Journal of Organizational and End User Computing
Volume 36 • Issue 1 • January-December 2024

Drug Recognition Detection Based on Deep Learning and Improved YOLOv8

Dingju Zhu
https://orcid.org/0000-0002-5907-3349
South China Normal University, China

Zixuan Huang
South China Normal University, China

KaiLeung Yung
Hong Kong Polytechnic University, China

Andrew W. H. Ip
https://orcid.org/0000-0001-6609-0713
University of Saskatchewan, Canada

ABSTRACT

Identifying drugs from surveillance or other videos presents challenges such as small target sizes, class imbalance, and similarities to other objects. Additionally, the hardware used to capture videos and the video resolution and clarity limit model scalability, leading to poor detection accuracy in traditional models. To address this issue, we propose an improved YOLOv8s-based model. The experimental outcomes reveal that the improved YOLOv8s model attains a precision of 95.1% and a mAP@50 of 87.4% in drug detection and identification, representing improvements of 3.0% and 2.2% over the original YOLOv8s model. The proposed improvements to YOLOv8s effectively boost detection accuracy and recognition rates while preserving high efficiency. This model demonstrates superior overall detection performance compared to other algorithms, providing fresh perspectives and methods for advancing research and applications in drug detection and recognition.

KEYWORDS

Attention Mechanism, Drug Detection, Inner-Shape IoU, Large Separable Kernel Attention, SA-NET, YOLOv8s

INTRODUCTION

Drugs typically refer to substances capable of inducing both psychological and physiological dependence in individuals, posing significant risks to the physical and mental health of those who use them. This category includes not only legally prescribed medications that are misused, but also various natural plants, compounds, and organic solvents without medical use. According to the "World Drug Report 2021" released by the United Nations Office on Drugs and Crime, around 275 million people globally are involved in drug usage, resulting in at least 500,000 deaths every year, while over 36 million individuals develop mental health disorders because of drug abuse. Across the world, drug-related criminal activities are increasing, as evidenced by incidents such as drug-impaired driving accidents and a rise in violent crimes driven by addiction. Drug addiction results in both physical and psychological dependance, affecting bodily mechanisms and potentially reducing lifespan. Irrespective of the method of drug ingestion, drugs pose significant harm to the human body and

DOI: 10.4018/JOEUC.359770

contribute to the dissemination of infectious diseases. Moreover, drug-related activities disrupt societal peace and escalate criminal behavior, posing a serious threat to social cohesion. The drug issue is a global challenge that endangers human health, undermines social stability, and hinders economic development, making it a pressing concern for nations around the world.

Social media has now become the main channel through which people communicate with others in their social networks. The rapid evolution of internet technology has not only propelled the growth of the self-media and live broadcasting industries, but also facilitated drug trafficking by offering new platforms and channels to connect people with drugs. Consequently, more and more drug transactions are occurring online, posing significant challenges to drug enforcement efforts. Whether internet users are adolescents or adults, everyone should be aware of the dangers posed by drugs and learn to identify them. Drug abuse not only harms individuals' physical health but also affects societal security and stability. Therefore, drug detection is of the utmost importance. Currently, drug detection primarily relies on chemical composition analysis and odor detection, which are the most effective methods available for detecting drugs. However, conducting such screenings can be arduous and complex, necessitating specialized personnel and equipment, making large-scale detection and identification unfeasible. If it were possible to scan and detect potentially illicit substances in video images, analyzing the features and objects within the footage could facilitate the identification and investigation of suspected narcotics. This approach might significantly aid in the recognition and screening of drugs. For example, heroin packages are often grayish-white rectangular blocks sealed with plastic film and wrapped in yellow opaque adhesive tape. Thus, discovering small paper packets with this specific packaging should raise serious concerns. However, even though image detection can spot potentially sensitive images, it cannot conclusively determine whether these items are drugs, as solely relying on video images is inadequate for precise drug identification. Ultimately, determining whether something is a drug still requires traditional drug detection methods for final confirmation. These methods may include chemical composition analysis, odor detection, and more, which can provide more precise and reliable detection results. Therefore, image detection of drugs can serve as an initial screening method, enhancing the efficiency and speed of drug detection and identification. Meanwhile, traditional detection methods can be used as a final confirmation measure to ensure the accuracy and reliability of detection results.

In recent years, the rapid advancement of deep learning technologies (Song et al., 2024), particularly convolutional neural networks (CNNs) and computer vision-based detection methods, has led to remarkable progress, enabling their successful application in various domains. For example, deep learning techniques have been widely employed in sentiment analysis models (Samir et al., 2023). This technological progress has also offered innovative solutions for drug detection and screening. Deep learning technologies have achieved significant breakthroughs in image processing and visual recognition (Chao et al., 2024), with deep learning-based object detection algorithms becoming more refined. These algorithms demonstrate substantial advantages in both speed and accuracy, requiring only training on labeled datasets. Compared to traditional object detection methods, they offer robust feature extraction capabilities. In contrast to methods based on chemical composition and odor traits, deep learning-powered drug detection algorithms highlight exceptional speed and efficiency. Representative deep learning algorithms include AlexNet, GoogLeNet, ResNet, a Faster Region-based-CNN (R-CNN), DenseNet, and the YOLO series. Object detection algorithms, as one of the fundamental tasks in computer vision, accurately identify and classify objects in images, finding applications in various real-world scenarios. Chen et al. (2022) argued that intersection over union (IoU) variance impacts imbalance optimization, leading to potential performance bottlenecks. To address this, they proposed an enhanced Faster R-CNN for high-quality iterative object detection, which iteratively samples and gathers target boxes from loop steps, increasing the number of high-IoU training samples. Feng et al. (2024) presented an improved YOLO model that breaks through the limitations of the presence of necks, allowing the model to access semantic and structural information. Wei et al. (2024) set out to replace the traditional category regression loss with

the contrast loss of text and region, breaking the category limitation of the model. Hou et al. (2024) proposed a simple transformer-style CNN by studying the internal structure of self-attention. Li et al. (2023) introduced the concept of bidirectional-path aggregation network-feature pyramid network in YOLOv8 to improve feature fusion across different scales and replaced certain convolutional modules with the GhostblockV2 structure, achieving significant accuracy improvements. However, it struggled to outperform other models in small object detection tasks. Moon et al. (2024) proposed a rotating bounding box multilevel feature pyramid transformer to optimize the performance of object detection. Gao et al. (2022) projected an improved Faster R-CNN algorithm, incorporating a feature pyramid and adding deformable convolutions to the backbone. They replaced region of interest pooling with region of interest alignment to prevent the loss of feature details. Although the algorithm enhanced detection performance in extreme conditions, its high detection time limited its ability to meet the demands for fast detection. Wu et al. (2024) presented a point transformer V3 that prioritized performance and simplicity, further improving the model's capabilities. Deng et al. (2023) added efficient channel attention mechanisms to each constraint satisfaction problem unit in the YOLOv5 backbone. They made lightweight improvements to the feature fusion module, proposing an improved attention-YOLOv5-Ghost algorithm that addressed issues like large parameter sizes, redundant gradient computation, and slow detection speed.

YOLOv8 is one of the fundamental models in the YOLO series, and compared to previous anchor-based detection methods, its adoption of anchor-free techniques offers higher detection accuracy and speed. YOLOv8s is a lightweight version of YOLOv8. Although the model is smaller in scale, it still performs well in terms of accuracy. It is well-suited for fast detection tasks and can maintain high detection speeds even in resource-constrained environments. The structure of YOLOv8s is designed for flexible adjustments, allowing easy integration and experimentation to verify the impact of different module changes on performance, so for this study, we selected the YOLOv8s model for experiments. However, due to the similarities and complexities among drugs and medications, the YOLOv8 detection algorithm still suffers from drawbacks, such as inadequate target perception and localization errors. In response to the issues identified in the preceding studies, this paper proposes an improved YOLO model based on YOLOv8s. The main contributions of this model are:

1. We introduce a large separable kernel attention (LSKA) module, combined with the C2f module at the neck-end to construct the C2f-LSKA module. This enhancement enables the network to better capture and leverage information within the image, thereby improving the accuracy of the object detection model while reducing computational complexity and memory usage.
2. We incorporate an attention module called the shuffle attention network (SA-Net), which adeptly combines two attention mechanisms. This module partitions the channel dimension into multiple sub-features to decrease the computational burden. Subsequently, the shuffle units merge the complementary channel and shuffle channel modules for each sub-feature set. This strategy delivers enhanced performance with reduced model complexity.
3. We utilize the LSKA module to refine the single spatial pyramid pooling fusion (SPPF) module within the backbone network and incorporate attention mechanisms to revamp the pyramid pooling layer. This enhancement significantly improves the feature extraction prowess of the backbone network, enhances recognition capabilities for detected targets, and bolsters feature fusion capabilities.
4. We introduce the shape-IoU loss and the inner-IoU loss and combine them to construct an inner-shape IoU loss function, replacing the original complete intersection over union (CIoU) loss function. This innovative adjustment enhances the model's generalization capabilities and detection accuracy while improving the regression performance of detection bounding boxes.

