# DATASET AND IMPROVED YOLOV7 FOR TEXT-BASED TRAFFIC SIGN DETECTION

Xiuyuan Chi[1], He Huang[1], Junxing Yang[1]*, Junxian Zhao[1], Xin Zhang[1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China
chi_xiuyuan@163.com, huanghe@bucea.edu.cn, yangjunzing@bucea.edu.cn, 2464745491@qq.com, zx2806404866@163.com

**KEY WORDS:** CTTSD, Improved YOLOv7, BiFormer, Prune, Traffic sign, Detection.

**ABSTRACT:**

Traffic sign detection is an important part of autonomous driving technology, and it is also important to have a large-scale dataset applicable to Chinese traffic scenarios. The article proposes a text-based self-labelled traffic sign dataset which consists of 3153 images, of which 2903 images are used for training and 250 images are used for validation. And an improved YOLOv7 algorithm is provided that incorporates the BiFormer attention mechanism into the YOLOv7 network to enhance its ability to detect small objects. This approach has the advantage of improved accuracy but may increase runtime. To mitigate this problem, the improved YOLOv7 network undergoes model pruning to compress the model size and increase its speed. Experimental results show that the improved YOLOv7 network in this paper improves the average accuracy by 2.9% while maintaining almost the same speed as the original network. After testing, the model has a real-time effect and practical significance. In conclusion, the text-based self-annotated dataset and the improved YOLOv7 network proposed in this paper have important reference values for text-based traffic sign recognition in automatic driving assistance systems.

## 1. INTRODUCTION

Autonomous driving technology is widely regarded as a technology with tremendous development prospects, as it can significantly reduce traffic accidents, improve traffic efficiency, and enhance travel comfort (Yang et al., 2018). The foundation of achieving autonomous driving lies in vehicles being able to autonomously acquire semantic information from the traffic environment. Traffic signs, as elements containing rich semantic information, provide important warnings, instructions, and prohibitions, effectively alleviating traffic congestion and reducing the occurrence of traffic accidents. Additionally, traffic sign recognition algorithms play a vital role in Advanced Driver Assistance Systems (ADAS) (Fu and Huang, 2010).

Traffic sign recognition involves two key steps: detection and classification of traffic signs. Traditional traffic sign recognition algorithms heavily rely on manually designed features. Some algorithms, such as (Broggi et al., 2007), transform the image into the RGB colour space and utilize colour features of traffic signs for detection. However, these algorithms are susceptible to interference from factors like lighting, weather conditions, and reflectivity of the sign surface. To overcome these challenges, improved algorithms have employed different approaches, including Hough transform (Escalera et al., 2007), distance transform (Gavrila and Philomin, 1999), genetic algorithm (Escalera et al., 2003), and fast radial symmetry algorithm (Keller et al., 2008), which utilize shape features of traffic signs for detection. Once a traffic sign is detected, the traffic sign recognition system proceeds to classify it and extract its semantic information. Traditional traffic sign classification algorithms include template matching algorithms (Miura et al., 2000) and support vector machine algorithms (Gil-Jimenez et al., 2007), among others. However, due to the limitations of manually designed features, these algorithms struggle to meet the increasing demands of complex traffic environments. In 2012, with the emergence of AlexNet (Krizhevsky et al., 2012), many Convolutional Neural Network (CNN)-based methods were introduced for traffic sign recognition (Zou et al., 2019). Subsequently, with the rapid development of deep learning, numerous CNN-based object recognition algorithms were incorporated into traffic sign recognition. These algorithms can be classified into single-stage recognition algorithms and two-stage recognition algorithms. Single-stage recognition algorithms, such as SSD (YOU et al., 2020) and the YOLO series (Garg et al., 2019), perform the detection and classification of traffic signs in a single feature extraction step, resulting in faster processing speeds. However, single-stage algorithms may suffer from issues like inaccurate localization, missed detections, and false positives, leading to relatively lower recognition accuracy. In contrast, two-stage recognition algorithms, such as the Faster-RCNN series (Nguyen et al., 2014), adopt a staged strategy. These algorithms first generate candidate regions and then perform feature extraction and classification on these regions. Although two-stage algorithms offer higher accuracy, their processing speed is slower due to multiple stages of computation.

For autonomous driving systems and advanced driver assistance systems, it is crucial to promptly recognize and convey the semantic information of traffic signs to the decision-making system or the driver. Therefore, traffic sign recognition algorithms need to be capable of quickly and accurately detecting small traffic signs in images. The YOLO series algorithms are renowned for their excellent real-time performance and generalization abilities, and they have been widely applied in the recognition of fine-grained instances such as traffic signs (Liu and Xiong, 2020). However, since the input images usually have high resolutions, recognition algorithms often preprocess the images by resizing them before inputting them into the recognition network. This compression of traffic sign pixels poses a significant challenge in traffic sign detection tasks. Furthermore, as the image size decreases, the details and features of the traffic signs may be lost, resulting in decreased detection accuracy. Therefore, even the YOLO series algorithms struggle to achieve outstanding levels of accuracy in traffic sign recognition.

