EasyChair Preprint

№ 12860

Study on the Plate Detection Method Based on the Data Generated by Significance Detection

Linzhong Fang, Xuanlai Tang, Shuyon Gao, Yicheng Song and Shengfeng Li

March 31, 2024

Study on the Plate Detection Method Based on the Data Generated by Significance Detection

*Note: Sub-titles are not captured in Xplore and should not be used

Linzhong Fang
School of Mechanical and Energy Engineering
Shanghai Technical Institute of Electronics&Information
Shanghai, China
ahfanglz@163.com

Xuanlai Tang
KEENON Robotics Co., Ltd
Shanghai, China
Tangxl@keenon.com

Shuyon Gao
School of Computer Science
Fudan University
Shanghai, China
sygao18@fudan.edu.cn

Yicheng Song
School of Computer Science
Fudan University
Shanghai, China
ycsong@fudan.edu.cn

Shengfeng Li
School of Communication and Information Engineering
Shanghai Technical Institute of Electronics&Information
Shanghai, China
lsf7679@163.com

Abstract: **In the smart restaurant environment, vision-based plate detection technology is the core of automated dish management. This study focuses on the identification of the existence of dishes on the plate, and constructs a dataset of dishes on the plate by collecting data from the actual restaurant scene. In order to complete the detection of dinner plates, a plate detection model based on Yolov5 was designed. Due to the limitation of the available restaurant dishes, the dataset has a single dish variety on the plate, which leads to the low performance of the trained plate detection model. In order to solve this problem, this paper proposes a plate data generation method based on saliency detection, which uses its category-independent characteristics to extract a variety of dish data from a variety of scenarios, and completes the annotation of the generated data by designing prompts and combining with Grounding DINO, which effectively solves the problem of small amount of plate data and single type. Experimental results show that the proposed plate detection model can effectively detect the presence of dishes on plates in a variety of restaurant environments, and the data generation method based on saliency detection significantly improves the quality of the dataset and the performance of the plate detection model.**

Keywords: plate detection, salient object detection, image editing, data generation

I. INTRODUCTION

With the rapid development of computer vision technology, the application of computer vision technology in smart restaurants becomes more and more extensive. In smart restaurants, the plate detection technology is particularly critical, which can help the restaurant to accurately manage the food supply, improve the quality and efficiency of service, and can be used in automatic settlement, customer diet structure suggestions, residual food calculation, automatic task process and other specific scenarios. However, the challenges for smart restaurants include accurately identifying a variety of dishes in complex environments, which requires computer vision technology to be adapted to the restaurant's varied application scenarios. Although the concept of smart restaurant has gradually popularized, there is still a relative lack of research on plate testing from the perspective of restaurant service. Therefore, this study aims to explore the plate detection task in the smart restaurant environment, aiming to improve the technical level and service quality of the smart restaurant. In order to accomplish the above tasks, our research needs to be conducted from two aspects:

First. Construction of dish detection data set. Build a data set of dishes with a wide variety of dishes and covering the various restaurant scene.

Second. Construction of the plate detection model. Accurately identify a variety of dishes on the plate, not being disturbed by a variety of information in the actual scene.

In terms of the construction of the plate detection data set, there is no public plate detection work at present, and the variety of dishes is various and complex, and it is difficult for several restaurant scenes to cover many dishes. Therefore, much richer datasets can be built based on conditionally generated data and real scenario data.

The current food image analysis work depends on the classification of dishes. Food images are processed into masks with the same number of categories by using segmentation model, which represent the confidence that the location is classified into the category. Once the category changes, the model needs to be retrained to accommodate the new category. To this end, the mainstream practice is to train larger models in data sets with larger data volumes and more data categories.

The food101 proposed in literature [1] is the first large-scale dataset in the field, which contains 110241 images from 101 food categories, and this work presents a baseline model for classifying a large number of food images based on machine learning methods, but its data are limited to Western dishes. Literature [2] focuses on the categories of Chinese food and proposes a data set containing 172 categories of Chinese dishes. It is not difficult to find that the dishes in different regions are quite different, and the visual characteristics of different cuisines have different bias. It is a

This research was funded by the National Natural Science Foundation of China (Grant No. 62072112).

very challenging task to solve the visual tasks of various cuisines widely through a model. The largest food data set is proposed in literature [3] at present. The data set presented in this work divides food products into 2000 classes, and more than one million finely screened pictures of food are provided, which can cover the vast majority of common food categories. For this purpose, a model of visual feature enhancement is designed. At the same time, literature [3] also points out that such large-scale visual task is extremely challenging, many downstream tasks are also worth further exploration. In addition, when the data category changes frequently or the long tail of the data problem intensifies, the model is usually difficult to show good results.