RELATED WORK

Object Detection

Object detection is one of the core tasks in computer vision, and in recent years, the application of CNNs in this field has gradually increased. The development of target detection has gone through two stages, and the first stage is the traditional target detection algorithm, which mainly relies on manual feature extraction. The second stage is the detection algorithm based on deep learning. The latter is further categorized into anchor-based detection algorithms and anchorless detection algorithms based on whether or not anchors are used. Anchor-based algorithms include the RetinaNet (Lin et al., 2017), Faster R-CNN (Girshick, 2015), YOLO (Redmon et al., 2016; Redmon & Farhadi, 2018; Bochkovskiy et al., 2020; Wang et al., 2023), and Single Shot Multi-Box Detector (SSD) (Liu et al., 2016). Anchor-free detection algorithms include the Fully Convolutional One Stage Object detection (FCOS) (Tian et al., 2022), CornerNet (Law & Deng, 2018), and CenterNet (Duan et al., 2019). In these object detection algorithms, the bounding box regression loss function is crucial for allowing detectors to precisely localize targets, improving the overall detection accuracy of the model.

YOLOv8 Model

The YOLOv8 model, launched by Ultralytics in January 2023, offers five variants—v8x, v8l, v8m, v8s, and v8n—ranging from largest to smallest. As the model size increases, so does its accuracy, making it suitable for various tasks such as object detection, image classification, instance segmentation, and keypoint detection. Compared to the YOLOv5 model, the YOLOv8 replaces the original C3 module in the backbone network with the C2f module, which allows for improved gradient flow through additional skip connections and split operations while maintaining a lightweight design. It also introduces a decoupled-head structure by separating the localization and classification branches, eliminating parameter sharing between the two tasks, and resolving conflicts arising from joint training. This enhances overall model performance. Additionally, the YOLOv8 model incorporates distribution focal loss alongside CIoU loss for regression, further improving the model's precision and effectiveness.

Attention Mechanism

The significance of attention mechanisms has been widely explored and applied in previous studies. It tends to allocate the most informative feature representations while suppressing less relevant ones. Self-attention mechanisms compute the weighted sum of contextual information for a specific position by considering all positions within the image. In the SE, Hu et al. (2018) modeled the relationships between channels using two fully connected layers. Wang et al. (2018) introduced the non-local module, which generates an attention map by calculating the correlation matrix between spatial points in the feature map. The efficient channel attention (ECA) net (Wang et al., 2020) employs a 1D convolution filter to generate channel weights, significantly reducing the complexity of the SE model. The convolutional block attention module (Woo et al., 2018), the global context network (Cao et al., 2019), and the search generative experience (Li et al., 2019) sequentially combined spatial and channel attention, while the dual attention network (Fu et al., 2019) adaptively integrated local features with their global dependencies by adding two attention modules from different branches.

Bounding Box Regression Losses

Object detectors based on bounding box regression loss have been widely adopted in computer vision due to their simplicity and efficiency. The accuracy of localization algorithms within the loss function significantly affects the average precision of detection results. The Ln-norm loss (Girshick, 2015), a type of bounding box regression loss, is highly sensitive to changes in the scale of bounding boxes. Subsequently, to address this deficiency, the IoU loss (Yu et al., 2016) was introduced as a replacement for Ln-norm loss, offering more accurate regression results for predicted boxes.

However, the IoU loss struggles with the gradient vanishing problem for non-overlapping samples. The generalized-IoU loss (Rezatofighi et al., 2019) compensates for this by introducing a minimum enclosing box. The distance-IoU loss (Zheng et al., 2020) incorporates distance constraints, adding the normalized distance between the center points of the predicted box and the ground truth (GT) box as a new loss term, improving convergence speed and localization accuracy. CIoU loss further enhances the regression process by considering the shape similarity between boxes and adding a shape penalty to distance-IoU loss. The efficient IoU loss (Zhang et al., 2022) addresses imbalance during training with focal loss and redefines the shape loss, further enhancing detection performance.

METHOD

The YOLOv8s algorithm boasts a relatively high accuracy and fast detection speed among its series. However, when confronted with drug detection and recognition, challenges such as small target sizes, class imbalances, and similarities with other objects make it difficult to extract features of certain drugs. The original YOLOv8s model exhibits shortcomings like missed detections, subpar accuracy, and hefty computational demands. Therefore, this paper introduces a drug recognition detection model based on an enhanced YOLOv8s architecture. The overall network structure of the improved model is depicted in Figure 1.

Modification and Optimization of C2f Module

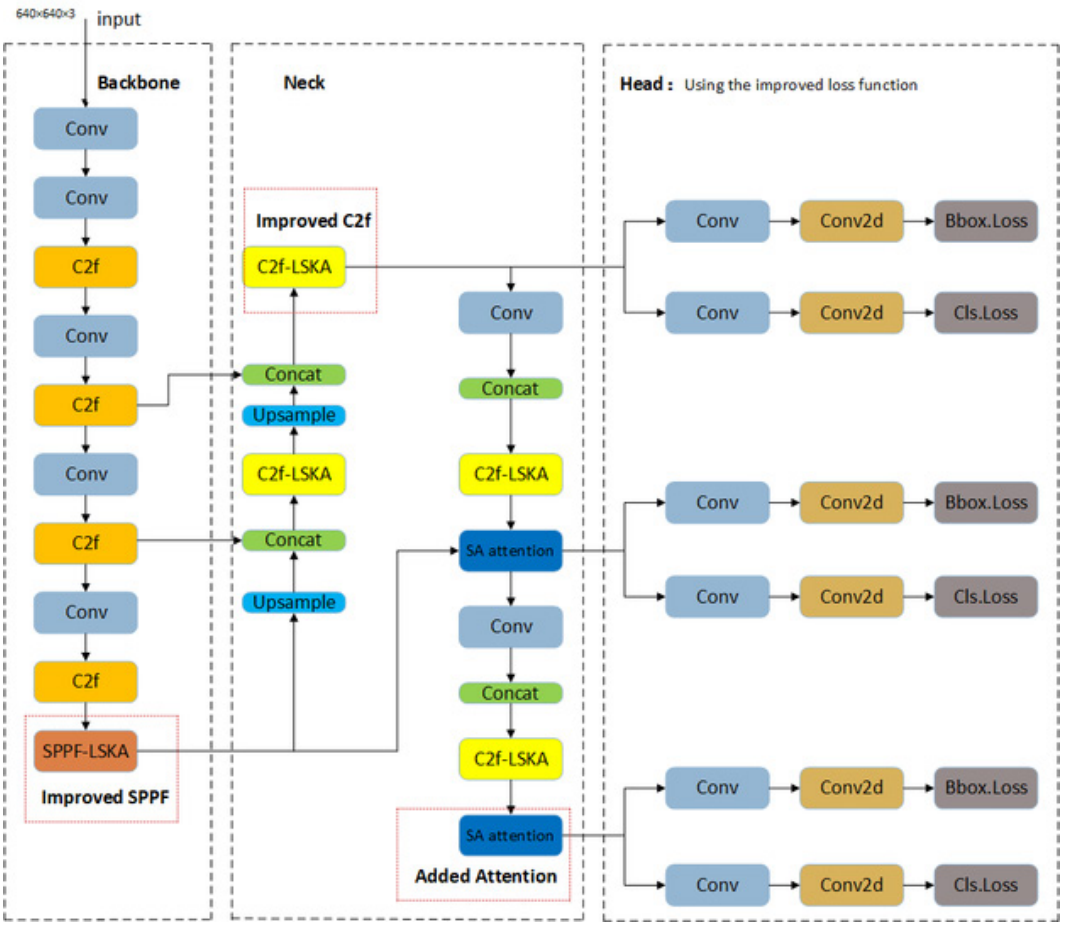
The LSKA (Lau et al., 2023) module enhances the large-kernel attention (LKA) module. It achieves this by decomposing the two-dimensional convolution kernels in depthwise convolution layers into stacked, one-dimensional kernels, applied separately in the horizontal and vertical directions. This decomposition allows large kernels to be directly employed within the attention mechanism, eliminating the need for extra blocks and lowering both the computational complexity and memory usage. As the convolutional kernel size increases, this approach significantly reduces computational overhead. Traditional convolution extracts features by sliding a kernel of fixed size across the input image. In contrast, large kernel separable convolution uses larger kernels to capture broader spatial information, breaking the kernel down into smaller components that independently perform convolutions in horizontal and vertical directions. This decomposition reduces the number of parameters in the model, thus improving computational efficiency while preserving spatial information. As a result, the model's capacity to interpret input images is enhanced, which leads to improved performance and generalization abilities.