Deep learning-based methods cannot be separated from a huge dataset, and the widely used traffic sign datasets are GTSDB, TT100K, and CCTSDB.GTSDB (German Traffic Sign Detection Benchmark) (Houben et al., 2013) is a widely practised and used dataset in the field of traffic sign detection. The dataset includes 900 images, 1206 labelled instances, and the labelling content contains "instruction", "prohibit", "warning", etc. The training set contains 600 images, and the training set contains 600 images. The training set contains 600 pictures with 846 instances, and the training set contains 300 pictures with 360 instances. The Chinese traffic sign dataset TT100K (Tsinghua-Tencent 100K) was jointly produced by Tsinghua and Tencent in 2016 (Zhu et al., 2016), which contains 100000 pictures but contains only one instance. images, but only 9176 images contain instances, and a total of 24717 instances are categorized into 128 classes for labelling, of which 6105 training set images contain 16527 instances, and 3071 training sets contain 8790 instances.2022 Changsha University of Science and Technology enriched the CCTSDB they created (Changsha University of Science and Technology Chinese Traffic Sign Detection Benchmark) dataset (Zhang et al., 2022), which contains 17856 images, the traffic signs in the image are divided into mandatory, prohibitory and warning according to their meanings. The training set is 16,356 and the test set has a total of 1,500 pictures, and the rest of the details are not yet published.

| Dataset | Picture(object) | Train(object) | Test(object) | Year |
|---|---|---|---|---|
| GTSDB | 900(1206) | 600(846) | 300(360) | 2013 |
| TT100K | 9176(24717) | 6105(16527) | 3071(8190) | 2016 |
| CCTSDB | 17856 | 16356 | 1500 | 2022 |

**Table 1.** Traffic Signs Dataset

Due to the particularity of Chinese traffic signs, foreign traffic sign detection algorithms based on datasets such as GTSDB often cannot be directly applied to Chinese traffic scenarios. To cope with the demands of autonomous driving systems and advanced driver assistance systems, it is necessary not only to accurately detect the main categories of traffic signs but also to accurately recognize the semantic information contained in traffic signs, especially text-based traffic signs. Therefore, constructing a dataset suitable for Chinese text traffic signs is crucial for algorithm development in this field.

In summary, traffic sign detection is an important component of autonomous driving technology. Currently, deep learning-based object detection algorithms have been employed to improve the accuracy and efficiency of traffic sign detection. However, the availability of datasets specifically designed for text-based traffic signs is relatively limited, posing challenges to the research and application of related algorithms. Additionally, detecting small-sized traffic signs remains a challenge. Therefore, by enhancing object detection algorithms and creating a larger-scale dataset suitable for Chinese traffic scenarios, we can facilitate the application and development of traffic sign detection algorithms in the field of autonomous driving technology. In light of these challenges, we have made the following contributions: We have created a dataset specifically for text-based traffic signs, distinct from GTSDB. This dataset, named CTTSD, is designed for Chinese text-based traffic signs and includes annotations for the text data. Additionally, we have made algorithmic improvements by integrating the BiFormer attention mechanism into YOLOv7. This enhancement aims to improve the accuracy of small object detection.

## 2. DATASET

In this section, we introduce the issues about the dataset, including the source of the dataset, how it is labelled, and how the final dataset is divided.

### 2.1 Data Collection

In our research, we aim to address the recognition and understanding of Chinese text-based traffic signs. To support this goal, we have created CTTSD (Chinese Text-based Traffic Sign Dataset), which is specifically designed for Chinese text-based traffic signs. The data for CTTSD is primarily sourced from the CCTSDB dataset. CCTSDB is a publicly available dataset that contains a large number of traffic sign images. However, this dataset only provides annotations for the types of traffic signs and does not include separate annotations for text information. Therefore, to meet our research needs, we have decided to enhance the CCTSDB dataset.