Significance detection as a pre-processing method for many visual tasks plays an important role in various visual tasks, which can extract the significance area in the image as a prior information to downstream tasks. Its important feature is category irrelevant, namely focus on the visual significant area, while ignoring the difference of category. Compared with the common food segmentation methods that rely on the classification and segmentation of dishes, the method based on significance detection shows more obvious advantages in a variety of scenarios. These scenes do not focus on the specific types of dishes, but focus on the ability to distinguish dishes from non-relevant areas through visual methods. Literature [4] improves foreground detection as a prior for target segmentation in weakly supervised segmentation tasks. The most common tests of significance usually include both fully supervised and weakly supervised significance tests. The fully supervised significance detection requires labeling the pixels one by one, which spends more time to process the data. For example, literature [5] is the classical fully supervised significance detection method, which uses the fully supervised data provided by the publicly available data set. Weak supervised detection can save the annotation time on the premise of sacrificing a small amount of detection accuracy. Method in literature [6,7,8] detect the significance models using picture-level categorical labels. For example, method in the literature [9,10,11,12] only needs a small number of pixels. With the help of the visual perception ability of the model, it can save a lot of annotation work at the cost of small errors. In this paper, the method of significance detection is applied to the disk data generation to realize the balance of image generation efficiency and ease of use. And we constructs the expanded data set for the disk detection based on this method.

In terms of the construction of the plate detection model, the common target detection models can be divided into single-stage model and two-stage model. Literature [13] proposed an early single-stage approach for SSD. Literature [14] and [15] are representative methods of the Yolo (You Only Look Once) series of target detection models. The multi-stage approach is mainly the work of R-CNN series. Literature [16] introduces the first R-CNN model. Fast R-CNN is introduced in literature [17]. And literature [18] designs the more optimized Faster R-CNN model. Literature [19] proposes a detection model for target segmentation. At present, Yolo series algorithms are widely used real-time target detection algorithms. End-to-end single-stage detection makes the Yolo series of algorithms achieve a good balance in accuracy and inference speed, and become one of the classical algorithms in the field of target detection. Taking RCNN series model as an example [16], the dual-stage target detection algorithm is reflected in the first stage of using the algorithm or small neural network to generate a large number of candidate boxes (region proposal). In the second stage, the model will screen and streamline the candidate boxes (bounding box regression), and determine the category of the objects in the box. The training and inference of the dual-stage target detection algorithm take a long time, but the prediction is more accurate. The single-stage target detection algorithm represented by Yolo series integrates target detection and category determination in a network, which greatly improves the model inference speed at the expense of a small amount of detection accuracy. The one used in this paper is the Yolov5 model, which achieves a good balance in detection accuracy and inference speed, and can be adapted to specific target detection tasks by artificially modifying different modules in the model. This paper introduces the attention mechanism module in Yolov5 for the disk detection, aiming to extract the global semantic information and context information of the input picture to a greater extent, adapt to the complex and multiple detection environment, and then improve the reliability and accuracy of the disk detection algorithm.

Traditional detection methods are difficult to adapt to the detection task of open set targets, and the detection model using open set can greatly reduce the work of dataset annotation. Literature [20] proposes the unsupervised detector DINO that can adapt to the detection task of open sets. In the usual annotation work, different classification standards should be established for multiple data sets, and annotators need to be familiar with the classification methods of different data sets, which undoubtedly affects the efficiency of data set construction. The CLIP model proposed in literature [21] can embed text as query input for visual tasks. The Grounding DINO introduced in literature [22] provides open detection methods ranging from natural text to target detection. With the constant change of datasets and the development of dataset scale, how to provide efficient and accurate detection methods for dataset construction tasks is equally challenging.

The main contributions of this paper can be summarized as follows:

Firstly. Aiming at the plate identification problem of automatic catering equipment, the plate detection task is proposed, and the Yolov5-based plate detection model is designed.

Secondly. The meal plate detection dataset was constructed and the detection box was annotated. The meal disk detection model performs extensive experiments on the dataset, and the results show the validity of the dataset.

Thirdly. The plate data synthesis method is proposed based on significance detection, which uses the category-independent characteristics of significance detection to extract diversified dish data from various scenarios, and effectively solves the problem of small amount of plate data and single type.

Fourthly. The disk data annotation method is proposed based on Grounding DINO. By designing effective prompt annotation, it can provide relatively high-quality labeling box level annotation, and manual annotated plate detection data sets collected in combination with actual scenarios, which effectively improves the performance of the disk detection model.

II. Disk Detection Dataset Construction

2.1 Real-world Scene Plate Detection Data Collection

Using the camera perspective of the automatic catering equipment of a domestic intelligent catering supply chain enterprise, the plate data in real restaurant scenes was collected. After manual filtering of the pictures with high degree of repetition and information that exposed personal privacy, 3500 plate data were finally selected, as shown in the figure 1 below. Later, the method of manual annotation is adopted to mark the location of the plate. Some data samples and the annotation are shown in Figure 1.