In the original YOLOv8s network, the C2f module consists of standard convolutions (Conv) and multiple bottleneck blocks. These blocks are interconnected with numerous skip connections and additional split operations, contributing to a more complex network structure and increased computation. For this problem, we use the LSKA_Attention module to replace the original bottleneck module in the C2f module. This integration forms a new C2f-LSKA module by combining the C2f backbone network module with the LSKA_Attention module, as shown in Figure 2. Compared to the original C2f module, the improved C2f-LSKA module achieves optimization in detection speed and computational complexity. This enhancement enables the network to capture and utilize information from images more effectively, thereby increasing the precision of the object detection model while reducing computational complexity and memory usage.

Structural Adjustment of SPPF Module

In the YOLOv8 model, the main objective of the SPPF module is to extract features from different scales and fuse them to improve the performance and accuracy of target detection. The core advantage of the SPPF module is its ability to capture features across different scales and combine them, thereby strengthening the model’s capability to represent various targets. This multi-scale feature extraction and fusion strategy improves the performance and robustness of object detection, enabling the model

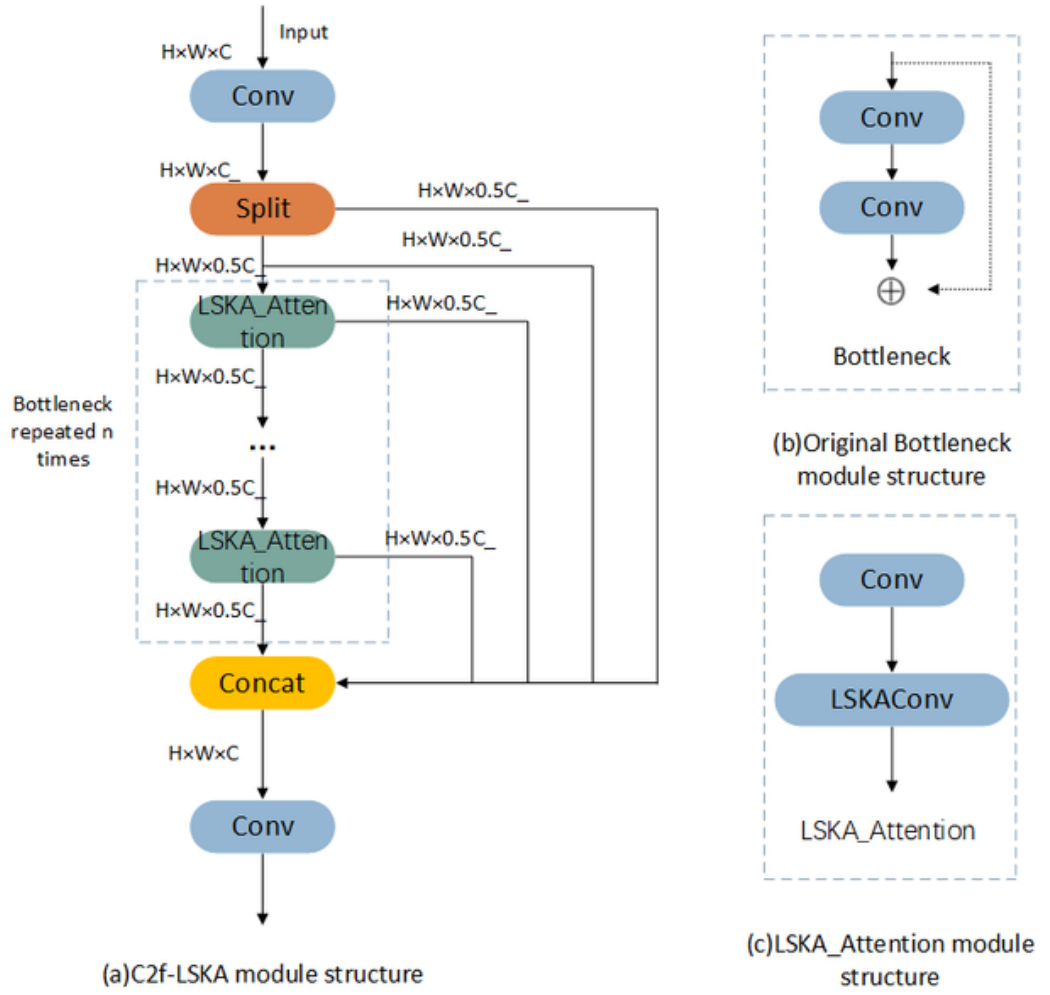
Figure 1. Improved YOLOv8 network structure diagram



to better adapt to targets of various sizes and proportions. The SPPF module processes input through three consecutive 5×5 max-pooling layers, concatenating their outputs to obtain multi-scale features. Additionally, the LSKA module addresses the challenge of large parameter growth in traditional convolutions by dividing a k×k convolution into separable kernels of kx1 and 1xk. These are applied in a cascading fashion, enabling more efficient processing of input features while maintaining accuracy and computational efficiency. This approach mitigates the computational cost typically associated with large convolutional kernels.

Integrating the LSKA module with the SPPF module, as shown in Figure 3, involves feeding the concatenated outputs from the multiple pooling layers of the original SPPF module into an 11×11 LSKA convolution module. This setup utilizes large separable convolutional attention to capture extended dependencies, thereby expanding the receptive field for more comprehensive feature extraction. The features are then fused through standard convolution, adjusting the final output feature vector size of the backbone section of the model. Integrating the LSKA attention mechanism into the SPPF module does not impose a substantial parameter overhead. Yet, it enhances the model's perception of multi-scale features, providing richer contextual information and aiding in more effective feature fusion. We use a large-scale separable convolutional attention module to refine the singular SPPF module in the backbone network and integrate the attention mechanism into the modified