### 2.2 Data Annotation

We randomly selected 3153 images from the CCTSDB dataset as the base data for CTTSD. To ensure the diversity and challenge of the dataset, we specifically chose images with different types of traffic signs, including speed limits, prohibitions, warnings, and directions. This ensures that the dataset contains a variety of traffic signs with different shapes, colours, and styles. During the annotation process, we used the Labelimg annotation software to carefully annotate all traffic signs that contain text. Our annotation work focuses primarily on capturing the Chinese text information present in the signs. For each annotation box, we labelled the category as "text" to facilitate text detection and extraction tasks. We made efforts to ensure the quality and accuracy of the annotations in CTTSD. We paid particular attention to challenging factors present in the images, such as occlusions, blurriness, lighting variations, and tilting. This was done to ensure that the dataset reflects the complexity and diversity of real-world traffic signs, thereby enhancing the robustness and performance of models in practical applications. We hope that CTTSD will serve as a valuable resource for research on Chinese text-based traffic sign recognition and understanding. By utilizing this dataset, researchers can undertake various tasks such as text detection, text recognition, traffic sign classification, and more.

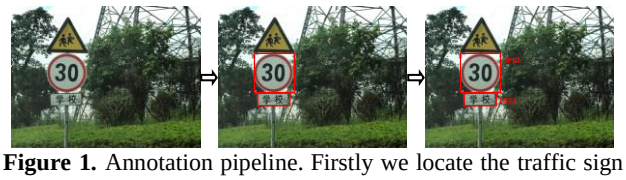


**Figure 1.** Annotation pipeline. Firstly we locate the traffic sign and draw its bounding box. Then the class label is attached.

(a) (b)

**Figure 2**. Labelling instructions. Like the image in (a), the part in the box is not labelled, and although we know that it is text sign, we cannot identify the text on the sign, which does not help us in our subsequent experiments, so that part is discarded. In addition, like in picture (b), the large sign has small sign nested in it, and all of them come with text. For complex sign like this, we only label the largest one outside, and the one inside is not labelled separately.

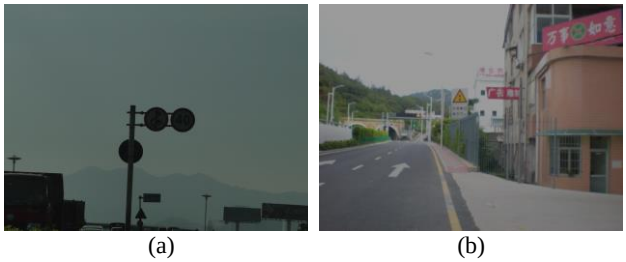
### 2.3 Dataset Statistics

The CTTSD dataset is a Chinese text traffic sign dataset containing 3253 images. Among them, 1885 images contain labelled instances, totalling 2781 labelled instances. For training and evaluation, we divide the dataset into a training set and a test set in a ratio of roughly 11:1. The training set contains 2903 images while the test set contains 250 images.

| Name | Image | Train | Test | Instance |
|---|---|---|---|---|
| CTTSD | 3153(1885) | 2903 | 250 | 2781 |

**Table 2.** CCTSD dataset. The CCTSD dataset contains a total of 3153 images, out of which a total of 1885 contain labelled instances, out of which a total of 2781 instances are labelled. The dataset is divided into a training dataset and a test dataset, the training dataset contains 2903 images while the test dataset contains 250 images.

The CTTSD dataset is diverse and authentic. We pay special attention to collecting challenging image samples, which include images in various situations such as blurry, cloudy, rainy and occluded. This diversity reflects the variety of environments and conditions that may be encountered in real-world scenarios in traffic, providing challenges to the robustness and generalization ability of the model. During the labelling process of the dataset, we strictly followed standard label specifications and accuracy requirements. Each label instance is annotated with proofs to ensure the accuracy of the text region and matched with the corresponding traffic sign. The CTTSD dataset was created to provide a valuable resource for the research and application of traffic sign-related tasks in Chinese text. By using this dataset, researchers can perform training and evaluation on tasks such as text detection and text recognition, thereby promoting research progress in related fields.



(a) (b)


(c) (d)

**Figure 3**. CTTSD Overview. Included in our dataset are overcast data in (a), blurry images in (b), obscured images in (c), and rainy conditions in (d).

## 3. METHOD

In this experiment, our main goal is to improve the detection accuracy and speed of text-based traffic signs, especially for small objects. To achieve this goal, we choose to conduct research based on the improved YOLOv7 algorithm (Wang et al., 2023), and introduce the BiFormer (Zhu et al., 2023) attention mechanism to further improve the accuracy of small object detection. The YOLOv7 algorithm was chosen as the basic algorithm because the algorithm is widely used and mature in the field of target detection, with fast detection speed and good performance. YOLOv7 is an improved version of the YOLO series. Compared with the previous version, it has been optimized in detail and has higher stability. The one-stage detection method is adopted, the detection speed is fast, and the real-time requirements are met. To improve the accuracy of small object detection, we introduce the BiFormer attention mechanism. Small objects usually have a small size and low signal-to-noise ratio, which puts higher requirements on the visual perception ability of the model. The BiFormer attention mechanism combines textual information and visual features and improves the recognition ability of small objects by fusing the two. In this way, the model can better focus on the details of small objects, thereby further improving the detection accuracy. The BiFormer attention mechanism is introduced into the YOLOv7 algorithm to make the model pay more attention to text information during the detection process. By fusing textual and visual features, the model can localize and recognize text-based traffic signs more accurately, especially for small-sized objects, and can better distinguish subtle differences between objects and backgrounds. However, introducing the BiFormer attention mechanism will increase the parameter amount of the model, which in turn affects the detection speed of the model. To solve this problem, we employ a model pruning algorithm. The application of the model pruning algorithm aims to reduce redundant parameters in the model, reduce the complexity of the model, and thus improve the inference speed of the model. Combining the above improvements, we expect to achieve better performance in the text-based traffic sign detection task.

