
5. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2018**, *128*, 261–318. [CrossRef]
6. Zhao, Z.; Zheng, P.; Xu, S.; Wu, X. Object Detection with Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 3212–3232. [CrossRef]
7. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
8. Qin, L.; Shi, Y.; He, Y.; Zhang, J.; Zhang, X.; Li, Y.; Deng, T.; Yan, H. ID-YOLO: Real-Time Salient Object Detection Based on the Driver's Fixation Region. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15898–15908. [CrossRef]
9. Tian, D.; Han, Y.; Wang, B.; Guan, T.; Wei, W. A Review of Intelligent Driving Pedestrian Detection Based on Deep Learning. *Comput. Intell. Neurosci.* **2021**, *2021*, 5410049. [CrossRef]
10. Liang, S.; Wu, H.; Zhen, L.; Hua, Q.; Garg, S.; Kaddoum, G.; Hassan, M.; Yu, K. Edge YOLO: Real-Time Intelligent Object Detection System Based on Edge-Cloud Cooperation in Autonomous Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25345–25360. [CrossRef]
11. Wang, X.; Ban, Y.; Guo, H.; Hong, L. Deep Learning Model for Target Detection in Remote Sensing Images Fusing Multilevel Features. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July 2019–2 August 2019; pp. 250–253.
12. Han, X.; Zhong, Y.; Zhang, L. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
13. Guo, W.; Yang, W.; Zhang, H.; Hua, G. Geospatial Object Detection in High Resolution Satellite Images Based on Multi-Scale Convolutional Neural Network. *Remote Sens.* **2018**, *10*, 131. [CrossRef]
14. Yang, R.; Yu, Y. Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis. *Front. Oncol.* **2021**, *11*, 638182. [CrossRef]
15. Rezaei, M.; Yang, H.; Meinel, C. Instance Tumor Segmentation using Multitask Convolutional Neural Network. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
16. Ito, S.; Ando, K.; Kobayashi, K.; Nakashima, H.; Oda, M.; Machina, M.; Kanbara, S.; Inoue, T.; Yamaguchi, H.; Koshimizu, H.; et al. Automated Detection of Spinal Schwannomas Utilizing Deep Learning Based on Object Detection from MRI. *Spine* **2020**, *46*, 95–100. [CrossRef]
17. Sande, K.E.; Uijlings, J.R.; Gevers, T.; Smeulders, A. Segmentation as selective search for object recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1879–1886.
18. Jiang, X.; Pang, Y.; Pan, J.; Li, X. Flexible sliding windows with adaptive pixel strides. *Signal Process.* **2015**, *110*, 37–45. [CrossRef]
19. Guo, M.; Cai, J.; Liu, Z.; Mu, T.; Martin, R.; Hu, S. PCT: Point cloud transformer. *Comput. Vis. Media* **2020**, *7*, 187–199. [CrossRef]

20. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Tay, F.; Feng, J.; Yan, S. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 538–547.
21. Wu, H.; Xiao, B.; Codella, N.C.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
22. Wang, Q.; Wu, B.; Zhu, P.F.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
23. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 2011–2023. [CrossRef]
24. Woo, S.; Park, J.; Lee, J.; Koeon, I. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2018.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1904–1916. [CrossRef]
26. Viola, P.A.; Jones, M.J. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; p. 1.
27. Patle, A.; Chouhan, D.S. SVM kernel functions for classification. In Proceedings of the 2013 International Conference on Advances in Technology and Engineering (ICATE), Mumbai, India, 23–25 January 2013; pp. 1–9.
28. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
29. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
30. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef]
31. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
32. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
33. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
34. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:abs/1804.02767.
35. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:abs/2004.10934.
36. Jocher, G. YOLOv5. 2023. Available online: https://github.com/ultralytics/yolov5 (accessed on 5 June 2023).
37. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
38. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:abs/2107.08430.
39. Potapova, E.; Zillich, M.; Vincze, M. Survey of recent advances in 3D visual attention for robotics. *Int. J. Robot. Res.* **2017**, *36*, 1159–1176. [CrossRef]
40. Nguyen, T.V.; Zhao, Q.; Yan, S. Attentive Systems: A Survey. *Int. J. Comput. Vis.* **2018**, *126*, 86–110. [CrossRef]
41. Han, D.; Zhou, S.; Li, K.; Mello, R. Cross-modality Co-attention Networks for Visual Question Answering. *Soft Comput.* **2021**, *25*, 5411–5421. [CrossRef]
42. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems, Montreal, QC Canada, 7–12 December 2015.
43. Wang, X.; Girshick, R.B.; Gupta, A.K.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
44. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global Second-Order Pooling Convolutional Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3019–3028.
45. Fu, J.; Liu, J.; Tian, H.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.
46. Everingham, M.; Gool, L.V.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]

47. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.
48. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 2633–2642.