Figure 2. Schematic diagram of C2f-LSKA module and sub-module

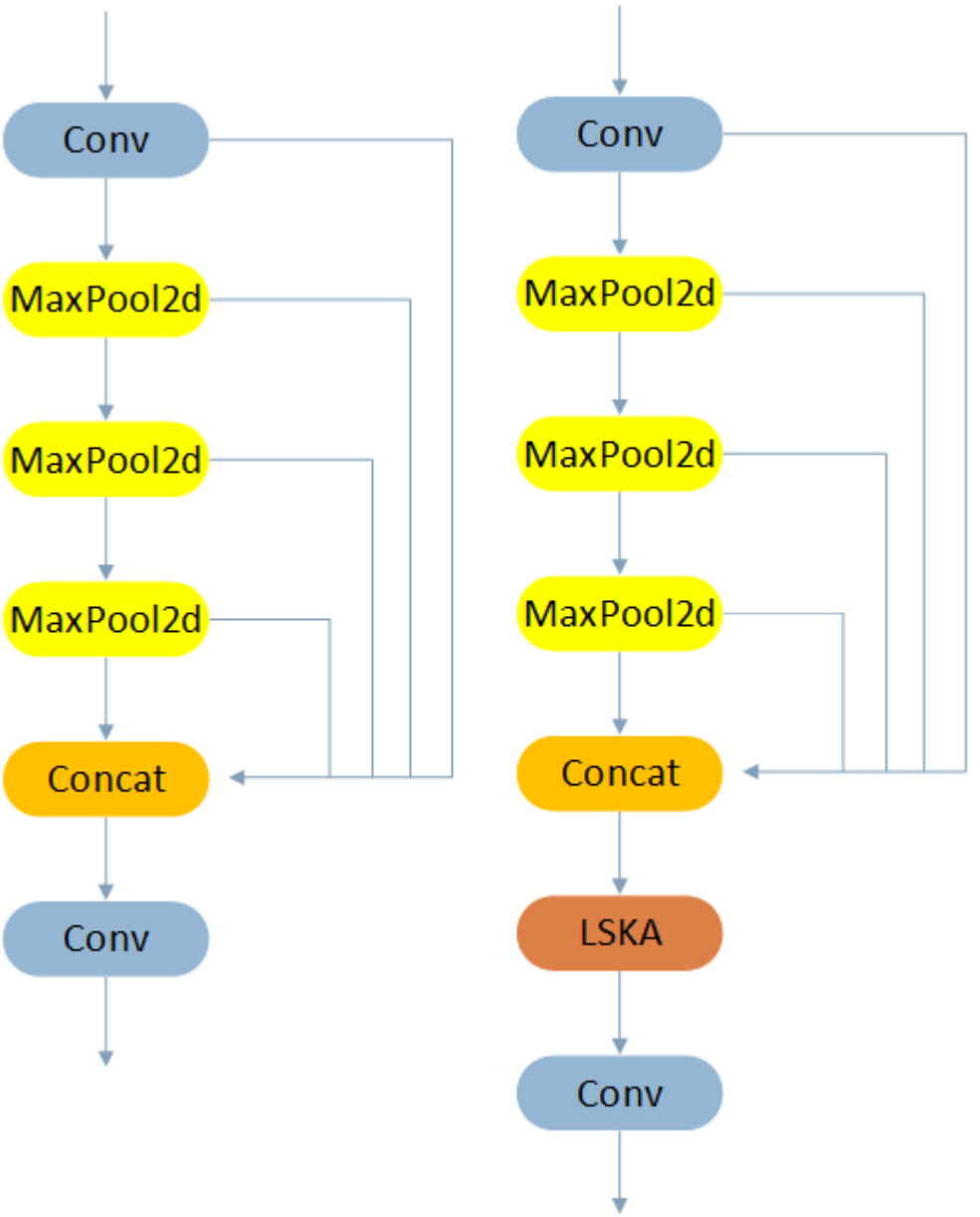


pyramid pooling layer. This enhancement effectively strengthens the feature extraction capability of the backbone network, thus enhancing its recognition ability to detect target objects.

SA-Net

Attention mechanisms provide significant flexibility and enhance the learning of discriminative feature representations, facilitating their seamless integration into algorithm backbone networks. Currently, attention mechanisms mainly fall into two categories: spatial attention mechanisms and channel attention mechanisms. The spatial attention, epitomized by the chemically aware model builder (CAMB) attention mechanism, focuses on establishing cross-channel spatial information by leveraging semantic dependencies between feature map spatial dimensions and channel dimensions. On the other hand, the channel attention, as demonstrated by the SE attention mechanism, explicitly models interactions between different channels to capture channel-specific attention. The primary aim of these mechanisms is to capture both pixel-level relationships and inter-channel dependencies. While using both types of attention mechanisms simultaneously can lead to improved performance, it also increases computational complexity.

Figure 3. Structure of SPPF and SPPF-LSKA



Consequently, this paper introduces the SA-NET, an attention mechanism that effectively combines two different attention mechanisms, as shown in Figure 4. The SA module is proposed to address this issue, effectively integrating two types of attention mechanisms using a shuffle unit. In detail, the SA module first partitions the channel dimension into several sub-features and processes them concurrently. Then, for each sub-feature, the SA module employs a shuffle unit to capture feature dependencies across both spatial and channel dimensions. Afterward, all sub-features are aggregated and merged, utilizing the "channel shuffle" operator to enable communication among different sub-features.

Figure 4. Structure of SA-NET

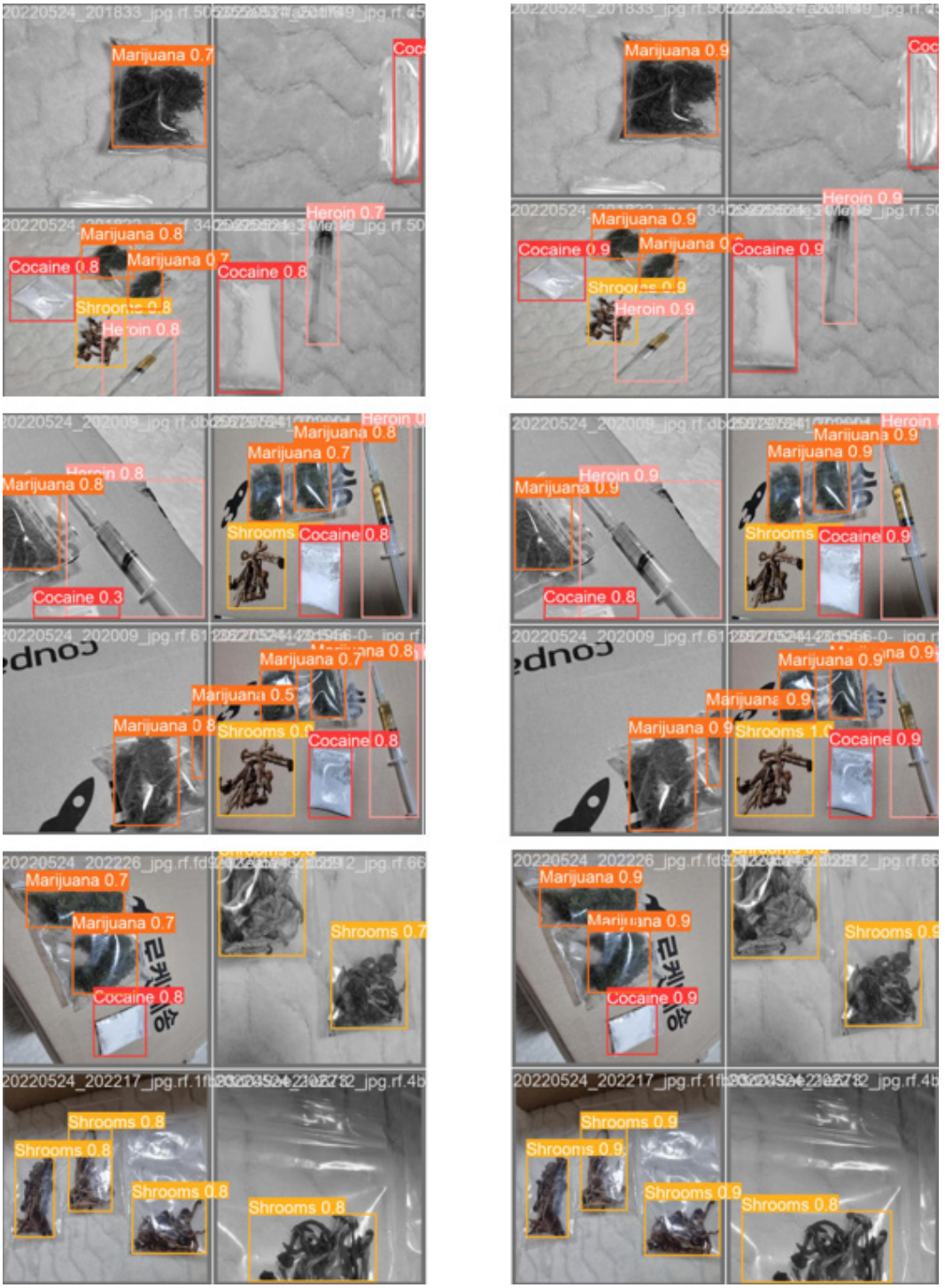


Figure 4 illustrates that the SA module employs a "channel split" approach to concurrently process

each group of sub-features. For the channel attention branch, the SA utilizes global average pooling to embed global information, generating channel-wise statistics, as $s \in \mathbb{R}^{\frac{C}{2G}\times 1\times 1}$, shown in Equation 1.

$$s = \Phi_{gp}(X_{k1}) = \frac{1}{H\times K}\sum_{i=1}^{H}\sum_{j=1}^{W}X_{k1}(i,j) \tag{1}$$

Then, a gate mechanism with a sigmoid activation function is utilized to create concise features, defined by Equation 2.

$$X_{k1}^{'} = \sigma\left(\Phi_c(s)\right)\cdot X_{\mathrm{K1}} = \sigma\left(W_{1\mathrm{s}} + \mathrm{b}_1\right)\cdot X_{\mathrm{k1}} \tag{2}$$

where W_1, $\mathrm{b}_1 \in \mathbb{R}^{\frac{C}{2G}\times 1\times 1}$, using this pair of parameters to scale and shift the channel vector, respectively.