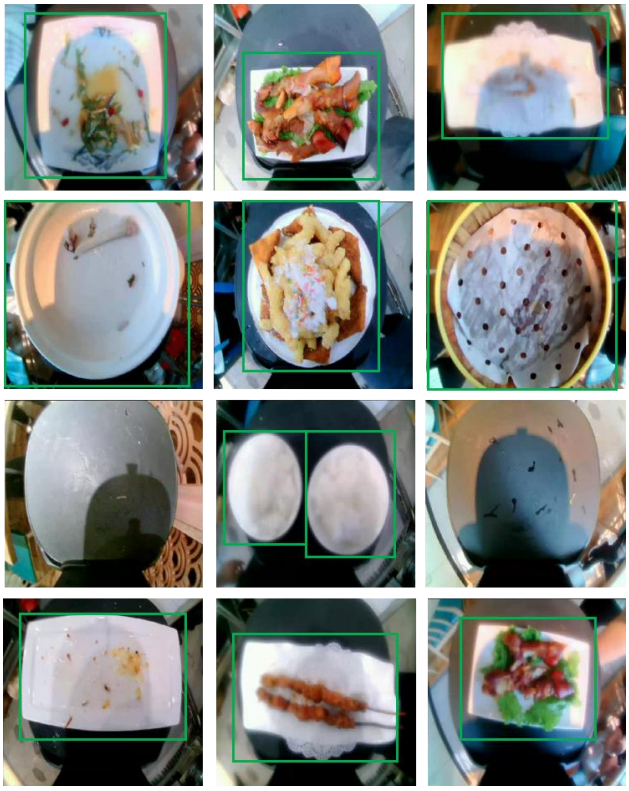

Figure 1. Examples of the meal plate detection data set

The disk detection data set constructed in this paper contains a variety of situations in the actual scene, including images taken by cameras with different resolutions, blurred shooting caused by water vapor, lighting in different restaurant environments, empty disk, multiple shadows caused by angles, large targets on the disk, multiple disk targets, etc. These situations cover a variety of difficult scenes under the camera of the actual restaurant automated catering equipment, and also increase the difficulty of plate detection. Collecting multiple scenarios to improve the generalization performance of the detection model. In the subsequent experimental section, 80% (2800 images) of the data were used as the training set for training the model, and 20% (700 images) of the data were used as the test set for testing and comparing the model performance.

2.2 Synthesis of meal plate data based on the significance detection model

2.2.1 Method of dish extraction based on the significance detection model

As shown in Figure 2, the task framework of the downstream tasks performed by the unified dish extraction method is displayed. Due to the limited types and number of dishes available in real plate scenes, in order to expand the data for training the dish detection model, this paper designs a dish detection data synthesis method based on significance target detection. The main reason is that, although there are currently data sets related to dishes, dishes and plates are not in the automated equipment trays and cannot be used to train the plate identification model proposed in this paper. Moreover, the dishes are rich and diverse, so the current semantic segmentation or instance segmentation method can not be used to extract by using fixed species. Considering the category-independent nature of the saliency target detection, the saliency detection model is used in this paper to help the model eliminate the interference from the background, leaving only information related to the dishes. Then, multiple empty plates in the plate detection data set constructed in Section 2.1 are used as the background to synthesize diverse plate data. To increase the diversity of synthetic data, data augmentation is performed using random sampling.

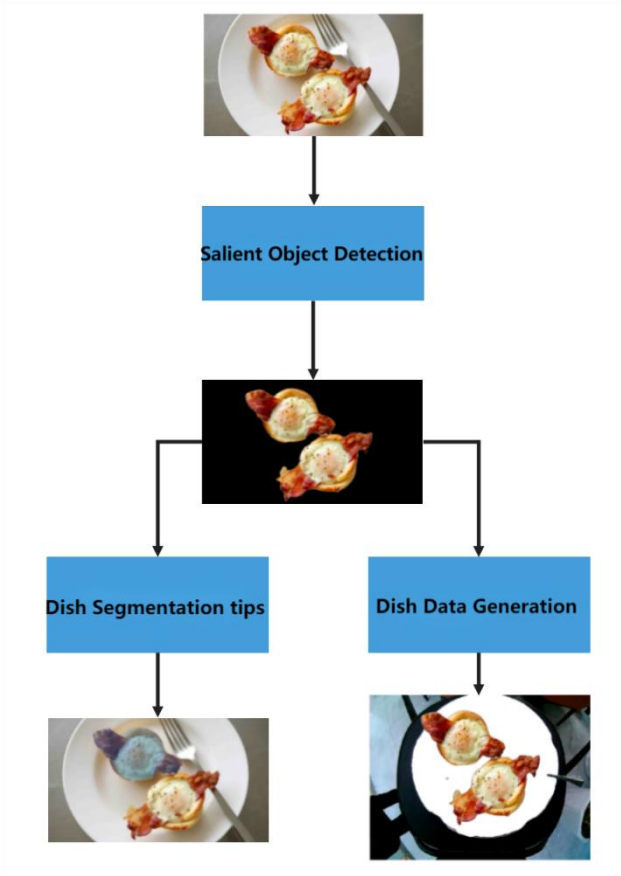

Figure 2. Dish image extraction method and downstream tasks

There are many downstream applications of dish extraction methods. One application of the downstream task of food detection is the prompt of food segmentation. In recent years, many instance segmentation and semantic segmentation algorithms use the significance detection results as the prior of image segmentation to produce more accurate results. The prior of significance detection can help the model to eliminate the interference of background and improve the effect of final detection. Dish data set generation scene to solve the difficulty of some data set construction in many business scenarios cannot get dishes samples in the real scene, need through the scene and dishes image synthesis method to obtain data sets. The method of significance detection can quickly extract the pixel area of the dish part, make it merge with the background area for

image fusion, and efficiently generate a large number of training samples with real details.

