applied sciences

MDPI

Article

Improved Lightweight Multi-Target Recognition Model for Live Streaming Scenes

Zongwei Li [1], Kai Qiao [1], Jianing Chen [1], Zhenyu Li [2,*] and Yanhui Zhang [3]

[1] School of Economics and Management, Shanghai Institute of Technology, Shanghai 200235, China; lzw0118@163.com (Z.L.); qk19961014@gmail.com (K.Q.); chappyjn@163.com (J.C.)
[2] School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, China
[3] Business School, East China University of Science and Technology, Shanghai 200237, China; yanhui415@163.com
* Correspondence: zhenyu081@163.com

Featured Application: This study applies to object capture of traditional live streaming scene frames to help the live e-commerce field to obtain more data to provide more marketing strategies.

Abstract: Nowadays, the commercial potential of live e-commerce is being continuously explored, and machine vision algorithms are gradually attracting the attention of marketers and researchers. During live streaming, the visuals can be effectively captured by algorithms, thereby providing additional data support. This paper aims to consider the diversity of live streaming devices and proposes an extremely lightweight and high-precision model to meet different requirements in live streaming scenarios. Building upon yolov5s, we incorporate the MobileNetV3 module and the CA attention mechanism to optimize the model. Furthermore, we construct a multi-object dataset specific to live streaming scenarios, including anchor facial expressions and commodities. A series of experiments have demonstrated that our model realized a 0.4% improvement in accuracy compared to the original model, while reducing its weight to 10.52%.

Keywords: model optimization; object detection; attention mechanism; live streaming

Citation: Li, Z.; Qiao, K.; Chen, J.; Li, Z.; Zhang, Y. Improved Lightweight Multi-Target Recognition Model for Live Streaming Scenes. *Appl. Sci.* **2023**, *13*, 10170. https://doi.org/10.3390/app131810170

Academic Editor: Shiyang Yan

Received: 20 July 2023
Revised: 28 August 2023
Accepted: 6 September 2023
Published: 10 September 2023

1. Introduction

Live e-commerce has emerged as a prominent marketing trend, with its role as a powerful sales-boosting tool being widely embraced globally [1]. Since 2019, leading global retailers like Amazon and QVC have established their own live video shopping platforms. In particular, China has witnessed a significant surge in the user base of live e-commerce, reaching a staggering 469 million in 2022, indicating its immense commercial potential. The utilization of real-time marketing strategies in live streaming scenarios effectively conveys sensory cues to viewers, thereby stimulating consumer purchases [2]. Consequently, the ability to capture these sensory cues during live streaming has become increasingly crucial.

In live streaming, the primary focus of consumers' visual attention is centered around the anchor and the commodity, as these factors play a crucial role in influencing their purchasing decisions. To extract such visual cues effectively, object detection algorithms in machine vision have proven to be invaluable. Machine vision, a mainstream field within deep learning, encompasses various subfields including scene recognition, object recognition, object detection, and video tracking [3]. Among these subfields, object detection models based on deep learning have undergone significant advancements since the occurrence of Region-based Convolutional Neural Networks (R-CNN), resulting in notable improvements in both accuracy and speed [4]. Traditional object detection techniques can be partitioned into two groups: single-stage and two-stage object detection. The former, such as RCNN, Fast RCNN, etc., are lightweight and offer fast processing speeds. Conversely, the other techniques achieve higher accuracy but require significant computational

resources. The YOLO algorithm, widely employed in practical applications, serves as an excellent example of a single-stage object detection technique that achieves comparable accuracy to two-stage methods [5].

The introduction of attention mechanisms into machine vision has been a great success. The attentional mechanism in general is a dynamic selection process, adaptively weighting the input features, resulting in significant performance and accuracy improvements in object recognition, but with a relatively larger computation. Attention mechanisms, such as Shuffle Attention (SA), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA), have been developed to achieve lightweight enhancements and can be easily integrated into mobile network modules [6]. In recent years, researchers have been actively exploring lightweight modules such as GhostNet, MobileNetV3, and BlazeFace [7,8]. Additionally, many scholars have been attempting to refine the backbone section of YOLOv5 with lightweight modules and incorporate attention mechanisms, aiming to strike a balance between accuracy and computational efficiency.

Qi et al. [9] integrated the Squeeze and Excitation (SE) attention mechanism into YOLOv5 for tomato virus disease identification, achieving higher accuracy. However, this modification resulted in an increased inference time compared to the original model, and the attention mechanism consumed a significant amount of computational resources. Xu et al. [10] substituted the YOLOv5 backbone network with ShuffleNetV2 and integrated the CA attention mechanism, achieving a favorable balance between the indicators for mask detection. Li et al. [11] enhanced the backbone of the YOLOv5 model using GhostNet and incorporated the CA attention mechanism to detect anchor expressions in live streaming scenarios, yielding promising results. In live streaming scenes, relying solely on facial expressions is insufficient to capture the rich visual cues. To address this, we advise utilizing a new dataset that contains both anchor facial expressions and commodities. Furthermore, we enhance the YOLOv5 architecture by fusing it with MobileNetV3 and incorporating the CA attention mechanism, we replace the mobilenetv3 module in the backbone and neck layers, and continuously adjust the corresponding parameters to achieve the best results, meanwhile adding the CA attention mechanism to further improve the accuracy of the model. Our experiments demonstrate that the fusion of MobileNetV3 and CA attention mechanisms leads to improved performance in YOLOv5 models. The major contributions of our research are the following:

1. Our improved MobileNetV3-CA network architecture overcomes the limitations of YOLOv5, providing more possibilities for lightweight models and its weight has been greatly reduced, accompanied by an increase in precision.