In the spatial attention branch, the SA module employs group norms to generate spatial statistical data and then creates compact features similar to the channel branch based on this. Initially, spatial statistics for X_{k2} are obtained using group normalization (GN), followed by $\Phi_{\mathrm{c}}(\cdot)$ to enhance X_{k2} spatial attention, resulting in the final output, as shown in Equation 3.

$$X_{k2}^{'} = \sigma\left(W_2\cdot GN(X_{k2}) + b_2\right)\cdot X_{k2} \tag{3}$$

where W_2, $b_2 \in \mathbb{R}^{\frac{C}{2G}\times 1\times 1}$.

When finished, the two branches are connected with the number of channels equal to the number of inputs, as $X_k^{'} = \left[X_{k1}^{'}, X_{k2}^{'}\right] \in \mathbb{R}^{\frac{C}{2}\times H\times \Omega}$. Then all sub-features are aggregated, and finally, the "channel shuffle" operator is used to realize the information communication between different sub-features.

The SA attention mechanism is designed by integrating group convolution (to reduce computation), spatial attention mechanism (implemented with GN), channel attention mechanism (similar to SENet), and ShuffleNetV2 (using channel shuffle to blend information across different groups). It reduces computational load by introducing group convolution, applying spatial and channel attention to each group, and facilitating an information exchange using the channel shuffle operation. The advantage of the SA module lies in its ability to dynamically adjust the importance of feature maps at the channel level, while increasing the diversity and richness of feature maps through channel shuffling. This enhancement boosts the model's capacity for representation and adaptability to complex perceptual tasks.

Adjustment and Application of Loss Function

The loss function used for bounding box regression plays a crucial role in object detection tasks by measuring the discrepancy between predicted and GT bounding boxes. Similarly, the IoU loss function is commonly applied in computer vision tasks to directly measure the overlap between predicted and GT bounding boxes. During the bounding box regression process, this loss function not only assesses the accuracy of the regression but also facilitates gradient propagation by calculating the regression loss, which accelerates the convergence of the model. The mathematical definition of the IoU loss function is shown in Equation 4.

$$IoU\ Loss = 1 - IoU \tag{4}$$

where IoU stands for the ratio of the intersection area between the bounding boxes predicted by the model and the GT bounding boxes to their union area. The formula for calculating IoU is shown in Equation 5.

$$IoU = \frac{Area\ of\ Inter\ section}{Area\ of\ Union} \tag{5}$$

The intersection area of the bounding boxes refers to the region where the predicted bounding box overlaps with the GT bounding box, while the union area is the total area covered by both boxes combined minus the intersection area. The IoU loss function measure ranges between 0 and 1, where a value approaching 1 signifies a greater overlap between the predicted and GT bounding boxes, resulting in a lower loss. In contrast, a value approaching 0 indicates minimal overlap, leading to a higher loss. During the training process, the model aims to minimize the IoU loss function, thereby refining the alignment of predicted bounding boxes with the GT, which enhances the overall accuracy of object detection.

Within the YOLOv8 network, the CIoU functions as the bounding box regression loss function. It augments the computation of the IoU by introducing penalty terms for center point distance, aspect ratio difference, and area, resulting in a more accurate assessment of bounding box similarity. Consequently, it leads to an improved performance in object detection models. The formula for the CIoU is shown in Equation 6.

$$CIoU = IoU - \frac{\rho^2(C_{pred}\ C_{true})}{C^2} - \alpha \mathrm{v} \tag{6}$$

where $\rho^2\ C_{\mathrm{pred}}\ C_{\mathrm{true}}$ is the Euclidean distance between the center points of the predicted and GT bounding boxes, C^2 is the length of the diagonal that represents the smallest external rectangle, and v is a normalized aspect ratio difference term. The specific formulas are shown in Equations 7 and 8.

$$\mathrm{v} = \frac{4}{\pi^2} \cdot \left(\arctan\left(\frac{\mathrm{w}_{true}}{h_{true}}\right) - \arctan\left(\frac{w_{pred}}{h_{pred}}\right) \right)^2 \tag{7}$$

$$\alpha = \frac{v}{1 - IoU + v} \tag{8}$$

where w and h denote the width and height, respectively.

Current IoU-based edge regression methods tend to focus on speeding up convergence by introducing additional loss terms, often ignoring the inherent limitations of the IoU itself. While the IoU loss theoretically offers a robust representation of bounding box regression status, its lack of adaptability to diverse detectors and detection tasks hampers its generalization. Inner-IoU (Zhang et al., 2023) proposes a method of computing the IoU loss by utilizing auxiliary bounding boxes, with a scale factor ratio dictating the generation of auxiliary boxes across different scales for loss calculation, thereby improving the model's generalization. The specific computational formulas are shown in Equations 9 through 13.

$$b_l = x_c - \frac{w * ratio}{2}, b_r = x_c + \frac{w * ratio}{2} \tag{9}$$

$$b_t = y_c - \frac{h * ratio}{2}, b_b = y_c + \frac{h * ratio}{2} \tag{10}$$

$$inter = \left(\min\left(b_r^{gt}, b_r\right) - \max\left(b_l^{gt}, b_l\right)\right) * \left(\min\left(b_b^{gt}, b_b\right) - \max\left(b_t^{gt}, b_t\right)\right) \tag{11}$$

$$union = (w^{gt} * h^{gt}) * (ratio)^2 + (w * h) * (ratio)^2 - inter \tag{12}$$

$$Io\,U^{inner} = \frac{inner}{union} \tag{13}$$

where b^{gt}, b^{pred}denote the computed results of the truth and predicted bounding box, respectively, and w and h denote the width and height, respectively.

The inner IoU loss, while sharing some similarities with the traditional IoU loss, introduces unique aspects. It computes the IoU between auxiliary bounding boxes and ranges from 0 to 1. When the ratio of the auxiliary bounding box size to the actual bounding box size is less than 1, the auxiliary boxes are smaller, resulting in a reduced effective regression range compared to the IoU loss. This condition produces a larger gradient, facilitating faster convergence for high-IoU samples. Conversely, when the ratio exceeds 1, indicating larger auxiliary boxes, the effective regression range increases, which is advantageous for low-IoU regressions. Thus, employing smaller auxiliary bounding boxes can accelerate convergence for high-IoU cases, while larger boxes are more beneficial for low-IoU cases.

In current boundary box regression losses, the primary focus is on the geometric relationship between the predicted and GT boxes. These losses are determined by assessing their relative positions and shapes, often neglecting the intrinsic characteristics of the bounding boxes, such as their dimensions and proportions, and how these properties affect the regression process. Shape-IoU (Zhang & Zhang, 2023) proposes calculating the loss by focusing on the shape and scale of the boundary boxes themselves, making boundary box regression more precise. The specific calculation formulas are shown in Equations 14 through 19.

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{14}$$

$$ww = \frac{2 \times (w^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{15}$$

$$hh = \frac{2 \times (h^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{16}$$

$$\text{distance}^{shape} = hh \times \left(x_c - x_c^{gt}\right)^2 / c^2 + ww \times \left(y_c - y_c^{gt}\right)^2 / c^2 \tag{17}$$

$$\Omega^{shape} = \sum_{t=w,h} (1 - e^{-wt})^{\theta}, \theta = 4 \tag{18}$$

$$\begin{cases} \Omega_w = hh \times \dfrac{|w - w^{gt}|}{\max(w, w^{gt})} \\ \Omega_h = ww \times \dfrac{|h - h^{gt}|}{\max(h, h^{gt})} \end{cases} \tag{19}$$

where the scale factor, denoted as scale, is related to the size of the targets in the dataset, while ww and hh represent the weight coefficients in the horizontal and vertical directions, respectively, which are related to the shape of the GT box. The corresponding boundary box regression loss calculation formula is shown in Equation 20.

$$L_{shape-IoU} = 1 - IoU + \text{distacne}^{shape} + 0.5 \times \Omega^{shape} \tag{20}$$

When the GT bounding box is not square, meaning there is a difference between its length and width, the shape and scale variations of the regression samples lead to differences in their IoU values. For samples of the same scale, the box's shape influences the IoU values, with more noticeable effects observed along the shorter side of the box. Conversely, for samples with the same shape, smaller-scale samples experience a greater impact on their IoU values due to the shape of the GT box compared to larger-scale samples.

