**Frontiers in Computing and Intelligent Systems**
ISSN: 2832-6024 | Vol. 7, No. 2, 2024

# Research on the Detection of Traffic Flow based on Video Images

**Jian He, Wei Teng *, Zeyu Zhao, Binche Liu, Bing Qin and Jun Jiang**

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan Liaoning, 114051, China
*** Corresponding author:** Wei Teng

**Abstract:** Based on the current level of social development, everyone's demand for cars has increased rapidly. At present, the total number of motor vehicles and drivers in China ranks first in the world. With the rapid development of deep learning, the method of vehicle flow statistics based on video can directly use the existing traffic monitoring camera to realize the detection of vehicles, and some traffic flow detection based on YOLOv1, YOLOv2, YOLOv3, YOLOv4 and other algorithms have problems such as insufficient accuracy and low efficiency. Therefore, this paper proposes to use YOLOv5 to replace the original algorithm to achieve object detection, tracking, and processing. I improve the efficiency of the statistics of the traffic flow.

**Keywords:** Video-based Traffic Statistics; YOLOv5; Object Detection; Tracking; Processing.

## 1. Introduction

The total number of motor vehicles and drivers in China ranks first in the world, and the detection of traffic flow is a crucial study in urban traffic management and planning. With the increase of urban population and the continuous growth of car ownership, the problem of traffic congestion is becoming increasingly prominent, which brings great challenges to the sustainable development of cities. Understanding and mastering traffic flow can provide important data support for traffic management departments, helping to optimize decision-making in road planning, traffic light control, traffic congestion mitigation, and more. Traffic flow detection plays an important role in urban traffic management and planning, which can provide accurate data support for traffic management departments, help optimize traffic conditions, improve road traffic efficiency, and promote the sustainable development of cities. However, many algorithms have serious problems such as low detection accuracy and low detection efficiency. To better improve the accuracy and efficiency, we use the YOLOv5 algorithm to improve it to achieve real-time monitoring [1].

## 2. Introduction of YOLOv (1,2,3,4) and Their Shortcomings

The Hinton team designed AlexNe by using a convolutional neural network, which works better on the ImageNet dataset than other traditional methods. So far, convolutional neural networks have gradually become popular in the field of computer vision, and more algorithms for image detection and processing have been proposed.

YOLOv1 is the first version of YOLO, presented in 2015. It takes an input size of 448x448 and divides the image into 7x7 grids, each predicting two bounding boxes, each containing a target. Each bounding box also predicts the category probability of the target. YOLOv1 was released by Joseph et al., which not only has a high degree of generalization ability, but also can meet the requirements of real-time detection. However, because YOLOv1 does not use a full convolutional network, the size of the input is fixed, and the image often needs to be cropped, which has a worse prediction effect on small targets in large images, and also has a certain impact on comprehensive object recognition. FCN network structure is divided into two parts: full convolution part and deconvolution part. The full convolution part is some classic CNN networks (such as VGG, ResNet, etc.), which is used to extract features; the deconvolution part is to get the original size of the semantic segmentation image through up-sampling. The input of FCN can be a color image of any size, and the output size is the same as the input size, and the number of channels is n (number of target categories) + 1 (background). The FCN network structure is as follows:

In order to improve the problem of YOLOv1, YOLOv2 was proposed in 2016. YOLOv2 uses the Darknet-19 network as the backbone network and introduces Anchor Boxes to better handle targets at different scales. In addition, YOLOv2 uses multi-scale training and testing to improve the detection ability of small targets. Compared with YOLOv1, YOLOv2 has improved in terms of accuracy and speed. However, due to the small number of grids, the positioning accuracy for large-scale targets is not high.

YOLOv3 was introduced in 2018 and is the largest version of the YOLO series. YOLOv3 uses a deeper Darknet-53 backbone network and adds detection at multiple scales. YOLOv3 also introduces three target prediction layers of different sizes for detecting targets of different sizes. In addition, YOLOv3 also uses FPN (Feature Pyramid Network) to extract multi-scale features. FPN is a top-down feature fusion method, but FPN is a multi-scale object detection algorithm, that is, there is not only one feature prediction layer. Although some algorithms also adopt multi-scale feature fusion to carry out target detection, they often only use the features of one scale obtained after fusion. Although this approach can combine the semantic information of top-level features with the details of low-level features, it will cause some deviations in the process of feature deconvolution. The prediction using only the features obtained after fusion will have a bad effect on the detection accuracy. Based on the above problems, FPN method can make prediction on multiple fusion features of different scales to maximize the detection accuracy.