2. We evaluate multiple model variations on a self-built dataset and find that our improved model achieves the best balance of metrics with extreme lightness.

3. The combined recognition of anchor expressions and commodity categories offers more efficient technical support for enhancing marketing strategies in the live e-commerce industry.

2. Related Work

2.1. Deep Learning and Emotion Recognition in Live Streaming Scenarios

Emotions, as fundamental human behaviors, play a significant role in information processing and can trigger corresponding actions [12]. The impact of emotions on human behavior has been commonly demonstrated among various domains such as online comments, advertising marketing, TV shopping, and live commerce [13–15]. The generation of emotions in live streaming scenarios is complex, with sensory cues being important factors in emotional arousal. As a result, sensory marketing has gained increasing attention [16]. Some researchers have focused on manipulating emotions through sensory cues, such as smell and music [17,18]. The influence of rich sensory stimuli on consumer emotions can lead to impulsive buying, and it has been confirmed that impulsive buying behavior is primarily driven by emotions [2]. Therefore, it remains an important topic to explore how to evoke consumer emotions through sensory cues to promote sales in live streaming environments.

Emotion recognition has been provided favorable technical support by the development of deep learning. The emergence of Convolutional Neural Networks (CNNs) has made significant strides in object detection models that recognize emotions from facial expressions [19]. Real-time or near real-time speech emotion recognition algorithms have also greatly improved with the development of deep learning, moving away from old-fashioned frameworks such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) [20]. In live streaming scenarios, viewers are often attracted by the anchors in the broadcasting room, and visual cues, as the most intuitive influencing factor, should be carefully considered. According to the theory of emotional contagion [21], the emotions of the anchor will undoubtedly exert a significant influence on the emotions of the audience to a certain extent.

2.2. Application of Yolov5 Algorithm for Object Detection

Traditionally, feature extraction in object detection heavily relied on manual feature design, which often resulted in poor generalization. However, with the emergence of deep learning, Convolutional Neural Networks (CNNs) have emerged as the mainstream framework for machine learning in object detection, thanks to their remarkable performance and excellent feature extraction capabilities, starting from the introduction of R-CNN. One-stage and two-stage algorithms are the two main types of deep learning-based object detection algorithms. Firstly, one-stage algorithms directly predict the object's coordinates and class through regression, offering faster recognition speeds. On the other hand, two-stage algorithms employ region generation for target classification and calibration, leading to higher accuracy. However, the two-stage approach comes with increased computational overhead, reducing the model's speed and hindering real-time monitoring [22,23]. Since 2015, the YOLO (You Only Look Once) family of single-stage deep learning algorithms has undergone continuous improvements. YOLO utilizes a convolutional neural network architecture to determine the location and type of objects in an image, enabling high-speed recognition. The yolov5 deep learning algorithm further enhances efficiency by adopting a more lightweight network architecture, significantly reducing the weight and improving the speed. The yolov5 family comprises four different architectures (YOLOv5x, YOLOv5l, YOLOv5m, and YOLOv5s), allowing flexibility in adapting to various object detection requirements by adjusting the extracted features' width and depth [24].

YOLOv5s, the lightest variant in the YOLOv5 series, boasts the fastest recognition speed and finds widespread application in various scenarios. Wang et al. [25] utilized a YOLOv5s model with channel pruning to achieve remarkable results in fast apple fruit detection. Guo et al. [26] optimized the backbone network of the YOLOv5s and integrated the SE attention mechanism, significantly improving the model's accuracy compared to YOLOv5s and YOLOv4. Li et al. [27] employed YOLOv5s in an industrial setting for forklift monitoring, enhancing the backbone section with the GhostNet and incorporating the SE attention mechanism. Li et al. [11] pioneered the application of YOLOv5 in a live streaming scenario for real-time monitoring of anchor expressions. The improved YOLOv5s model incorporates the GhostNet module and the CA attention mechanism, achieving a superior balance between precision and speed.

The previous YOLOv5 model has found extensive applications in various commodity environments. However, as the live streaming scene is still a nascent industry, there is significant potential to explore more applications for YOLOv5 in this domain. While Li et al. [11] achieved effective recognition of anchor expressions through an improved model, our focus extends beyond expressions to encompass other elements within the live streaming scene. Therefore, the re-application of the model for further improvements becomes particularly crucial.

2.3. The Development and Application of Attention Mechanism in Deep Learning

Inspired by human perception, the attention mechanism is implemented. When humans visually perceive objects, they tend to focus on specific parts that are relevant

or important to them. This selective observation allows humans to efficiently extract important information from a substantial quantity of visual data using limited cognitive resources. The attention mechanism mimics this process, enhancing the efficiency and accuracy of perceptual information processing. It serves as an effective solution to tackle the challenge of information overload. By incorporating the attention mechanism into computer vision tasks, the substantial computational workload can be effectively reduced. As a result, the attention mechanism has gained significant traction in the realm of deep learning, becoming a standard component in neural network architectures [28].