2.2.2 Significant target extraction

In terms of extracting foreground dishes, this paper selects a significance detection model using in literature [10], which applies saliency maps and salient target corresponding edges for supervision. It constructs a two-stream asymmetric network to extract salient target features and edge features respectively. The method presented here consists of a backbone encoder network, a significant target decoding network, an edge decoding network, and a fusion decoding network. Here the ResNet-50 is used as an encoder. For the remaining output of each stage, each side output is converted into 64 channels by using two stacked convolution layers. For the significant target decoding streams, global guide features can automatically adapt to each layer-scale features to guide feature fusion. Finally using a 3×3 single-layer convolution merges features as single-channel significant plots. Since the edges of a salient target contain both fine edge features and location information, the edge features are generated by using directly global guidance features to guide the shallow features, which has also greatly reduced the number of parameters. So the edge stream only contains only the deepest and shallow two-level features and finally using a 3×3 single-layer convolution merge feature as a single-channel edge graph. After obtaining the complementary features, the features are aggregated to output the final saliency figure under the guidance of the global guiding features at the fusion stage.

2.2.3 Pixel-level optimization of dense condition random fields

In this paper, the dense condition random field (denseCRF) [23] is used as the post-processing of the food image, so as to get more refined edge information. The dense conditional random field constantly corrects the local details of the image through multiple iterations of the energy field function. Its energy-field function can be expressed as follows:

$$E(z) = \sum_i \varphi_u(X, z_i) + \sum_{i<j} \varphi_p(X, z_i, z_j), \quad (1)$$

where X represents the image to be processed, the first term is the energy function related to the value of the pixel itself, and the second term calculates the category and pixel information difference between pairs of pixels. The energy field between the pairs of pixels is expressed as:

$$\varphi_p(X, z_i, z_j) = \mu(z_i, z_j) \sum_{m=1}^{k} w^{(m)} k^{(m)}(f_i, f_j), \quad (2)$$

where the kernel function k calculates the weighted similarity and difference between the different pixels. The kernel function k is given as:

$$k^{(m)}(f_i, f_j) = w^{(1)} \exp\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_\alpha^2} - \frac{\left|I_i - I_j\right|^2}{2\theta_\beta^2}\right) + w^{(1)} \exp\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_\gamma^2}\right), \quad (3)$$

where p represents the pixel location of the 2D coordinate, I indicates the pixel value, θ_α and θ_β are the control pixel distance and θ_γ is the control contrast. Using dense conditions random fields makes the resulting food image with finer edges.

2.2.4 Disk data synthesis and automatic annotation

After completing the extraction of food image, this study adopts the random sampling method based on probability distribution (such as Gaussian distribution) to realize the random change of food position in the blank plate image, so as to synthesize a new image. In this method, the specific location of dishes in the plate is controlled by the random algorithm, which simulates the diversity of dishes placed in the real scene. This process not only increases the naturalness and diversity of the synthetic images, but also provides more rich and real training data for the deep learning model, which helps to improve the recognition and generalization ability of the model for the placement of different dishes.