### 3.1 YOLOv7

The detection process of the YOLOv7 network is divided into four classic parts, namely the input (Input) terminal, the backbone (Backbone) network, the neck (Neck) network and the detection head (head) module. On the input side, adaptive image padding and Mosaic data augmentation techniques are used. Adaptive image filling can perform dynamic filling operations according to the size of the input image so that the network can handle image inputs of different sizes. Mosaic data enhancement technology randomly arranges four images into one image by randomly cropping, rotating, scaling and other

operations, thereby enhancing the diversity and generalization ability of the dataset. The main function of the backbone network is to extract the feature information in the image. It consists of CBS, ELAN and MP modules. The CBS module consists of a convolutional layer, a normalization layer, and an activation function layer, and is mainly responsible for feature extraction and channel transformation. ELAN is a feature extraction module composed of multiple convolution modules. By controlling the longest gradient path, the network can learn more features. The MP module performs downsampling through pooling and convolution operations with a step size of 2, thereby reducing the size of the feature map and improving computational efficiency. The neck network part contains PANFPN, SSPCSPC, ELAN-W and UpSample structures. PANFPN integrates the precise location information at the bottom layer with the abstract semantic information at the high level, so that the semantic information and positioning information in different layers are fully integrated, and the positioning accuracy of the model for multi-size objects is further improved. The SSPCSPC structure is a pyramidal feature pooling structure. It further enhances the network's perception of targets of different scales through the fusion of feature pooling at different scales. ELAN-W is a feature extraction module, which is specially used to extract deep semantic features of the network. The UpSample structure is used for upsampling operations to restore the size of the feature map to the original input size for subsequent detection. The detection head network part performs object detection by meshing the extracted features to achieve precise positioning and classification of objects in the image.

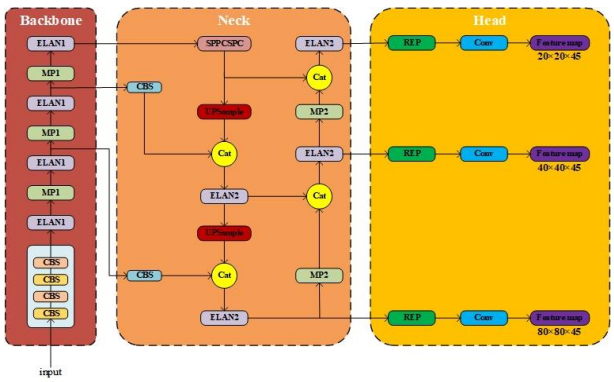



**Figure 4**. The network structure of YOLOv7, the image data input into the network after size processing, is sent to the Backbone stage, through a series of convolution operations to extract features, the extracted features are sent to the Neck stage, through some pooling layer as well as the upsampling layer, extracted to the network of the deep semantic features, and finally after the Head part of the features for detection or classification.

### 3.2 BiFormer

The BiFormer attention mechanism is a new visual Transformer model that combines two-layer routing attention and BRA (Bi-directional Routing Attention). Its core principle is to calculate attention weights by iteratively transferring information to achieve dynamic and query-adaptive sparsity, thereby improving the efficiency and accuracy of the attention mechanism. Compared with other attention mechanisms, the BiFormer attention mechanism has Dynamic and adaptable, as well as the comprehensiveness of two-way transmission, so it can handle complex feature relationships with a lower amount of calculation.

The two-layer routing attention mechanism is an important part of BiFormer, which updates the attention weights through multiple iterations. At each iteration step, attention weights are computed by comparing query and key similarities, and these weights are applied to values to compute a weighted sum. Through multiple iterations, the two-layer routing attention can adaptively adjust the attention weights according to the needs of the query to better capture the dependencies and contextual information among key features. BRA is a bidirectional routing attention mechanism, which plays an important role in BiFormer. BRA utilizes region-to-region routing of directed graphs to bidirectionally shift attention weights from query to key and from key to query, which can capture the correlation between features more comprehensively. By shifting the attention weights in different directions, BRA can enable the attention mechanism to understand and represent the features of the input more comprehensively. BRA also uses a token attention mechanism, which helps the model understand images better.