Currently, the two most common attention mechanisms applied to machine vision are spatial attention and channel attention [6]. The emphasis on the former is on the location of the object within the deep learning information and spatially transforms this location information. The spatial transformer network (STN) [29] is an example of spatial attention. Additionally, channel attention emphasizes the content information of the object. The SE network, introduced by Hu et al. [30], is a notable channel attention mechanism. The SE attention module enhances target recognition by adaptively calibrating channel weights, filtering important features, and using global average pooled features for computations.

As deep learning neural networks continue to evolve, researchers have developed hybrid attention mechanisms that combine both spatial and channel attention to improve the precision and efficiency of feature recognition within large feature maps [31]. The CBAM is capable of feature map recognition through both spatial and channel attention. It starts by applying global pooling operations to the feature map, generating channel attention features. Subsequently, spatial attention features are generated by concatenating and downsampling the channels. Finally, the input features are combined with the final features [32].

The CA mechanism integrates spatial coordinate information by embedding location details into channel attention, decomposing channel attention into two parallel one-dimensional feature encodings. This approach differs from CBAM as it does not forcibly compress the channels. The two one-dimensional feature encodings allow for more comprehensive extraction of spatial information and optimize feature extraction efficiency [33].

Another efficient replacement attention mechanism is SA. SA combines channel attention and spatial attention using shuffling units. This lightweight and efficient attention mechanism has demonstrated better performance and lower complexity compared to CBAM and SE attention mechanisms on public datasets [34].

3. Method

3.1. Data Pre-Processing

To enhance machine vision applications in live streaming scenarios, we have constructed expression–commodity datasets. In this process, we referred to well-known datasets including CK+ and MMI, which contain video sequences capturing facial expressions and are suitable for facial expression recognition and detection in videos [35,36]. While Li et al. [11] developed their own data pool to address the lack of anchor facial expressions data in live e-commerce and we observed that professional anchors tend to convey positive emotions to enhance the ambiance of the broadcasting room and boost consumers' purchase intention [37]. Based on this, we categorized anchor expressions into two main categories: emotional and normal.

Regarding the selection of goods, given the diverse range of commodities available in live streaming rooms, it is challenging to establish a direct connection between goods and anchor expressions across multiple categories. Therefore, we divided the goods into two categories: utilitarian and hedonic. Research on the hedonic and utilitarian categories of goods primarily focuses on scale development, as the same commodity can exhibit variations in both dimensions based on marketing strategies and subjective consumer factors. To ensure the dataset's generalizability, we employed the classification criteria proposed by Voss et al. [38] and classified the goods within our dataset accordingly. We selected representative goods to include in the final dataset to enhance its applicability. Table 1 illustrates the composition of our commodity data.

Table 1. Examples of utilitarian and hedonic commodities.

Utilitarian	Hedonic
Milk, bread, cookies, raw meat, socks, underwear...	High-grade liquor, luxury goods, cosmetics, high-grade leather boots...

3.2. Improved Yolov5s Model

To achieve our goal of an extremely lightweight model, we chose yolov5s, the weightless version, as the base network for improvement, on which we made a series of improvements to achieve our desired results.

In our study, the improved model framework consists of four main components. First, the feature map is entered in the input section, Then, the Backbone component optimizes different input image features to obtain a large amount of semantic and location information. Secondly, in the Neck structure, FPN and PAN are included, and FPN can fuse the features extracted from the Backbone to enhance the semantic features. And the feature pyramid structure of PAN enhances the ability of the model to convey precise location features, which helps the model to be able to perform target detection at different scales. Finally, the Prediction part is able to map the corresponding information to the corresponding images.

In summary, by leveraging the lightweight YOLOv5s model as a foundation and implementing enhancements, including the replacement of the Backbone and Neck layers with MobileNetV3 modules, as well as the integration of the CA mechanism, we have achieved the successful development of an exceptionally lightweight model. Remarkably, this model retains a high level of performance in object detection. For a visual representation of the network architecture, please refer to Figure 1.