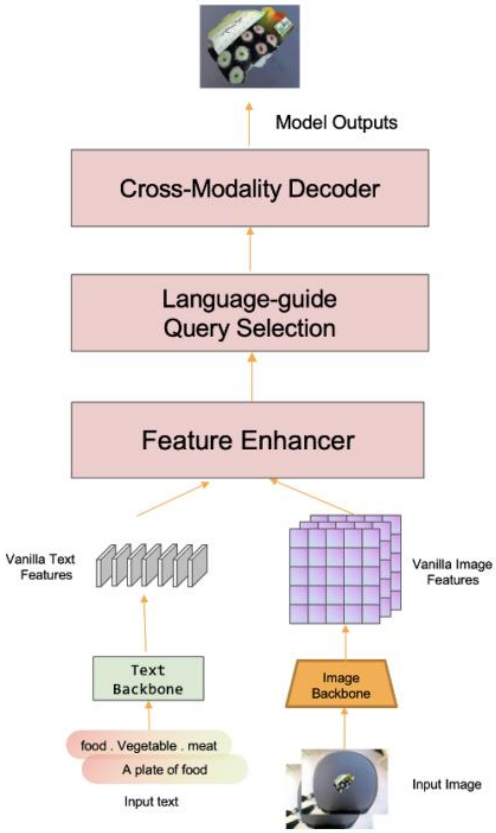

FIGURE 3. GENERATION OF AUTOMATIC ANNOTATION BOXES BASED ON THE GROUNDING DINO

As shown in Figure 3, the generation of automatic annotation box applies advanced methods such as grounding DINO to automatically box-select the items in the picture according to the text prompt. Compared with the time-consuming and laborious traditional manual annotation method, it provides a more efficient and cost-effective scheme for the generation of neural network training set. This automated process significantly improves the speed of processing large amounts of data, while significantly reducing the cost of the data preparation phase. Because automatic annotation can ensure the consistency and quality of data sets and reduce the subjective bias and inconsistency in manual annotation, this method not only enhances the

reliability of data processing, but also improves the replicability of the annotation process.

III. Disk Detection Model Construction

3.1 Target detection based on Yolo

In order to make the automatic catering equipment intelligently identify the existence of the booked dishes in the plate, and the follow-up tasks are automatically executed after the dishes are picked up. In this paper, we designed the disk detection model based on Yolov5, and introduced the attention module based on the model structure of Yolov5, which can introduce additional context information in the process of model training and reasoning, which also helps the model to pay better attention to important features and improve the detection performance.

Yolov5 is a general model widely used in target detection tasks. Its structure consists of three parts: backbone layer, neck layer and detection head. The computational module of this model contains the Conv layer, the C3 layer, and the SPPF layer. The model structure for the improved Yolov5 is shown in Figure 4. The Conv layer is the most basic module in the network structure, including the convolution layer, the normalization layer and the activation function, which work together on the input to extract more abstract semantic features. The C3 layer is used for the segmentation of intermediate features and semantic extraction at different scales. By dividing the input into two parts, one branch goes through the above Conv layer and bottleneck structure for semantic feature extraction, the other branch is processed only through the Conv layer, and then the two branch features are combined for output. The SPPF layer is used to fuse the feature map under different feelings. The input passes through the Conv layer for three maximum pooling operations, and then the output after each pooling is merged.

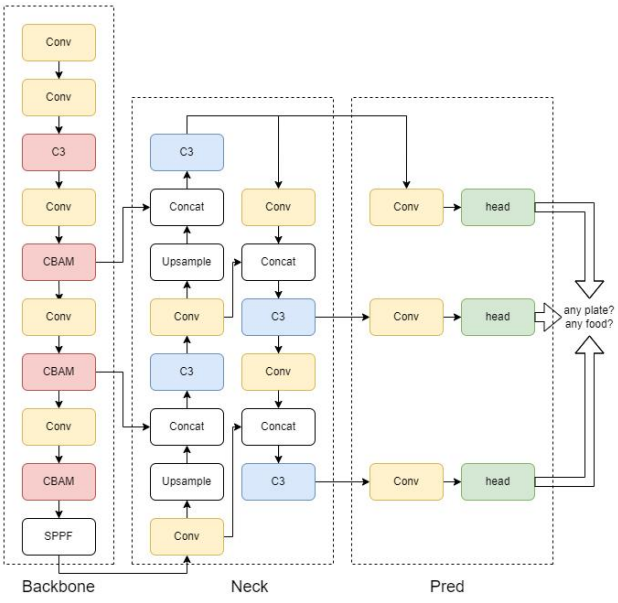

Figure 4. Improving the model structure of the Yolov5

3.2 Attention mechanism optimization of Yolo

In the traditional object detection task based on the convolution module, the model may be unable to fully use the global information in the training data to improve the detection performance due to the lack of context information guidance. Therefore, the model proposed in this paper introduces an attention mechanism based on Yolov5, which enables the model to selectively focus on specific information when processing the input data and better understand the target position, shape and context in the image.

In this paper, we tried CBAM and SE attention modules on the structural basis of Yolov5. The SE (Squeeze and Excitation) [24] module first reduces the feature graph to the channel level scale through global average pooling, and then obtains the correlation weight between different channels through two fully connected layers. Finally, the coefficient is weighted to the features of each channel, so that the model pays more attention to the channel with more information. CBAM (Channel Block Attention Module) [25] first process all channels through maximum pooling and average pooling, then the two processed feature graphs are passed through the same fully connected network, so as to obtain the correlation weight of the various channels of the different feature graphs for the current task. Finally, the two are summed to act on the various channels of the input features so that the model focuses on the information of the channels that are more relevant to the current task.

3.3 Loss function design of the target detection model

The model loss function is the weighted sum of classification loss, localization loss and confidence loss.

The classification loss measures the distance between the classification label and the true value, which is measured by the binary cross-entropy loss. The calculation formula is as follows:

$$y_i = Sigmoid(x_i) = \frac{1}{1+e^{-x_i}}, \tag{4}$$

$$L_{class} = -\sum_{n=1}^{N}[y_i^* \times \log(y_i) + (1-y_i^*)\log(1-y_i)], \tag{5}$$

where x_i represents the predicted value of the current category and y_i is the true value of the current category.