To implement the token-to-token attention mechanism, it is first necessary to gather tensors of key and value.

$$K^g = gather(K, I^T), V^g = gather(V, I^T) \tag{1}$$

where Kg is the tensor of the aggregated key, and Vg is the tensor of the aggregated value.

Then, use the attention operation on the aggregated K-V pairs.

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{2}$$

where LCE(V) is a local context enhancement.

Therefore, the BiFormer attention mechanism achieves the ability to process complex feature relationships and contextual information in visual tasks by combining dual-layer routing attention and BRA. This mechanism can dynamically adjust attention weights, improve the expressiveness and flexibility of the model, and comprehensively capture the correlation between features. By introducing the BiFormer attention mechanism, the model can better understand and represent the features of the input, thereby improving the performance and generalization ability of the model in vision tasks.

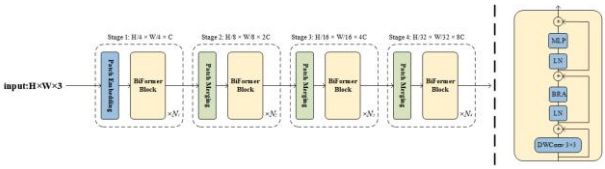



**Figure 5**. Left: The overall architecture of BiFormer. Right: Details of a BiFormer Block. In the *i* stage, the input spatial resolution is reduced using overlapping patch embedding (when $i = 1$) or using the patch merging (when $i = 2,3,4$) module while increasing the number of channels. The input features are then transformed using the BiFormer block that uses $N_i$ connected. Three BiFormers of different sizes were then instantiated by scaling the network width (the number of basic channels C) and depth ( the number of BiFormer blocks used in each stage $N_i$, $i = 1,2,3,4$ ).

### 3.3 Improved YOLOv7

To make YOLOv7 achieve better performance in text-based traffic sign detection tasks, improve the accuracy of small target detection, and enhance feature expression and generalization capabilities. We add the BiFormer attention mechanism to the backbone of YOLOv7.

Incorporating the BiFormer attention mechanism into the backbone of YOLOv7 brings multiple benefits. First, the backbone network is responsible for extracting features from the input image in YOLOv7. By introducing the BiFormer attention mechanism in the backbone part, the model's perception of key features and contextual information in the image can be enhanced. The BiFormer attention mechanism can adaptively adjust the attention weight according to the image content, and better focus on small objects and important features, thereby improving the accuracy of small object detection. Second, the choice of adding BiFormer to the backbone part is based on the important role of the backbone network in feature extraction. The backbone network is at the front end of the model and has a decisive impact on feature extraction. By introducing the BiFormer attention mechanism in the backbone part, the attention mechanism can be used to model the relationship between the target and the background in the early stage of feature extraction, which helps to extract more discriminative feature representations. This early attention mechanism enables the model to better distinguish objects from backgrounds and improve object detection accuracy. In addition, the backbone part is one of the most critical components in YOLOv7, which plays a key role in the performance and speed of the entire detection model. By adding the BiFormer attention mechanism in the backbone part, the spatial and semantic information in the image can be fully utilized, and the representation ability of the object can be enhanced during the feature extraction process. This helps to improve the generalization ability and robustness of the model while reducing the dependence on other stages.

Therefore, adding the BiFormer attention mechanism to the backbone of YOLOv7 can improve the detection accuracy of the model for small targets, and introduce an attention mechanism in the feature extraction stage to effectively use the key features and context information in the image. This design choice can improve the performance of object detection and provide more accurate feature representation for subsequent stages of processing, leading to better detection results.

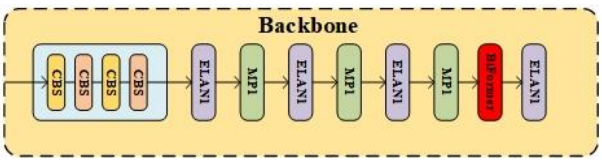



**Figure 6**. YOLOv7 with BiFormer added.

### 3.4 Model Pruning

Model pruning is a commonly used optimization technique, which aims to reduce the number of parameters of the model, reduce storage requirements, and improve the inference efficiency of the model. Usually divided into parameter, channel, structure, iteration and other pruning methods. This paper adopts the L1 pruning method in parameter pruning.