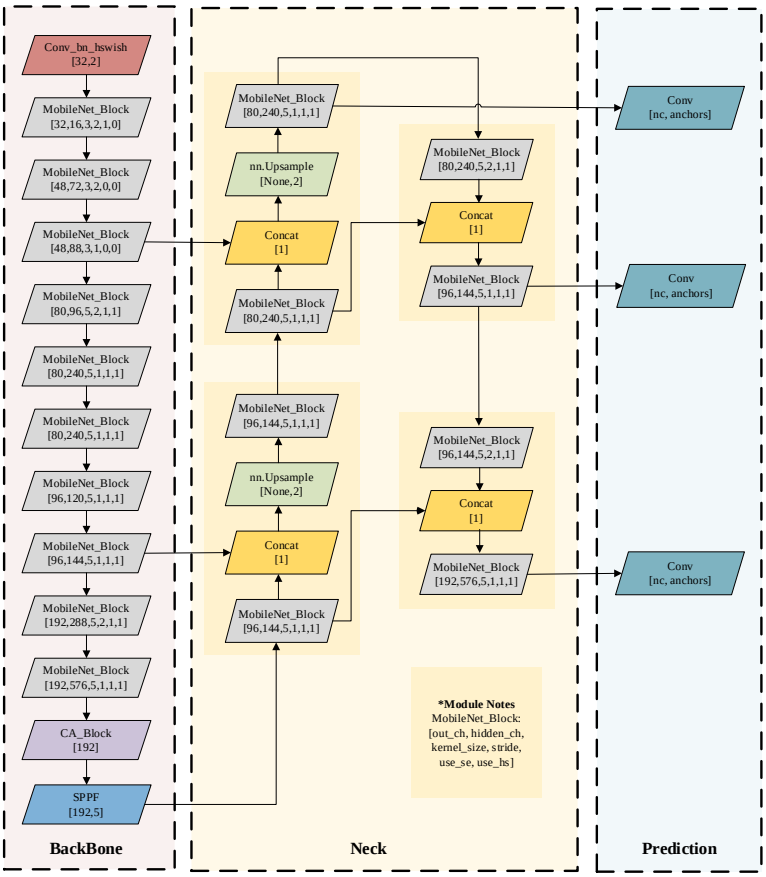


Figure 1. Improved yolov5s-MobileNetV3-CA network architecture (* MobileNet_Block: [out_ch, hidden_ch, kernel_size, stride, use_se, use_hs]).

3.2.1. MobileNetV3 Modules

MobileNet model is a lightweight CNN introduced by Google in 2017 [39]. It revolutionizes the computation of convolutional layers by introducing Depthwise Separable Convolution. The technique significantly reduces the model's number parameters without compromising accuracy.

Building upon the successes of MobileNetV1 and MobileNetV2, MobileNetV3 incorporates depth-separable convolutions and a residual structure with linear bottlenecks. Additionally, it introduces the SE channel attention mechanism, which assigns higher weights to important feature channels. To optimize network layers, MobileNetV3 combines the NAS (Network Architecture Search) algorithm [40] and the NetAdapt algorithm [41]. NAS optimizes individual network layers to explore the global structure, while NetAdapt locally optimizes network layers to determine the optimal number of convolutions.

In terms of network structure, MobileNetV3 brings significant improvements to the tail structure. It removes layers before average pooling and employs 1×1 convolutions to compute feature maps. This modification reduces computational effort and latency while preserving high-level features. Furthermore, MobileNetV3 replaces the original swish activation function with a new hard-swish function, which enhances the quantization process and speeds up model inference:

$$Hard - swish[x] = x\frac{ReLU6(X+3)}{6} \tag{1}$$

The SE attention module in MobileNetV3 performs channel-wise pooling of the output feature matrix. To generate output vectors, the pooled values are passed through two fully connected layers. The former has a number of vectors that is one quarter of the quantity of channels, and it applies the ReLU activation function. As for another, it employs the h-swish activation function and produces an output with the same amount of channels. The structure of MobileNetV3, as illustrated in Figure 2, incorporates the aforementioned SE attention module.

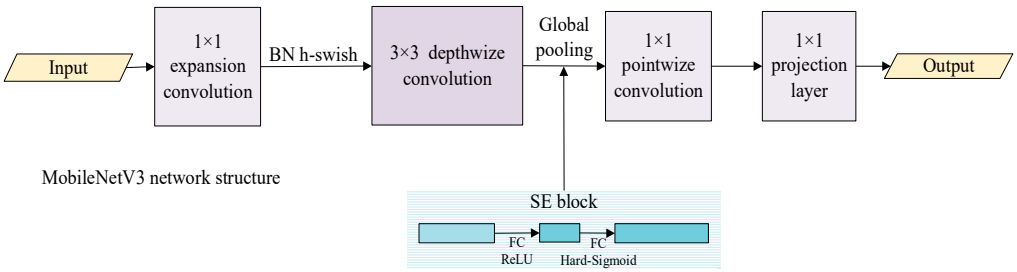


Figure 2. MobileNetV3 network structure.

3.2.2. CA Attentional Mechanisms

In order to maintain high accuracy even after extreme compression, we have opted to incorporate the CA mechanism, which encompasses both spatial and channel attention. This mechanism significantly aids in precise object localization and identification. It leverages 1D global pooling to gather directional feature maps in both horizontal and vertical directions. This enhances the representation of learned features within mobile networks by capturing location coordinates. The CA attention mechanism has garnered considerable attention in the realm of mobile networks. Its flexibility and lightweight nature mean that it can be easily integrated into classic building blocks of mobile networks, such as MobileNetV3 modules.