Location loss measures the direct distance between the position of the output bounding box and the real bounding box, which uses the CIoU loss calculation formula as below:

$$L_{CIoU} = CIoU(B, B_{gt}) = IoU(B, B_{gt}) - \frac{\rho^2(B, B_{gt})}{c^2} - \alpha v, \tag{6}$$

$$v = \frac{4}{\pi}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2, \tag{7}$$

$$\alpha = \frac{v}{1 - IoU(B, B_{gt}) + v}, \tag{8}$$

where v is the normalization of the difference in the length and width ratio of predictive boxes and the real boxes.

Binary cross-entropy loss is used for confidence loss:

$$p = Sigmoid(x_i) = \frac{1}{1+e^{-x_i}} \tag{9}$$

$$L_{obj} = -[y\log(p) + (1-y)\log(1-p)] \tag{10}$$

where p indicates the confidence of the current category prediction and y expresses the true label.

IV. Experimental Results and Analysis

4.1 Datasets and the experimental platform

The proposed algorithm based on Yolov5 is implemented on the Pytorch toolbox. In this experiment, 100 rounds will be trained on a Nvidia GeForce 3090, and the learning rate is set to 0.01 in the model training stage. Random gradient descent (SGD) is used to update parameters. Momentum equals 0.937, weight decay is equal to 5104, and batch size is set to 16.

4.2 Experimental setting and evaluation criteria

This experimental index will use traditional target detection indicators, including accuracy rate P, recall rate R, and average precision mAP. The accuracy rate and recall rate are calculated as follows:

$$P = \frac{TP}{TP + FP}, \tag{11}$$

$$R = \frac{TP}{TP + FN}, \tag{12}$$

where the definitions of TP, FP, and FN are shown in the following table (confusion matrix):

TABLE I. Definition of Confusion Matrix

Prediction \ Reference	Positive	Negative
Positive	True Positive (TP)	True Negative (TN)
Negative	False Positive (FP)	False Negative (FN)

The mAP will be used to obtain the weighted mean of the average precision of all categories, which can reflect the detection ability of the model on all categories. The calculation formula is as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{13}$$

$$AP_i = \frac{1}{M_i}\sum_{j=1}^{M} P_j, \tag{14}$$

where AP_i represents the average precision when detected for the *i-th* category, M_i is the total number of instances for the *i-th* category, and mAP represents the weighted mean of the average precision of all categories.

4.3 The presentation of data set and ablation experiments

4.3.1 Dish detection effect

As shown in Figure 5, the detection effect figure of our method on a certain dish data set is displayed. It is seem that the proposed method can be applied to various categories of dish identification tasks. Although the saliency detection model do not train these data, it still completes the segmentation of food pixels, and does not rely on the known classification information of the model, and does not need to retrain the new data, and exceeds the common food segmentation model in terms of detection efficiency and ease of use.

4.3.2 Effect of meal plate data synthesis

Figure 6 illustrates the dataset generated by the method presented in this paper. Although the study in this paper is only used to identify whether there is a tray on the tray, for the extenability of the process of establishing the data set.

Figure 5. Effect of significance detection of dishes

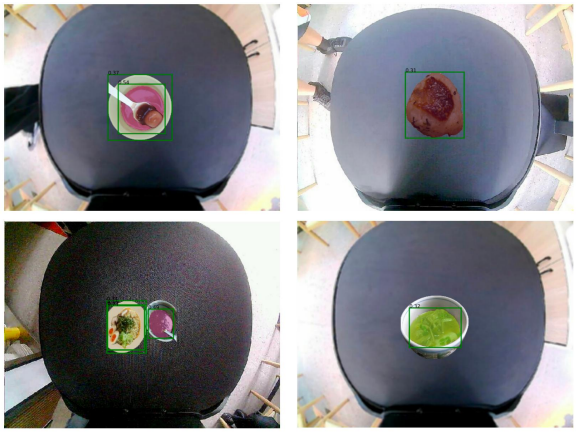

Figure 6. The exhibition of dish dataset

Although the study in this paper is only used to identify whether there is a plate on the equipment tray, but for the expansion of the subsequent data set, the tray carrier is divided into 18 categories in the process of establishing the data set, including: surplus board, bottle, drinking cup, empty tray, empty plate, empty tray, hand, kettle, and napkin, etc. However, in the study of this paper, it is only used to distinguish whether there is a dish plate on the automatic catering equipment tray. The images of detected dishes are marked as dishes.The rest are marked as empty tray, and the detection box is not output.