Combining the inner-IoU with the shape-IoU to construct the inner-shape IoU loss function, replacing the original CIoU loss function, allows for better focus on the shape and proportion of the

Table 1. Training parameter settings

parameter	setting	parameter	setting
epochs	200	close_mosaic	10
patience	50	warmup_epochs	3.0
batch	16	lr0	0.01
imgsz	640	lrf	0.01
workers	4	warmup_momentum	0.8
optimizer	SGD	weight_decay	0.0005

bounding boxes themselves, as well as the differences between the auxiliary boxes and the actual boxes when calculating the loss. This enhances the model's generalization ability and detection accuracy and improves the regression performance of the detection boxes.

EXPERIMENT

Experimental Environment and Parameter Settings

Regarding the experimental setup, computations were performed on NVIDIA A30 GPUs with 24GB of memory in a server environment. The code was developed, trained, and tested on the Red Hat Linux operating system (version 4.8.5). Programming was done using Python (version 3.8.12), and the PyTorch deep learning framework (version 2.0.0) was utilized for model construction. Additionally, model training acceleration was achieved through CUDA 11.7 and cuDNN 8.5.0, ensuring consistency in the hardware and software environment during training. For parameter configuration, the SGD optimizer, which is most commonly used in the field of computer vision, was used for gradient descent, and the specific parameter settings for the experimental environment are shown in Table 1.

Experimental Dataset

Acquiring datasets related to drugs is subject to regulations and legal restrictions due to the sensitive nature and potential involvement in illegal activities associated with images depicting drug-related content. Obtaining a large-scale drug dataset through legitimate channels is exceedingly challenging. Therefore, the dataset primarily used in the experiments is sourced from the drug dataset available on the dataset management platform Roboflow. Roboflow is a dataset management platform recommended by the YOLOv8 official website, providing free datasets and supporting the upload of custom datasets for format conversion. We carefully curated the experimental dataset by combining multiple datasets from Roboflow. We applied rigorous screening and cleaning processes to remove low-quality, duplicated, and redundant samples. Afterward, we employed specific data augmentation techniques to enhance the dataset's quality and accuracy. Our dataset comprises 5,560 images captured from various scenes and includes four types of drug detection targets: cocaine, heroin, marijuana, and mushrooms. The training, validation, and test sets consist of 4,580, 440, and 540 images, respectively. All models in the experiment were trained on this dataset with an input image size of 640×640, following the aforementioned specifications.

Evaluation Metrics

To objectively assess the model's detection performance, this experiment employs the mAP, precision rate (Precision), recall rate (Recall), floating-point operations (FLOPs), and frames per second (FPS) as evaluation metrics. Specifically, the formulas for the Precision and Recall values are shown in Equations 21 and 22.

Table 2. Ablation experiments of SA-NET

Model	mAP@0.5(%)	mAP@0.5:0.95(%)	P(%)
YOLOv8s	85.2	60.6	92.1
YOLOv8s+1SA-Net	85.7	60.6	**93.4**
YOLOv8s+2SA-Net	**86.0**	**60.8**	**93.4**
YOLOv8s+3SA-Net	85.6	60.1	93.0
YOLOv8s+4SA-Net	85.6	60.0	92.8

$$\text{Precision} = \frac{TP}{TP + FP} \tag{21}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{22}$$

where TP represents the count of correctly detected samples, FP represents the count of incorrectly detected samples, and FN represents the count of missed detections. The AP value represents the accuracy of a single category, while the mAP value denotes the average accuracy across all categories. The formulas for the AP and mAP values are defined in Equations 23 and 24, respectively.

$$AP = \int_0^1 P(\mathrm{r})d\mathrm{r} \tag{23}$$

$$\mathrm{m}AP = \frac{1}{N}\sum_1^N A P_{\mathrm{i}} \tag{24}$$

The FPS value denotes the number of images that can be detected in a second, which is a measure of the detection speed of the model, and the FLOPs value denotes the amount of computation for the model, which is a measure of the computational complexity of the model.

Comparison of Improvement Methods Effect

To explore the practical impact of the proposed enhancements in this paper, we individually improve the YOLOv8s neck module and loss function, optimizing and adjusting the parameters accordingly. Our experiments aim to compare the effectiveness of these various enhancement methods.

Experiments on Neck Improvement

The YOLOv8s model's neck segment consists of four C2f modules, with the latter three attached to three detection heads. To identify the best enhancement point for the neck network, the SA-Net was appended after each C2f module. Specifically, adding the SA-Net after the last three C2f modules establishes direct connections to the detection heads. The improved model's precision was then measured.

As shown in Table 2, incorporating the SA-Net attention mechanism enhances the model's accuracy, with the highest mAP@0.5 increasing by up to 0.8%. However, introducing too many SA-Net modules into the network can negatively impact detection accuracy, reducing the mAP@0.5 improvements. Based on the experimental results, this paper will add the SA-Net after the last two Cf2 modules in the YOLOv8 neck.

In the YOLOv8s model, the neck section includes four C2f modules. To explore the optimal position for introducing the LSKAs, we replaced each C2f module with the proposed C2f-LSKA module and tested the accuracy of the improved model.

Table 3. Ablation experiments of LSKA

Model	mAP@0.5(%)	mAP@0.5:0.95(%)	GFLOPs
YOLOv8s	85.2	60.6	28.4
YOLOv8s+1C2f-LSKA	85.5	60.3	27.9
YOLOv8s+2C2f-LSKA	85.8	60.6	27.1
YOLOv8s+3C2f-LSKA	**86.0**	**60.7**	26.4
YOLOv8s+4C2f-LSKA	**86.0**	**60.7**	**25.3**

Table 4. Ablation experiments of IoU

Model	mAP@0.5(%)	mAP@0.5:0.95(%)	P(%)
YOLOv8s+IoU	84.9	60.1	92.4
YOLOv8s+CIoU	85.2	60.6	92.1
YOLOv8s+Inner-IoU	85.9	60.6	94.0
YOLOv8s+Shape-IoU	86.3	61.0	93.6
YOLOv8s+inner-shape IoU	**86.6**	**61.3**	**94.3**

According to Table 3, incorporating the C2f-LSKA module can significantly enhance the model's detection accuracy, with the mAP@0.5 increasing by 0.8%. Additionally, it substantially reduces the model's computational complexity. As the number of C2f-LSKA modules increases, the model's detection accuracy improves, and its computational complexity decreases. Based on the experimental results, this paper replaces all four C2f modules in the YOLOv8s neck with C2f-LSKA modules.

Experiments on Loss Function Improvement

To explore the effectiveness of the inner-shape IoU loss function, we compared it with the inner-IoU, the shape-IoU, and the standard loss functions like the IoU and the CIoU. Table 4 shows that integrating the inner-IoU or the shape-IoU enhances model accuracy. As illustrated in Table 4, the YOLOv8s model with the inner-shape IoU loss function achieves higher detection accuracy. Compared to the original YOLOv8s model using the CIoU loss function, the mAP@50 increased by 1.4%, the mAP@0.5:0.95 improved by 0.7%, and the precision rose by 2.2%. This demonstrates that the inner-shape IoU loss function stabilizes the boundary box regression and improves prediction accuracy.

Ablation Study

To validate the effectiveness of the proposed algorithm improvements, we performed ablation experiments using the original YOLOv8s network as a baseline, as shown in Table 5. The experimental data show that replacing the original neck-end C2f module with the C2f-LSKA module increased the mAP@0.5 by 0.8 percentage points and reduced the computation by 3.1G. Substituting the original SPPF module in the backbone network with the SPPF-LSKA feature pyramid module resulted in an additional 0.2% increase in the mAP@0.5. Introducing the inner-shape IoU loss function instead of the original CIoU loss function added another 0.7 percentage points to the mAP@0.5. Incorporating the SA-Net further increased the mAP@0.5 by 0.5%, the mAP@0.5:0.95 by 0.7 percentage points, and the precision by 1.1%. Although the final reduction in computation was only 1.8G, the detection accuracy improved by a total of 2.2 percentage points, significantly boosting the comprehensive detection performance. Therefore, the improved YOLOv8s model proposed in this paper demonstrates higher accuracy in drug detection and recognition than the YOLOv8s baseline model, confirming the effectiveness and feasibility of the optimization modules.