The encoding process of the CA attention mechanism consists of two main processes: information embedding and attention generation. Firstly, in the process of coordinate information embedding, the global pool is decomposed to ensure the long-range interaction capture, and the channels are encoded along two horizontal vertical coordinates from two spatial ranges of (*H*,1) and (1,*W*), respectively, and the output between two directions is given by:

$$z_m^h(h) = \frac{1}{W}\sum_{0\le i\le W} x_m(h,i) \tag{2}$$

$$z_m^w(h) = \frac{1}{H}\sum_{0\le j\le H} x_m(j,w) \tag{3}$$

Afterwards, a one-dimensional direction-aware feature map is generated. This transformation captures the long-range dependencies in one spatial direction, while preserving the precise position information in another. In the process of coordinate attention generation, the output results of both directions are first sent to a shared 1×1 convolutional transform function:

$$f = RELU\left(F_{conv1\times1}\left[z^h, z^w\right]\right) \tag{4}$$

After that, f is divided into two independent tensors along the spatial dimension, and a tensor of the same channel count as the input X is obtained by a 1×1 convolutional transformation to obtain:

$$g^h = \sigma\left(F_{conv1\times1}\left(f^h\right)\right) \tag{5}$$

$$g^w = \sigma(F_{conv1\times1}(f^w)) \tag{6}$$

Finally, the output expansions are used as attention weights to obtain the final weighted attentional feature map:

$$y_m(i,j) = x_m(i,j) \times g_m^h(i) \times g_m^w(j) \tag{7}$$

The CA attention mechanism avoids the compression of the tensor channel bits compared to CBAM and retains more features, which will enable the model to achieve further improvements in accuracy.

4. Experience

4.1. Data Set and Experimental Environment

In our research, we utilized a self-assembled dataset to evaluate and train the enhanced model. The dataset comprises a total of 1844 images, encompassing four distinct target types: utilitarian, emotional, normal, and hedonic.

For model deployment and experimentation, we used the lab hardware system shown in Table 2, which includes an NVIDIA GeForce RTX 3090 GPU, an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, the PyTorch deep learning framework, and CUDA 11.6 hardware acceleration.

Table 2. The lab hardware system of the training environment.

Item	Item Value
Operation system	Windows 10
CPU	Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz
GPU	NVIDIA GeForce RTX 3090
Hardware acceleration	CUDA11.6

4.2. Experimental Results

To evaluate the capabilities of our model, we conducted a sequence of experiments using a self-assembled data set on the hardware devices mentioned above. In this paper, we adopted the commonly used metrics of precision and recall, which are represented by

mAP@0.5 and mAP@0.5:0.95 [42]. They serve as indicators of the precision of the model. Furthermore, to describe the lightweight degree of our model, we applied parameters such as weights, GFLOPs, and the number of parameters, which provide strong evidence on a range of model performances. Additionally, we introduced another performance measure called precision density [43]. It is defined as the precision divided by the parameters. It can distinctly illustrate the equilibrium between model weights and precision. This criterion was utilized to measure the overall performance of our model. The results of our experiments are presented in the Table 3:

Table 3. Experimental results of the yolov5s and the Yolov5s-MobileNetV3-CA model.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Weights (MB)	GFLOPs	Parameters (M)	Accuracy Density	Time (ms)
Yolov5s	0.976	0.716	14.4	15.8	7020913	0.139	14.9
Yolov5s-MobileNetV3-CA	0.98	0.713	1.9	2.1	738619	1.32	22.2

From the table above, we can observe that our improved model achieved a 0.4% increase in accuracy compared to the yolov5s's mAP@0.5, and the accuracy density improved significantly by 849.64%. Furthermore, the size of our model was reduced to only 10.52% of the original model's size. It is important to note that our model sacrifices detection speed to some extent, but for the application scenarios we designed, the trade-off between extreme lightweight design and higher accuracy is well justified. During the training process, we recorded variations in the mAP for each epoch in Figure 3 and the accuracy of the improved model stabilizes with the original model after about 90 epochs; although the improved model does not converge as effectively as the original model, the final outcomes consistently exhibit stability and demonstrate slight improvements over the original model. Figure 4 provides a visual representation that effectively contrasts the values of mAP0.5 and accuracy density for the two models, serving as a clear manifestation of the heightened performance achieved by the improved model. Furthermore, Figure 5 presents a detailed depiction of our model's losses, effectively highlighting the exemplary convergence that our model has achieved.

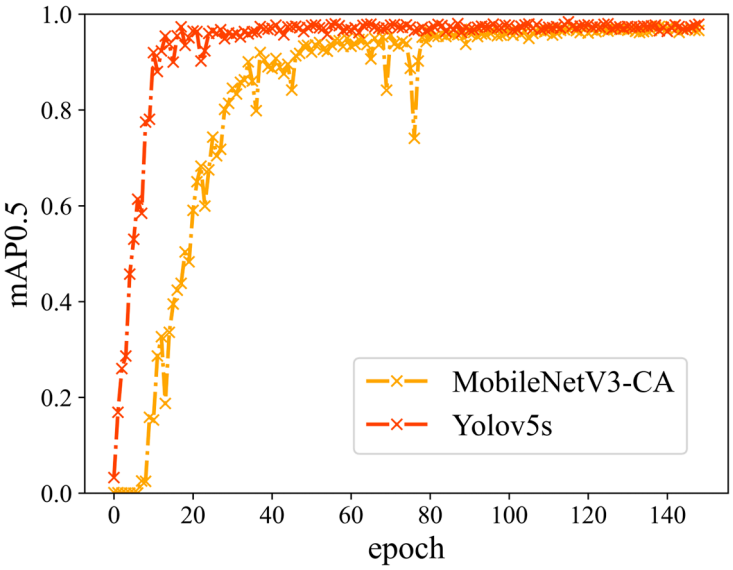


Figure 3. The mAP0.5 histories of the yolov5s and the Yolov5s-MobileNetV3-CA model.