TABLE II. Contrast Effects of the Dataset Generation Methods

Methods	Whether a manual interaction is required	Time consuming	Whether to rely on classification
Food2k	no	<0.1s	yes
SAM	yes	>1s	no
Diffusion model	yes	>1s	no
Our method	no	<0.1s	no

As shown in Table Ⅱ, the advantages and limitations of the current mainstream data set generation method and the proposed method are demonstrated .Three different mainstream methods are compared, including the detection model proposed for large-scale dish data in [3], the SAM model proposed by [26], and the diffusion model proposed by [27] that can be synthesized close to real detail images according to text prompts. Table Ⅱ mainly evaluates the use experience of these methods in the disk data generation task from three dimensions. Whether the need for human interaction to evaluate the method and whether the need for artificial input as a prompt determines whether the corresponding task can automatically process large amounts of data. Food2k classification method can be automated the selection of the required dishes according to the category. The method of SAM and SD requires artificial input prompts. Our method can automatically separate the dishes from the background areas. In terms of time consuming, the discrimination class methods are less time-consuming, while interaction class methods usually take a long time for interaction and algorithm iteration. Except that Food2k depends on the classification of images, the other methods are not limited to the classification of images, do not require training on domain-specific data, and have better generality. According to the analysis, it can be seen that compared with some general methods in terms of ease of use and performance. According to the analysis, it can be seen that our method has great advantages in ease of use and performance by comparing with some general methods.

4.3.3 Effect of target detection

Using 2800 training images in the target detection experiment, we test 700 images (common target detection indicators) and add the results of synthetic data training. The performance data of Yolov5 + CBAM model under different data sets are compared in the following Table Ⅲ, where +grounding DINO_word indicates that the new synthetic dataset annotation is a word as the required prompt word for grounding DINO. In this paper, the model performance is evaluated separately under two augmented datasets.

TABLE III. TARGET DETECTION RESULTS OF THE WORD PROMPT GENERATION DATASET

Model (test on +grounding DINO_word)	P(%)	R(%)	mAP(%)
Yolov5+CBAM (train on original dataset)	56.3	41.8	49.5
Yolov5+CBAM (train on +grounding DINO_word)	84.2	57.7	61.2
Yolov5+CBAM (train on +grounding DINO_phrase)	96.2	71.7	85.8

Table Ⅲ presents the model performance obtained from different training sets on new datasets generated by grounding DINO with words as prompt words. Among them, the average multi-category precision (mAP) of the model trained on the original dataset is lower at 49.5%, while the average precision of the augmented model trained with words and phrases as prompt words increase, with the former increasing to 61.2% and the latter to 85.8%.

TABLE IV. TARGET DETECTION RESULTS OF THE PHRASE PROMPT GENERATION DATASET

Model (test on +grounding DINO_phrase)	P(%)	R(%)	mAP(%)
Yolov5+CBAM (train on original dataset)	63.1	31.5	44.4
Yolov5+CBAM (train on +grounding DINO_word)	87.7	54.0	59.7
Yolov5+CBAM (train on +grounding DINO_phrase)	88.4	57.9	63.5

Table Ⅳ shows that the model performance obtained from different training sets on new datasets generated by grounding DINO with phrases as prompt words. The distribution pattern of the three models on the multi-category average accuracy is consistent with the data in Table Ⅲ. The above experimental results show that the model trained only on the original dataset does not have ideal generalization performance on the augmented dataset, while the model trained on the augmented dataset has better performance, and the model trained on the +grounding DINO_phrase dataset has higher multi-category precision than the model trained on the +grounding DINO_word dataset.

In order to avoid the above test results affected by model under-fitting or overfitting. In this paper, we separately visualize the category loss, confidence loss, prediction box loss of the above three models during the respective training process, The visualization results all indicate that the model has converged during the training process. And by comparing the three sets of curves, it can be obtained that the third set of curves became smooth after 60 rounds of training, while the first two groups of curves were still significant downward trend at the 60 rounds of training. It can be concluded that the model trained under the +grounding DINO_phrase dataset not only showed the highest multi-category precision during the test phase, but also have much faster convergence rates during the training phase. See Figure 7~9 for visualization results.