L1 pruning cuts out parameters with small or unimportant weights by evaluating the important indicators of the parameters, such as the absolute value of the parameters (L1 norm). Specifically, L1 pruning zeros or removes parameters with smaller weights, thereby achieving model sparsity. This pruning operation can effectively reduce redundant parameters in the model, thereby reducing the storage space requirement of the model and improving the reasoning efficiency of the model. In the process of model pruning, it is first necessary to evaluate the model and calculate the importance index of each parameter. By ranking the importance of parameters, it is possible to determine which parameters should be pruned. A pruned model needs to be retested and fine-tuned to maintain its performance and accuracy. The advantage of using the L1 pruning method is that the importance of parameters can be accurately measured, and the calculation efficiency is high. Through the pruning operation, we can significantly reduce the number of parameters of the model, thereby reducing the storage requirements of the model and saving storage space. In addition, the pruned model also has faster inference speed, because the calculation amount is correspondingly reduced after removing redundant parameters. Ultimately, the pruned and fine-tuned model maintains high performance across retests. Model pruning can not only reduce the storage overhead and computational load of the model but also improve the efficiency of the model in the inference stage. Therefore, using the L1 pruning method to optimize the model can effectively improve the efficiency and practicability of the improved model.

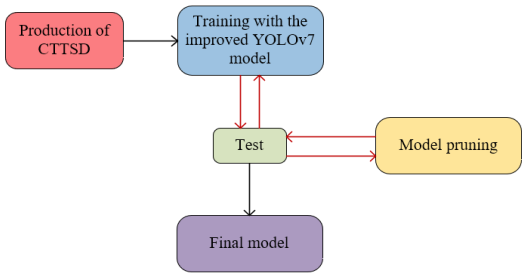



**Figure 7**. Overall figure. First, make the CTTSD dataset, after the production is completed, use the dataset to train the improved YOLOv7 model, after the training is completed, we get the model we need and then test it, after the test is completed, model pruning is performed, after model pruning the model should be fine-tuned one or more times for training to ensure the model accuracy, and finally get the model in the experiments of this paper.

## 4. EXPERIMENTAL RESULTS AND ANALYSIS

### 4.1 Experimental Environment

The training and testing datasets in this paper choose the CTTSD dataset, because this dataset is specially labelled for text-based traffic signs, and has higher applicability compared with other datasets. The focus of this study is on the detection of text-based traffic signs, while other datasets do not label this data in detail, so they do not meet the needs of this study. Since the CTTSD dataset meets the requirements of this experiment, it is selected as the training and testing dataset.

In the training phase, this article uses Ubuntu 22.04 as the operating system, and it is equipped with NVIDIA Corporation GA102GL [RTX A5000] graphics processor. The Python version is 3.9.1, the Torch 1.8.1 framework is used, and the training is performed under the cuda 11.1 accelerated environment. In the testing phase, this article uses Windows 10 as the operating system and uses a GeForce GTX 1080 graphics processor. The Python version is 3.7, using the torch 1.10.1 framework, and tested in the cuda 10.2 environment.

Through the selection of the above operating system, graphics processor, Python version and acceleration environment, this paper can be trained and tested on different platforms, ensuring the reliability and reproducibility of the experimental results.

### 4.2 Parameterization and Evaluation Indicators

For experiments, we set the input image size to the default 480×480 pixels. In the training process, the Adam optimizer is used as the optimization algorithm, the learning rate is set to 0.012, and the momentum is set to 0.937. To adjust the learning rate, a cosine annealing algorithm is used. The size of each batch is 8, and the training time is set to 1000 epochs. In terms of evaluating detection accuracy, this paper uses mean Average Precision (mAP) as the main evaluation index. Specifically, we set the IoU (Intersection over Union) threshold to be greater than 0.5, that is, mAP@0.5. The mAP indicator is the result of averaging the average precision (Average Precision, AP) of all target categories. By calculating the mAP value, we can understand the detection performance of the model on the entire dataset. The specific mAP calculation method is as follows:

$$mAP = \frac{1}{c}\sum_{(i=1)}^{c} AP_i \tag{3}$$

Among them, $c$ represents the number of detection target categories, and $i$ is the category index. The AP value is the area enclosed by the curve drawn by the precision rate (Precision) and the recall rate (Recall) between 0 and 1 and the coordinate axis. It can comprehensively measure the precision rate and recall rate of a certain class. The definitions of precision and recall are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Among them, $TP$ is the number of true cases, $FP$ is the number of false positive cases, and $FN$ is the number of false negative cases. The precision rate can reflect the proportion of the true examples among the positive examples predicted by the model, and the recall rate reflects the proportion of the positive examples correctly predicted by the model to the total positive examples. In terms of detection speed, the parameters of the model (Params) and the number of frames processed per second (Frames Per Second, FPS) are used for evaluation.