Table 5. Ablation study

C2f-LSKA	SPPF-LSKA	SA-Net	inner-shape IoU	mAP@0.5(%)	mAP@0.5:0.95(%)	P(%)	R(%)	GFLOPs
				85.2	60.6	92.1	79.2	28.4
√				86.0	60.7	94.7	78.8	**25.3**
	√			85.8	60.5	94.5	79.7	29.3
		√		86.0	60.8	93.4	78.8	28.4
			√	86.6	61.3	94.3	**79.9**	28.4
√	√			86.2	60.7	94.4	79.8	26.2
√	√	√		86.3	60.5	93.3	78.1	26.2
		√	√	86.6	61.2	94.2	79.1	28.4
√	√		√	86.9	61.0	94.0	79.7	26.3
√	√	√	√	**87.4**	**61.7**	**95.1**	79.6	26.6

Table 6. Comparative experiments

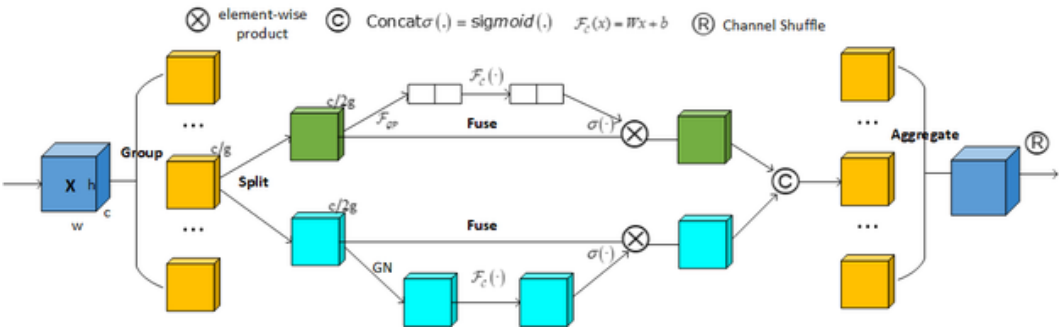
Model	GFLOPs	P(%)	R(%)	mAP@0.5(%)	mAP@0.5:0.95(%)	FPS
YOLOv5s	16.0	93.7	78.7	85.3	58.7	243.9
YOLOv9-c	238.9	93.3	78.6	86.7	60.7	32.57
YOLOv3	155.3	93.5	79.7	86.5	60.3	64.51
YOLOv7	105.2	93.9	79.1	86.1	60.8	68.49
Gold-yolo	32.1	94.1	77.4	85.8	60.4	135.13
Rtdetr-l	108.1	94.9	**81.0**	86.7	**62.1**	68.66
YOLOv8n	8.2	92.6	77.1	84.8	59.6	277.78
YOLOv8s(baseline)	28.4	92.1	79.2	85.2	60.6	156.25
YOLOv8m	79.1	93.9	80.7	86.3	61.0	98.03
YOLOv8s-Wrold	39.9	93.5	79.8	85.7	60.8	113.78
Ours	26.6	**95.1**	79.6	**87.4**	61.7	144.54

Comparison Experiments

To evaluate the detection performance of the improved model and comprehensively verify the advantages of the proposed algorithm, we compared the improved model with representative networks such as YOLOv5s, YOLOv7, YOLOv9c, and other enhanced YOLO models, as shown in Table 6.

The experimental results show that while the YOLOv8n model has the fastest detection speed and the lowest computational load, its detection accuracy is insufficient. YOLOv9-c, YOLOv3, and Rtdetr-l, despite their higher detection accuracy, suffer from excessive computational requirements and slow detection speeds, failing to meet rapid detection demands. The proposed improved model, although slightly slower in detection speed, significantly outperforms other mainstream models in detection accuracy, precision, and computational efficiency, with a precision rate of 87.4%, which is 2.2 percentage points higher than the original YOLOv8s model. Considering computational complexity, detection accuracy, and speed, the proposed algorithm excels compared to numerous mainstream algorithms.

Figure 5. Comparison of object detection results



Visualization Comparison

To intuitively verify the effectiveness of the improved algorithm, we conducted a visual comparison by training both the original YOLOv8s model and the improved model under the same experimental conditions and parameters. The left side of Figure 5 shows the detection results of the original model, while the right side displays the results of the improved model.

From the visual comparison, it is clear that both models accurately identified the drug types in the depicted scenarios. However, the improved algorithm shows significantly higher detection confidence than the original YOLOv8s algorithm. Based on the visual comparison and the previous analysis, it is apparent that the enhanced algorithm substantially improves the detection accuracy over the existing method.

DISCUSSION AND CONCLUSIONS

This article proposes an efficient drug detection and recognition model based on an improved YOLOv8s. By integrating large kernel separable convolution into the C2f modules in the neck, the model enhances multi-scale fusion capabilities while reducing computational complexity and memory usage. Adding the LSKA to the SPPF improves semantic fusion across different feature layers. Incorporating the SA-Net attention mechanism increases model robustness and reduces computation load. Using the combined inner-IoU and shape-IoU as the inner-shape IoU loss function for bounding box regression enhances detection accuracy and generalization ability, improving the bounding box regression performance. The improved model, which is smaller and faster than the YOLOv8m model, achieves performance that surpasses the YOLOv8m in nearly all aspects, meeting the requirements for rapid drug detection with its compact size and high accuracy. Compared to the original YOLOv8s, the improved model shows a 2.2% increase in the mAP@0.5 and a 1.1% increase in the mAP@0.5:0.95, while maintaining detection speed, making it highly practical. In online marketplaces like TaoBao, the model can help detect and prevent the illegal sale of drugs, which is a significant issue in some regions. The model could also help monitor user-uploaded images to detect disguised or hidden drugs in product photos. The model can be deployed in high-risk environments like nightclubs, concerts, or festivals, where drug usage is a concern. It could be integrated with security scanning systems to detect hidden or disguised drugs in bags, clothing, or personal items, helping security personnel prevent the entry of illegal substances into venues. However, the improved model occasionally misses detections in complex environments. In drug detection, "complex environments" refer to challenging conditions such as cluttered backgrounds, hidden drugs, or items designed to deceive detection systems. For example, drugs hidden in inconspicuous items or mixed with other substances may make accurate detection difficult. The model might struggle to differentiate drugs when they are camouflaged or purposefully obscured in a variety of real-world settings because it may not recognize subtle differences

in appearance, which requires enhanced feature detection or enhanced feature detection combined with other scanning technologies such as x-rays. Lack of proper lighting reduces the model's ability to distinguish drug-related features, leading to potential false negatives. Addressing this might require using auxiliary technologies like infrared imaging or night-vision sensors alongside the model. Future work will further optimize the algorithm structure to improve its detection accuracy and precision.

AUTHOR NOTE

Dingju Zhu (https://orcid.org/0000-0002-5907-3349)
Andrew W. H. Ip (https://orcid.org/0000-0001-6609-0713)
The authors of this publication declare there are no competing interests.
This research is partially supported by the Research Centre for Deep Space Explorations (RCDSE) of the Hong Kong Polytechnic University.