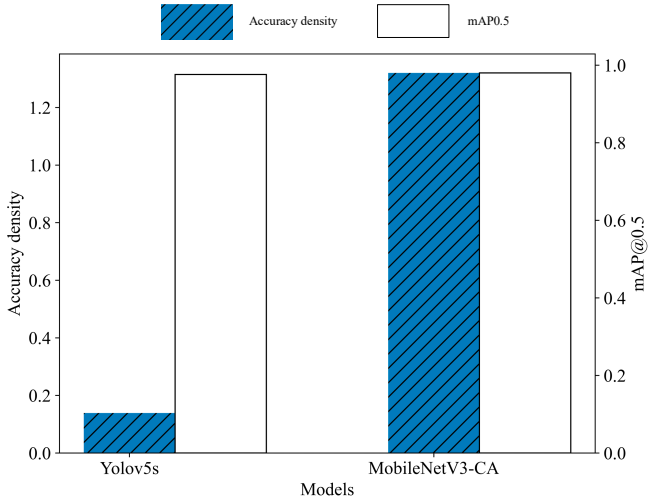


Figure 4. Comparison of accuracy density and mAP0.5 parameter values for the two models.

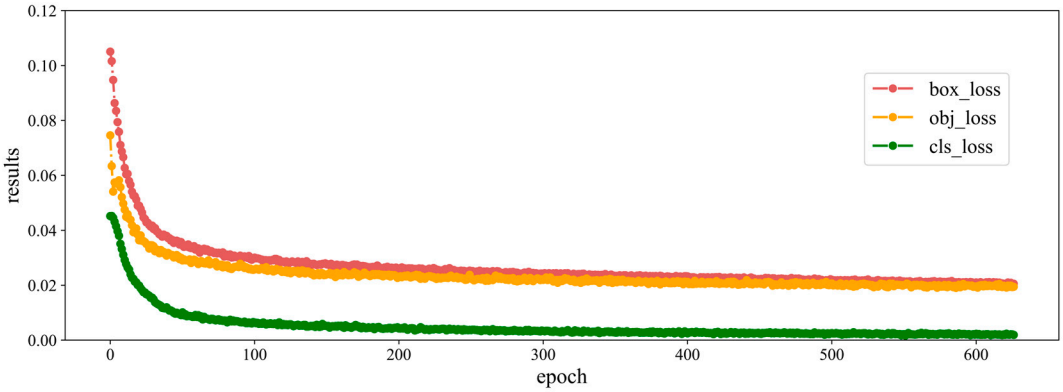


Figure 5. The losses of the Yolov5s-MobileNetV3-CA model.

The experimental results demonstrate the significant weight reduction in our model compared to the original model with a small amount of accuracy improvement, which is of great help in the application of live streaming scenarios. In the live streaming scenario, the anchor's facial expression has continuity, and neither the product nor the expression changes instantaneously, so it is acceptable for our model to sacrifice a certain amount of speed, as using less resources to obtain accurate live streaming information is the main purpose of our model improvement.

4.2.1. Ablation Experiments

To provide a clearer illustration of the role played by different modules in the Yolov5s-MobileNetV3-CA model, we performed a sequence of ablation experiments to evaluate individual modules. The evaluated models include the YOLOv5s model, the replaced model with the MobileNetV3, the YOLOv5s model with CA, and the improved model with both the MobileNetV3 module and the CA attention mechanism.

From Table 4, we can observe that the incorporation of the MobileNetV3 led to a significant decline in parameters and GFLOPs, and the model weights reduced from 14.4 MB to 1.8 MB, and the number of parameters reduced to 10.11% of the original size. The accuracy of both the YOLOv5s model and the improved MobileNetV3 model is improved by adding the CA attention mechanism, while the size of the model remains almost unchanged. This highlights the effectiveness of the CA in enhancing model performance.

Table 4. Ablation experiment results of each module.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Weights (MB)	GFLOPs	Parameters (M)	Accuracy Density
Yolov5s	0.976	0.716	14.4	15.8	7.020913	0.139
Yolov5s-CA	0.98	0.716	14.4	15.8	7.045521	0.139
Yolov5s-MobileNetV3	0.978	0.718	1.8	1.9	0.709883	1.37
Yolov5s-MobileNetV3-CA	0.98	0.713	1.9	2.1	0.738619	1.32

4.2.2. Comparison Experiments

To prove the efficacy of the CA attention mechanism further, we incorporated additional attention mechanisms, such as CBAM and SA, into our network structure and conducted a series of comparative experiments.

The results of the attention mechanism comparison experiments are displayed in Table 5. It revealed that the addition of the CA had a substantial positive impact on the model's accuracy improvement. Conversely, the inclusion of CBAM and SA attention mechanisms resulted in a significant decrease in model accuracy. Specifically, the model with the CA attention mechanism achieved a 0.82% and 0.92% higher mAP@0.5 compared to the models with CBAM and SA attention mechanisms, respectively. Moreover, the model with the CA attention mechanism exhibited faster detection speed.