### 4.3 Ablation Experiment

To verify the effectiveness of the improved model proposed in this paper, we conducted multiple sets of ablation experiments on the CTTSD dataset, and the experimental results are summarized in Table 3. Through ablation experiments, we can evaluate the contribution of each improvement and determine its importance to the model performance. Table 3 shows the results of the ablation experiments for each group. We eliminated or modified specific model components or algorithms in each set of experiments, and recorded the corresponding performance metrics, such as mAP@0.5. By comparing the experimental results, we can analyze the impact of each improvement on the performance of the model.

| Group | BiFormer | Prune | Params(M) | Precision(%) | Recall(%) | mAP@0.5(%) | FPS |
|---|---|---|---|---|---|---|---|
| **1** | \ | \ | 36.48 | 0.915 | 0.857 | 0.902 | 66 |
| **2** | √ | \ | 56.83 | 0.948 | 0.833 | 0.931 | 54 |
| **3** | √ | √ | 41.02 | 0.946 | 0.835 | 0.931 | 64 |

**Table 3**. Ablation experiments. Group1 is the original algorithm, Group2 is the algorithm after adding the BiFormer attention mechanism, and Group3 is the result of the model after adding the BiFormer attention mechanism and pruning.

" √ " in the table indicates that this part of the improvement is included, while "\" indicates that it is not included. According to the data in the table, it can be observed that in the experiment without the BiFormer attention mechanism, the number of parameters of the model is 36.48M, the precision rate is 0.915, the recall rate is 0.857, the average precision is 0.902, and the real-time performance of 66 FPS is achieved. This is in line with the consistent approach of the YOLOv7 algorithm, which is to improve real-time performance while ensuring a certain accuracy. However, as can be seen from the second row of data, when the BiFormer attention mechanism is integrated, the precision rate and average precision increase by 3.3% and 2.9%, respectively. However, this is also accompanied by an increase in the number of parameters and a decrease in real-time performance, with FPS dropping from 66 to 54. The increased accuracy does not offset the time cost very well. To solve this problem, this experiment carried out parameter pruning on the fused model, hoping to reduce the number of parameters and improve FPS in this way. According to the last row of data in the table, it can be seen that after pruning, the number of parameters of the model is reduced to 41.02M, the FPS is increased to 64, and the average accuracy remains unchanged. Based on all the data, the method proposed in this paper is effective. Based on ensuring real-time performance, the average precision is increased by 2.9% and the accuracy rate is increased by 3.1%. The experimental results show that the improved algorithm can correctly detect text-based traffic signs in the data inference stage (the data used in the inference is randomly selected from the TT100K dataset, and has not participated in any training and testing stages). In the case of interference with similar signs such as symbolic traffic signs and other billboard texts, it can still be detected correctly, and there are no false detections or missed detections.

(a) (b)


(c) (d)

**Figure 8**. Results of Improved YOLOv7. (a) A visual overall demonstration of the functionality of this algorithm, which detects only text-based traffic sign. (b) Used to demonstrate that the detection of small targets is effective. (c) To show that there is no false detection when there are symbol-based traffic signs that are very similar to text-based traffic signs. (d) To show that the presence of many similar text boxes such as billboards, cabs, etc., will not be missed or misdirected.

**Figure 9**. YOLOv7 before improved and after comparison. And as derived from the experimental results, this improvement in accuracy is especially reflected in the detection of small targets. In the left figure, the result of inference using the original algorithm did not detect the target (in the circle), and in the right figure, our improved algorithm can be seen to have successfully detected the target.

To feel the detection effect of the improved model on small targets more intuitively, we used a heat map to visualize it. Obviously, the improved model (right picture) will pay more attention to the small model and have higher accuracy.


**Figure 10.** Heatmap comparison. The left image shows the unimproved algorithm, which does not recognize the target and incorrectly attends to other places, and the right image shows the improved YOLOv7 algorithm, which can be seen to detect the target and does not incorrectly attend to the background.

To verify that the pruned model reduces the number of parameters and improves inference speed while maintaining accuracy, we conduct further tests and visualize the results. The visualization results clearly show the performance of the model before and after pruning. By comparing the left and right images, it can be seen that the models before and after pruning have no significant difference in detection accuracy. This shows that the pruning operation did not negatively affect the accuracy of the model and successfully maintained the accuracy of the model.


**Figure 11.** Before and after pruning comparison. The left figure shows the unimproved algorithm and the right figure shows the improved YOLOv7 algorithm, both of which have the same effect.