PROCESS DATES

October 18, 2024
Received: June 14, 2024, Revision: October 3, 2024, Accepted: October 4, 2024

CORRESPONDING AUTHOR

Dingju Zhu (China, zhudingju@m.scnu.edu.cn)

REFERENCES

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *ArXiv.org, arXiv preprint arXiv:2004.10934*. https://doi.org//arXiv.2004.10934DOI: 10.48550

Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). GCNet: Non-Local networks meet squeeze-excitation networks and beyond. *IEEE Xplore*, 1971–1980. DOI: 10.1109/ICCVW.2019.00246

Chao, Y., Zhu, H., & Zhou, Y. (2024). Integrating visual transformer and graph neural network for visual analysis in digital marketing. *Journal of Organizational and End User Computing*, *36*(1), 1–28. DOI: 10.4018/JOEUC.342092

Chen, X., Li, H., Wu, Q., Meng, F., & Qiu, H. (2022). Bal-R2CNN: High quality recurrent object detection with balance optimization. *IEEE Transactions on Multimedia*, *24*, 1558–1569. DOI: 10.1109/TMM.2021.3067439

Deng, L., Liu, Z., Wang, J., & Yang, B. (2023). ATT-YOLOv5-Ghost: Water surface object detection in complex scenes. *Journal of Real-Time Image Processing*, *20*(5), 97. Advance online publication. DOI: 10.1007/s11554-023-01354-z

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). CenterNet: Keypoint Triplets for Object Detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6568–6577. DOI: 10.1109/ICCV.2019.00667

Feng, Y., Huang, J., Du, S., Ying, S., Yong, J.-H., Li, Y., Ding, G., Ji, R., & Gao, Y. (2024). Hyper-YOLO: When visual object detection meets hypergraph computation. *ArXiv Preprint ArXiv:2408.04804*. https://doi.org//arxiv.2408.04804DOI: 10.48550

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3141–3149. DOI: 10.1109/CVPR.2019.00326

Gao, X., Chen, L., Wang, K., Xiong, X., Wang, H., & Li, Y. (2022). Improved traffic sign detection algorithm based on faster R-CNN. *Applied Sciences (Basel, Switzerland)*, *12*(18), 8948. DOI: 10.3390/app12188948

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440–1448. DOI: 10.1109/ICCV.2015.169

Hou, Q., Lu, C.-Z., Cheng, M.-M., & Feng, J. (2024). Conv2Former: A simple transformer-style ConvNet for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, ●●●, 1–10. DOI: 10.1109/TPAMI.2024.3401450 PMID: 38748521

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141. DOI: 10.1109/CVPR.2018.00745

Lau, K. W., Po, L.-M., & Abbas, Y. (2023). Large separable kernel attention: Rethinking the large kernel attention design in CNN. *SSRN*, 1–50. DOI: 10.2139/ssrn.4463661

Law, H., & Deng, J. (2018). CornerNet: Detecting objects as paired keypoints. *In Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750. DOI: 10.48550/arxiv.1808.01244

Li, X., Hu, X., & Yang, J. (2019). Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *ArXiv (Cornell University) Preprint ArXiv:1905.09646*. https://doi.org//arxiv.1905.09646.DOI: 10.48550

Li, X., Hu, X., & Yang, J. (2019). Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. arXiv preprint arXiv:1905.09646.

Li, Y., Fan, Q., Huang, H., Han, Z., & Gu, Q. (2023). A modified YOLOv8 detection network for UAV aerial image recognition. *Drones (Basel)*, *7*(5), 304. DOI: 10.3390/drones7050304

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. DOI: 10.1109/ICCV.2017.324

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *Lecture Notes in Computer Science*, *2016*(9905), 21–37. DOI: 10.1007/978-3-319-46448-0_2

Moon, J., Jeon, M., Jeong, S., & Oh, K.-Y. (2024). RoMP-transformer: Rotational bounding box with multi-level feature pyramid transformer for object detection. *Pattern Recognition*, *147*, 110067–110067. DOI: 10.1016/j.patcog.2023.110067

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. DOI: 10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *Arxiv.org*. https://doi.org//arXiv.1804.02767DOI: 10.48550

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666. DOI: 10.1109/CVPR.2019.00075

Samir, H. A., Abd-Elmegid, L. A., & Marie, M. M. (2023). Sentiment analysis model for airline customers' feedback using deep learning techniques. *International Journal of Engineering Business Management*, *15*, 18479790231206019. Advance online publication. DOI: 10.1177/18479790231206019

Song, Y., Du, H., Piao, T., & Shi, H. (2024). Research on financial risk intelligent monitoring and early warning model based on LSTM, transformer, and deep learning. *Journal of Organizational and End User Computing*, *36*(1), 1–24. DOI: 10.4018/JOEUC.353303

Tian, Z., Chu, X., Wang, X., Wei, X., & Shen, C. (2022). Fully convolutional one-stage 3D object detection on LiDAR range images. *Advances in Neural Information Processing Systems*, *35*, 34899–34911. DOI: 10.48550/arxiv.2205.13764

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7464–7475. . In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.DOI: 10.1109/CVPR52729.2023.00721

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11531–11539. DOI: 10.1109/CVPR42600.2020.01155

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/CVPR.2018.00813

Wei, G., Yuan, X., Liu, Y., Shang, Z., Yao, K., Li, C., Yan, Q., Zhao, C., Zhang, H., & Xiao, R. (2024). OVA-DETR: Open vocabulary aerial object detection using image-text alignment and fusion. *ArXiv Preprint ArXiv:2408.12246*.https://doi.org//arXiv.2408.12246DOI: 10.48550

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *Computer Vision – ECCV 2018. Lecture Notes in Computer ScienceComputer Vision, 11211*, 3–19. . Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer ScienceComputer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science.DOI: 10.1007/978-3-030-01234-2_1

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., & Zhao, H. (2024). Point transformer V3: Simpler, faster, stronger. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4840–4485. DOI: 10.48550/arXiv.2312.10035

Yu, J., Jiang, Y., Wang, Z., Cao, Z., & Huang, T. (2016). UnitBox: An advanced object detection network. *Proceedings of the 24th ACM International Conference on Multimedia*, 516–520. DOI: 10.1145/2964284.2967274

Zhang, H., Xu, C., & Zhang, S. (2023). Inner-IoU: More effective intersection over union loss with auxiliary bounding box. *ArXiv Preprint ArXiv:2311.02877*. https://doi.org//arxiv.2311.02877DOI: 10.48550

Zhang, H., & Zhang, S. (2023). Shape-IoU: More accurate metric considering bounding box shape and scale. *ArXiv Preprint ArXiv:2312.17663*. https://doi.org//arxiv.2312.17663DOI: 10.48550

Zhang, Y.-F., Ren, W., Zhang, Z., Jia, Z., Wang, L., & Tan, T. (2022). Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing*, *506*, 146–157. DOI: 10.1016/j.neucom.2022.07.042

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(07), 12993–13000. DOI: 10.1609/aaai.v34i07.6999

Dingju Zhu is a professor and doctoral supervisor at South China Normal University. He is also a convener of the artificial intelligence robot research team at the School of Artificial Intelligence, director of the Artificial Intelligence Robot Research Center at the School of Artificial Intelligence, and dean of the Guangdong Institute of Artificial Intelligence Robot Education Industry. He holds a doctorate from the University of the Chinese Academy of Sciences and is a postdoctoral fellow of Peking University, a postdoctoral fellow of the University of Macau, and a visiting scholar of Texas State University in the United States. In addition, Zhu is a visiting researcher of Shenzhen Advanced Technology Research Institute, Chinese Academy of Sciences, a Senior Member of China Computer Society, and he was awarded the titles of "Famous Teacher of South China Normal University," "Exemplary Individual of South China Normal University," "Outstanding Inventor of Guangdong Province," and "Local Talent of Shenzhen."

Zixuan Huang is a graduate student at the School of Artificial Intelligence, South China Normal University. He previously studied at Donghua University of Science and Technology, where he earned a bachelor's degree in software engineering. Zixuan's academic interests lie in computer vision and artificial intelligence. He aims to contribute to this field by developing advanced automatic detection systems by applying deep learning technology.

Kai-Leung Yung (kl.yung@polyu.edu.hk) is an Associate Head and Chair Professor of the Department of Industrial and Systems Engineering of The Hong Kong Polytechnic University. He received a BSc in Electronic Engineering at Brighton University, in 1975, a MSc, DIC in Automatic Control Systems at the Imperial College of Science, Technology, and Medicine, University of London, in 1976, and a PhD in Microprocessor Applications in Process Control at Plymouth University, in 1985, in the United Kingdom and became a Chartered Engineer in 1982. After graduation, he worked in the United Kingdom for companies such as BOC Advanced Welding Co. Ltd., the British Ever Ready Group, and the Cranfield Unit for Precision Engineering (CUPE). In 1986, he returned to Hong Kong to join the Hong Kong Productivity Council as Consultant and subsequently switched to academia to join the Department of Industrial and Systems Engineering of the Hong Kong Polytechnic University.

Wai-Hung Ip (wh.ip@polyu.edu.hk) has more than 30 years of experience in teaching, research, industry, and consulting. He received his PhD from Loughborough University in the UK, an MSc in Industrial Engineering from Cranfield University, and an LLB (Hons) from the University of Wolverhampton. After finishing postgraduate degrees in industrial engineering in the UK, his engineering research career began at that time with novel work in Industrial Engineering, Information, and Sensor Systems.