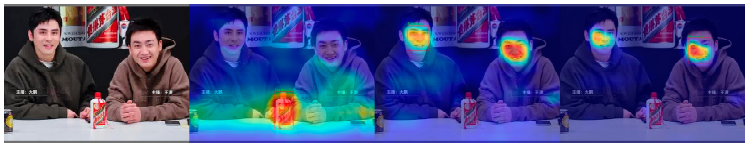
These findings strongly illustrate the effectiveness of our utilization of the CA attention mechanism and emphasize its superiority over alternative attention mechanisms, such as CBAM and SA, regarding accuracy enhancement and detection speed.

Table 5. Comparison experiment results of each module.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Weights (MB)	GFLOPs	Parameters (M)	Accuracy Density	Time (ms)
Yolov5s-MobileNetV3-SA	0.971	0.711	1.9	2.1	0.738619	1.314	22.4
Yolov5s-MobileNetV3-CBAM	0.972	0.708	1.9	2.1	0.737549	1.317	22.7
Yolov5s-MobileNetV3-CA	0.98	0.713	1.9	2.1	0.738619	1.326	22.2

In addition, we visualized the training results of the models with the three different attention mechanisms using heatmaps, as shown in Figure 6. This visualization technique enables us to validate if the models are overfitted and provides insights into whether the predictions are primarily driven by image features or influenced by the background.

MobileNetv3-CA



MobileNetv3-SA

MobileNetv3-CBAM

Figure 6. Heat map visualization results of the learning effects of different attention mechanisms.

Through the heatmaps, we can observe that the model is effective in recognizing all three objects in the picture, but is different. We observed that the allocation of attention in the SA was relatively scattered, and the recognition of the target objects was not sufficiently focused. Moreover, compared to CBAM, the CA was able to effectively concentrate its attention on the recognized target objects without causing excessive dispersion.

This analysis indicates that the CA performs better in terms of focusing on the identified target objects and avoiding unnecessary dispersion, as compared to SA and CBAM.

5. Conclusions and Future Research

In our research, we aimed to construct an object detection model that can achieve a remarkable balance between precision and lightweight design and applied it to live streaming scenarios. We have enhanced the yolov5s framework by incorporating the MobileNetV3 structure for the optimization of the Backbone and Neck layers. Additionally, we have introduced the CA attention mechanism to create an extremely lightweight model. Through a series of experiments, we have demonstrated that our model achieves significant improvements in terms of parameter reduction and weight while further enhancing accuracy. As a result, our model offers greater flexibility for deployment in various devices and occupies minimal space in this application scenario. These advancements provide robust technical support for a wide range of marketing strategies in live marketing. The selection of MobileNetV3 for this study, in comparison to our previous model enhancements [11], enabled us to achieve advancements in both the Backbone and Neck layers. This choice also provided us with a broader scope for parameter experimentation, and the results have proven this approach to be highly beneficial. The former model significantly contributed to balancing various parameters including accuracy, weight, and speed. Building upon previous work, our refined model further emphasizes the importance of reducing model weight and parameter count, making it exceptionally fitting for our application's live streaming scenario.

In future research, we aim to further optimize our model to address its detection speed limitations. Additionally, we will explore additional sensory cues, including but not limited to visual and auditory inputs, in live marketing strategies. The integration of multimodality is a research direction that deserves attention. Our goal is to build a more comprehensive object detection network specifically designed for live streaming scenarios. This will provide marketers and researchers in the field of live streaming with reliable technical support that meets diverse requirements and needs.

Author Contributions: Conceptualization, Z.L. (Zhenyu Li) and Z.L. (Zongwei Li); methodology, Z.L. (Zhenyu Li); software, Y.Z.; validation, J.C., K.Q. and Z.L. (Zongwei Li); formal analysis, Z.L. (Zhenyu Li); investigation, K.Q.; resources, Y.Z.; data curation, J.C.; writing—original draft preparation, K.Q.; writing—review and editing, K.Q.; visualization, K.Q.; supervision, Z.L. (Zhenyu Li); project administration, Z.L. (Zongwei Li). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, Study on the mechanism and spatial and temporal effects of international learning on the internationalization speed of manufacturing enterprises, grant number 71974130.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All data are publicly available.