## 5. CONCLUSION

Aiming at the relatively low availability of text-based traffic sign datasets and the fact that they do not meet the unique characteristics of traffic signs in China, this paper creates and annotates a Chinese-focused text-based traffic sign dataset called CTTSD. This dataset covers images under actual conditions such as occlusion, blurring, and different weather conditions. It is in line with actual application scenarios and can be used to test various target detection and text detection algorithms. Aiming at the difficulties existing in the detection of small target images, this paper proposes a small target detection algorithm that integrates the BiFormer attention mechanism based on the YOLOv7 algorithm and applies model pruning technology. By integrating the BiFormer attention mechanism into the backbone part, the feature extraction ability for small targets is enhanced, thereby improving the detection accuracy; then the model is pruned to reduce the number of parameters and further improve the detection speed. The experimental results show that the algorithm proposed in this paper is better than the current best YOLOv7 algorithm in the detection of text-based traffic signs. The algorithm significantly improves detection accuracy while maintaining high-speed inference. After model pruning, the amount of parameters is reduced, and the inference speed is further improved, reaching a real-time performance of 64 frames per second.

In summary, the CTTSD dataset and the improved YOLOv7 algorithm proposed in this paper have important reference values for the detection of text-based traffic signs. They can effectively improve detection accuracy and maintain the advantages of high-speed reasoning, providing a reliable solution for traffic sign detection in autonomous driving technology.

## ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (Grant Numbers 42201483), the China Postdoctoral Science Foundation (Grant Numbers 2022M710332) and Research on Orthophoto Generation Method Based on Multi-source Data Fusion(X23001).

## REFERENCES

A. Broggi, P. Cerri, P. Medici, P. P. Porta and G. Ghisio, "Real Time Road Signs Recognition," 2007 IEEE Intelligent Vehicles Symposium, Istanbul, Turkey, 2007, pp. 981-986, doi: 10.1109/IVS.2007.4290244.

C. G. Keller, C. Sprunk, C. Bahlmann, J. Giebel and G. Baratoff, "Real-time recognition of U.S. speed signs," 2008 IEEE Intelligent Vehicles Symposium, Eindhoven, Netherlands, 2008, pp. 518-523, doi: 10.1109/IVS.2008.4621282.

D. M. Gavrila and V. Philomin, "Real-time object detection for "smart" vehicles," Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 1999, pp. 87-93 vol.1, doi: 10.1109/ICCV.1999.791202.

Escalera S, Radeva P, Pujol O. Traffic sign classification using error correcting techniques[J]. VISAPP (2), 2007, 2007: 281-285.

ESCALERA A， ARMINGOL J M A， MATA M. Traffic sign recog- nition and analysis for intelligent vehicles［J］. Image and Vision Computing，2003，21（3）：247-258

Garg P, Chowdhury D R, More V N. Traffic sign recognition and classification using YOLOv2, faster RCNN and SSD[C]//2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2019: 1-5.

J. Miura, T. Kanda and Y. Shirai, "An active vision system for real-time traffic sign recognition," ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.00TH8493), Dearborn, MI, USA, 2000, pp. 52-57, doi: 10.1109/ITSC.2000.881017.

Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

Liu X F, Xiong F. A real-time traffic sign detection model based on improved yolov3[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2020, 787(1): 012034.

M. -Y. Fu and Y. -S. Huang, "A survey of traffic sign recognition," 2010 International Conference on Wavelet Analysis and Pattern Recognition, Qingdao, China, 2010, pp. 119-124, doi: 10.1109/ICWAPR.2010.5576425.

Nguyen B T, Ryong S J, Kyu K J. Fast traffic sign detection under challenging conditions[C]//2014 International Conference on Audio, Language and Image Processing. IEEE, 2014: 749-752.

P. Gil-Jimenez, H. Gomez-Moreno, P. Siegmann, S. Lafuente-Arroyo and S. Maldonado-Bascon, "Traffic sign shape classification based on Support Vector Machines and the FFT of the signature of blobs," 2007 IEEE Intelligent Vehicles.

S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," The 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 2013, pp. 1-8, doi: 10.1109/IJCNN.2013.6706807.

Symposium, Istanbul, Turkey, 2007, pp. 375-380, doi: 10.1109/IVS.2007.4290143.

Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7464-7475.

Yang, D., Jiang, K., Zhao, D. et al. Intelligent and connected vehicles: Current status and future perspectives. Sci. China Technol. Sci. 61, 1446–1471 (2018). https://doi.org/10.1007/s11431-017-9338-1

YOU S， BI Q， JI Y， et al. Traffic sign detection method based on improved SSD ［J］. Information，2020，11（10）：475

Zhu Z, Liang D, Zhang S, et al. Traffic-sign detection and classification in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2110-2118.

Zhu L, Wang X, Ke Z, et al. BiFormer: Vision Transformer with Bi-Level Routing Attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10323-10333.

Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey. arXiv[J]. arXiv preprint arXiv:1905.05055, 2019, 16.