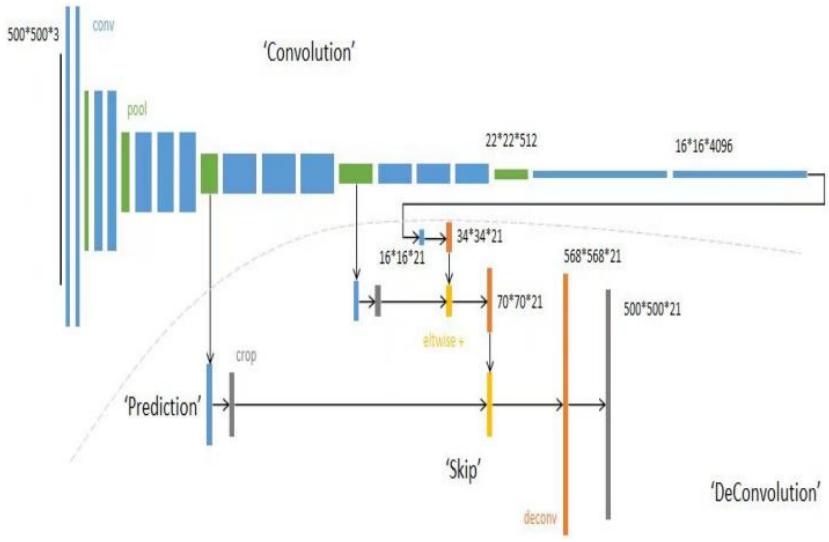


**Figure 1.** The FCN network structure

YOLOv4 was proposed in 2020. YOLOv4 has made a series of improvements in network structure and training strategy, including using CSPDarknet53 as the backbone network, using the SPP (Spatial Pyramid Pooling) module to process multi-scale information, and using PANet to fuse different hierarchical features. The network structure of PANet is shown in Figure 2, which consists of five core modules. Among them, (a) is an FPN[2], (b) is a bottom-up feature fusion layer added by PAN, (c) is an adaptive feature pooling layer, (d) is a bounding box prediction head of PANet, and (e) is a fully connected fusion layer for predicting masks.YOLOv4 has been significantly improved in terms of accuracy and speed. However, there is still the problem of poor handling of overlapping targets, identifying them as one target or missing some of them.