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

References

1. Zheng, S.; Chen, J.; Liao, J.; Hu, H.L. What motivates users' viewing and purchasing behavior motivations in live streaming: A stream-streamer-viewer perspective. *J. Retail. Consum. Serv.* **2023**, *72*, 103240. [CrossRef]
2. Zhang, X.; Cheng, X.; Huang, X. "Oh, My God, Buy It!" Investigating impulse buying behavior in live streaming commerce. *Int. J. Hum. Comput. Interact.* **2022**, *39*, 2436–2449. [CrossRef]
3. Morris, T. *Computer Vision and Image Processing*; Palgrave Macmillan Ltd.: London, UK, 2004; pp. 1–320.
4. Aziz, L.; Salam, M.S.B.H.; Sheikh, U.U.; Ayub, S. Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review. *IEEE Access* **2020**, *8*, 170461–170495. [CrossRef]
5. Diwan, T.; Anirudh, G.; Tembhurne, J.V. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimed. Tools Appl.* **2023**, *82*, 9243–9275. [CrossRef] [PubMed]
6. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
7. Bazarevsky, V.; Kartynnik, Y.; Vakunov, A.; Raveendran, K.; Grundmann, M. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv* **2019**, arXiv:1907.05047.
8. Jin, R.; Xu, Y.; Xue, W.; Li, B.; Yang, Y.; Chen, W. An Improved Mobilenetv3-Yolov5 Infrared Target Detection Algorithm Based on Attention Distillation. In *International Conference on Advanced Hybrid Information Processing*; Springer International Publishing: Cham, Switzerland, 2021; pp. 266–279.
9. Qi, J.; Liu, X.; Liu, K.; Xu, F.; Guo, H.; Tian, X.; Li, M.; Bao, Z.; Li, Y. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease. *Comput. Electron. Agric.* **2022**, *194*, 106780. [CrossRef]
10. Xu, S.; Guo, Z.; Liu, Y.; Fan, J.; Liu, X. An improved lightweight yolov5 model based on attention mechanism for face mask detection. In *Artificial Neural Networks and Machine Learning–ICANN 2022, Proceedings of the 31st International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2022, Part III*; Springer Nature: Cham, Switzerland, 2022; pp. 531–543.
11. Li, Z.; Song, J.; Qiao, K.; Li, C.; Zhang, Y.; Li, Z. Research on efficient feature extraction: Improving YOLOv5 backbone for facial expression detection in live streaming scenes. *Front. Comput. Neurosci.* **2022**, *16*, 980063. [CrossRef]
12. Clore, G.L.; Schwarz, N.; Conway, M. Affective causes and consequences of social information processing. *Handb. Soc. Cogn.* **1994**, *1*, 323–417.
13. Deng, B.; Chau, M. The effect of the expressed anger and sadness on online news believability. *J. Manag. Inf. Syst.* **2021**, *38*, 959–988. [CrossRef]
14. Bharadwaj, N.; Ballings, M.; Naik, P.A.; Moore, M.; Arat, M.M. A new livestream retail analytics framework to assess the sales impact of emotional displays. *J. Mark.* **2022**, *86*, 27–47. [CrossRef]
15. Lin, Y.; Yao, D.; Chen, X. Happiness begets money: Emotion and engagement in live streaming. *J. Mark. Res.* **2021**, *58*, 417–438. [CrossRef]
16. Krishna, A. An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *J. Consum. Psychol.* **2012**, *22*, 332–351. [CrossRef]
17. Gardner, M.P. Mood states and consumer behavior: A critical review. *J. Consum. Res.* **1985**, *12*, 281–300. [CrossRef]
18. Kahn, B.E.; Isen, A.M. The influence of positive affect on variety seeking among safe, enjoyable products. *J. Consum. Res.* **1993**, *20*, 257–270. [CrossRef]
19. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 443–449.
20. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* **2021**, *21*, 1249. [CrossRef]
21. Barsade, S.G. The ripple effect: Emotional contagion and its influence on group behavior. *Adm. Sci. Q.* **2002**, *47*, 644–675. [CrossRef]
22. Girshick, R.; 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; IEEE: Columbus, OH, USA, 2014; pp. 580–587. [CrossRef]
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE T. Pattern Anal.* **2017**, *39*, 1137–1149. [CrossRef]
24. Glenn, J. yolov5. Git Code. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 4 March 2023).
25. Wang, D.; He, D. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosyst. Eng.* **2021**, *210*, 271–281. [CrossRef]
26. Guo, K.; He, C.; Yang, M.; Wang, S. A pavement distresses identification method optimized for YOLOv5s. *Sci. Rep.* **2022**, *12*, 3542. [CrossRef]
27. Li, Z.; Lu, K.; Zhang, Y.; Li, Z.; Liu, J.B. Research on Energy Efficiency Management of Forklift Based on Improved YOLOv5 Algorithm. *J. Math.* **2021**, *2021*, 5808221. [CrossRef]
28. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]

29. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 2, pp. 2017–2025.
30. 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, 18–23 June 2018; pp. 7132–7141.
31. Hu, H.; Li, Q.; Zhao, Y.; Zhang, Y. Parallel deep learning algorithms with hybrid attention mechanism for image segmentation of lung tumors. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2880–2889. [CrossRef]
32. 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.
33. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
34. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2235–2239.
35. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6–8 July 2005; IEEE: Piscataway, NJ, USA, 2005; p. 5.
36. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn–Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101. [CrossRef]
37. Guo, J.; Wang, X.; Wu, Y. Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *J. Retail. Consum. Serv.* **2020**, *52*, 101891. [CrossRef]
38. Voss, K.E.; Spangenberg, E.R.; Grohmann, B. Measuring the hedonic and utilitarian dimensions of consumer attitude. *J. Mark. Res.* **2003**, *40*, 310–320. [CrossRef]
39. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
40. 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, 10 October–2 November 2019; pp. 1314–1324.
41. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. Mnasnet: Platform-aware neural architecture search for mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2820–2828.
42. Borisyuk, F.; Gordo, A.; Sivakumar, V. Rosetta: Large scale system for text detection and recognition in images. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 71–79.
43. Bianco, S.; Cadene, R.; Celona, L.; Napoletano, P. Benchmark analysis of representative deep neural network architectures. *IEEE Access* **2018**, *6*, 64270–64277. [CrossRef]