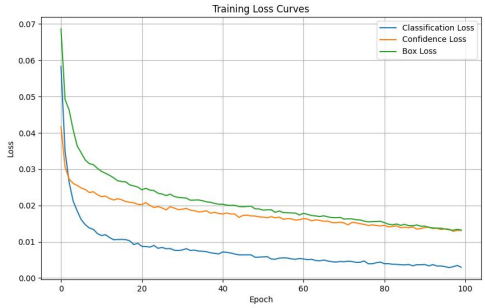

FIGURE 7. THE LOSS CURVES OF THE TRAINED MODEL UNDER THE ORIGINAL DATASET

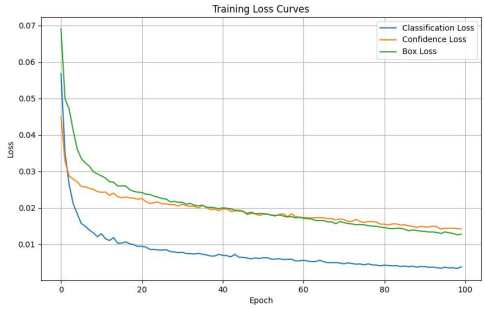

FIGURE 8. THE LOSS CURVES OF THE TRAINED MODEL UNDER +GROUNDING DINO_WORDL DATASET

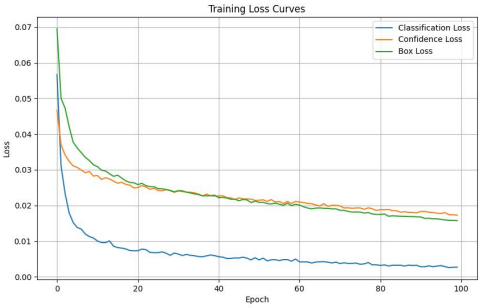

FIGURE 9. THE LOSS CURVES OF THE TRAINED MODEL UNDER +GROUNDING DINO_PHRASEL DATASET

4.3.4 Module ablation experiments of Yolov5

For the model structure itself, this paper conducted ablation experiments on the original data set on whether Yolov5 needs attention module, and the experimental results are shown in the following Table Ⅴ:

TABLE V. RESULTS OF THE ABLATION EXPERIMENTS OF YOLOV5

Model	P(%)	R(%)	mAP(%)
Yolov5	98.0	77.4	81.8
Yolov5+SE	93.7	74.3	72.4
Yolov5+CBAM	93.4	83.0	**83.7**
Yolov5+CBAM+SE	93.9	72.9	69.8

The above experimental results show that the mAP on the test set reached 83.7% when only the CBACM module was introduced in the Yolov5 model structure, which is better than the Yolov5 with an mAP of 81.8% in the test set. However, when the SE module is introduced in the model structure, the model performance decreases to different degrees. Thus the final model structure selected Yolov5 + CBAM as the plate detection model.

V. CONCLUSIONS

To meet the specific requirements of the plate detection in the smart restaurant environment, our study defines the plate detection task, and collects data from the actual restaurant scene to build the plate data set. To efficiently perform the plate detection, a Yolov5-based detection model was designed. In the face of the limitation of restaurant dishes, which leads to the uniformity of dishes in the data set, our study introduces a data synthesis method based on significance detection. By exploiting the variety of dishes data from different scenarios, it effectively solves the problems of small data amount and single type of dishes. In terms of the annotation of synthetic datasets, our method in this paper explores effective prompt word methods and achieves high-quality automatic annotation by using grounding DINO technology. The experimental results show that the plate detection model proposed in this study can effectively identify the dishes on the plate in the smart restaurant environment, and the data synthesis method based on significance detection significantly improves the quality of the data set and the performance of the plate detection model.

ACKNOWLEDGMENT

The authors are thankful to the anonymous reviewers for their valuable comments.

REFERENCES

[1] Bossard L, Guillaumin M, Van Gool L. Food-101 – mining discriminative components with random forests[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. Springer International Publishing, 2014: 446-461.

[2] Chen J, Ngo C W. Deep-based ingredient recognition for cooking recipe retrieval[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 32-41.

[3] Min W, Wang Z, Liu Y, et al. Large scale visual food recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[4] Xie J, Xiang J, Chen J, et al. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 989-998.

[5] Wei J, Wang S, Huang Q. F^3Net: fusion, feedback and focus for salient object detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12321-12328.

[6] Li G, Xie Y, Lin L. Weakly supervised salient object detection using image labels[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

[7] Wang L, Lu H, Wang Y, et al. Learning to detect salient objects with image-level supervision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 136-145.

[8] Piao Y, Wang J, Zhang M, et al. MFNet: Multi-filter directive network for weakly supervised salient object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4136-4145.

[9] Zhang J, Yu X, Li A, et al. Weakly-supervised salient object detection via scribble annotations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12546-12555.

[10] Gao S, Guo Q, Zhang W, et al. Dual-stream network based on global guidance for salient object detection[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 1495-1499.

[11] Liu Y, Wang P, Cao Y, et al. Weakly-supervised salient object detection with saliency bounding boxes[J]. IEEE Transactions on Image Processing, 2021, 30: 4423-4435.

[12] Zeng Y, Zhuge Y, Lu H, et al. Multi-source weak supervision for saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6074-6083.

[13] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 – 14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[15] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[16] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[17] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[18] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[19] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[20] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9650-9660.

[21] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[22] Liu S, Zeng Z, Ren T, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection[J]. arXiv preprint arXiv:2303.05499, 2023.

[23] Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials[J]. Advances in neural information processing systems, 2011, 24.

[24] Qi J, Liu X, Liu K, et al. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease[J]. Computers and electronics in agriculture, 2022, 194: 106780.

[25] An J, Putro M D, Priadana A, et al. Improved YOLOv5 Network with CBAM for Object Detection Vision Drone[C]//2023 IEEE International Conference on Industrial Technology (ICIT). IEEE, 2023: 1-6.

[26] Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[27] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.