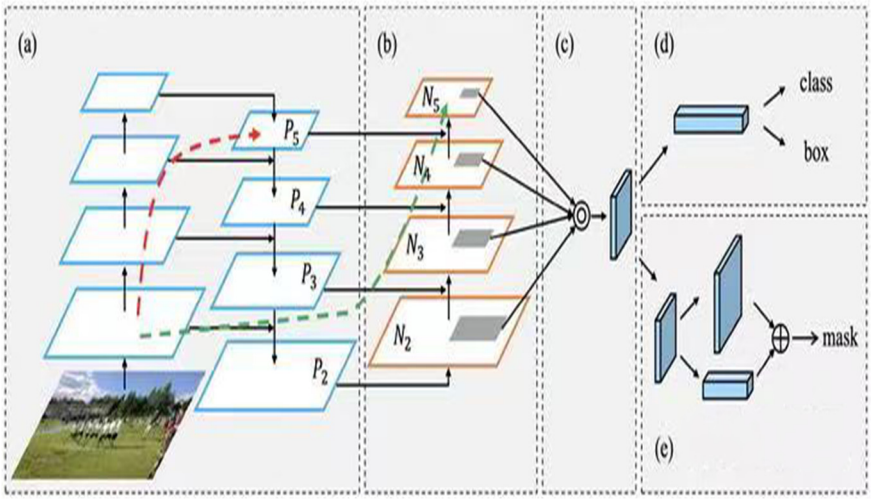


**Figure 2.** Network Structure of Panet

## 3. Introduction of YOLOv5

In order to make the tracking, detection, and processing of images better realized, and the efficiency of traffic flow statistics is greatly improved, we use the YOLOv5 algorithm to implement it, which is compared with other YOLO algorithms, YOLOv5 uses CSPNet as the backbone network. CSPNet is a new network structure, which can effectively reduce the redundancy of feature maps and improve the computational efficiency and accuracy, using feature maps of four scales, which are 40x40, 80x80, 160x160, and 320x320. (CSPNet structure by splitting the shallow feature map in two at the channel dimension, a portion propagates backwards via a feature extraction module such as a residual block, the other part is directly merged with the output of the feature extraction module through the cross-stage hierarchy. A richer gradient combination is achieved, and can be reduced on the basis of the same or improved accuracy 10% ~ 20% of network parameters.） This made it possible to increase the ability to detect small targets, and at the same time, a variable input size of 640x640 was used. This allows for adaptation to different scenarios and needs. And there are 876 million parameters. YOLOv5 increased the number of parameters to improve the expressiveness and generalization ability of the model, and the mAP and FPS of YOLOv5 on the COCO dataset were 36.2% and 140.0. YOLOv5 is superior to the rest in terms of accuracy and speed. Therefore, we use YOLOv5 to improve the accuracy and efficiency of traffic flow measurement[2].

YOLOv5 is a pre-trained object detection architecture and model series on the COCO dataset, which is an extension of the YOLO series. neck and head, among them, Backbone is responsible for feature extraction, Neck is responsible for feature fusion, and Head contains three detection heads,

which are responsible for outputting detection information, yolov5 has modified the four parts of the yolov4 network, and has achieved great improvements, using Mosaic data enhancement, adaptive anchor frame calculation, and adaptive image scaling on the input side. CBS module consists of a convolution layer Conv, a batch normalization layer BatchNorm and SiLu activation function layer. The specific structure diagram is shown below.

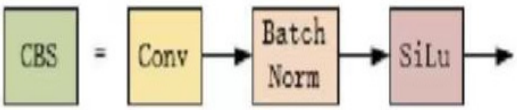


**Figure 3.** CBS module structure diagram

CBS module can extract the features in the image and sort out the extracted feature map. Conv of the CONvolution layer in CBS module changes the multiple of the feature map reduction by modifying the step size, and the image filling operation is calculated automatically by the model. In Backbone, the step size of Conv of CONvolution layer is 2, and the size of convolution kernel is 3 X3. Through this convolution layer, the width and height of feature map will be halved, and feature extraction of image is carried out at the same time. In Neck, convolution check feature maps with a size of 1X1 are mostly used to raise or reduce dimension. The BatchNorm layer is added after the convolution layer to normalize the extracted features, so that the network will not be unstable due to large data before the activation function, and to prevent the gradient explosion, disappearance and overfitting problems. At the very end, the SiLu activation function is used, adding smooth and non-monotonic features to increase the nonlinearity of the data.

Focus structure is used on the backbone side. (Focus structure is a special layer in YOLOv5, mainly used to improve the detection ability of the model on small targets. This design can make the model better focus on the features of small targets, thus improving the accuracy of target detection and recall rate. Especially for small target detection, the Focus structure can bring obvious advantages. In addition, the Focus structure can reduce the computational effort and memory footprint by breaking the input feature graph into multiple subgraphs and concatenating the channel dimensions instead of directly copying the spatial dimensions of the entire feature graph. Such operations can improve the overall performance of the model by reducing the number of channels in the feature graph while maintaining critical information) and CSP structures (CSP structures improve the performance of deep neural networks by dividing input features into two parts and then cross-connecting between the two parts. CSP structure can effectively improve the feature representation ability of the model, thus improving the accuracy and generalization ability of the model. FPN+PAN structure was added to the neck side, the loss function during training was improved on the head side, and GIOU_Loss was used to improve DIOU_nms for prediction box screening. [3]. Compared with the fixed size and proportion used in the traditional method, the adaptive anchor frame has many advantages, such as enriching the detection of the object's background and small target, and calculating the Batch As it uses Normalization, it computations data on four images at a time, so that the mini-batch size doesn't have to be as large as it uses a single GPU to get good results.

Adaptive anchor frame calculation

In both YOLOv3 and YOLOv4, anchor needs to be calculated by K-means clustering method in advance, and the anchor is fixed. However, in YOLOv5, although the anchor is set in advance, the optimal anchor in different training sets can be calculated adaptively during training, so as to update the anchor value. This feature can also be turned off manually, changed in train.py, set to False.

The specific implementation structure of YOLOv5 is as follows [4]:

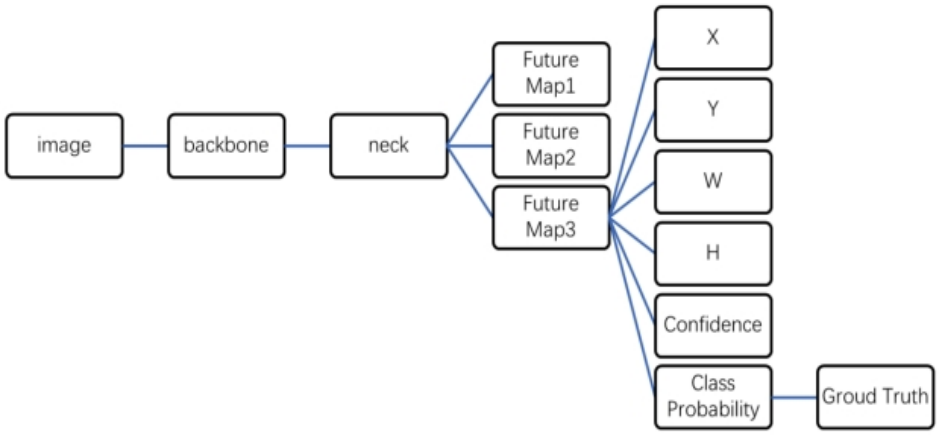


**Figure 4.** YOLOv5 structure diagram

## 4. Algorithm Design

The YOLOv5 algorithm is used to realize object detection, image segmentation, and target tracking to realize the statistics of traffic flow.

There are five versions of YOLOv5, namely YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l and YOLO5x. The network depth, width, parameter size and number of convolutional nuclei in different stages are different

1.YOLOv5s: This is the smallest model in the YOLOv5 series. "s" stands for "small". The model performs best on devices with limited computing resources, such as mobile devices or edge devices. YOLOv5s has the fastest detection, but the accuracy is relatively low.

2.YOLOv5m: This is a medium sized model in the YOLOv5 series. "m" stands for "medium". The YOLOv5m offers a good balance between speed and accuracy and is suitable for devices with a certain amount of computing power.

3.YOLOv5l: This is a larger model in the YOLOv5 series. "l" stands for "large". YOLOv5l is relatively accurate, but the

detection speed is slower. It is suitable for devices that require high accuracy and have strong computing power.

4.YOLOv5x: This is the largest model in the YOLOv5 series. "x" stands for "extra large." YOLOv5x performed best in terms of accuracy, but had the slowest detection speed. Ideal for tasks requiring extremely high accuracy and devices with powerful computing power, such as Gpus.

5.YOLOv5n: This is a variant in the YOLOv5 family, optimized for Nano devices such as the NVIDIA Jetson Nano. YOLOv5n delivers accuracy for edge devices while maintaining high speeds.

We used yolov 5s in this experiment.

First, we built a vehicle model and improved it by using the YOLOv5 network, which effectively alleviated the problem of target loss and target location. YOLOv5 is used to predict the position and category of the target in the image, and the prediction results are processed by YOLOv5. In the process of target tracking and detection, we conduct target tracking by studying the process of target positioning moving target in video over time. In the target tracking task in the real scene, considering the real-time and flexibility of the algorithm, we intend to adopt the online tracking method based on detection. Multi-target tracking technology (MOT) is designed to detect and obtain the position of the target that is interested or wants to track in the video frame, and assign an ID to each target, and keep each target ID unchanged during the process of target movement. After consulting a lot of data, we were able to reconstruct the feature extraction network of YOLOv5 using the improved EfficientNetv2, and the association fusion network solved the competition problem between multi-task learning in JDE algorithm, greatly reducing the number of parameters and calculation amount of multi-target tracking algorithm, and better meeting the real-time requirements of mobile devices. At the same time, a fast multi-object based algorithm based on semi-suppressed fuzzy clustering proposed by Zhang Junhui and others is used to greatly improve the real-time and reliability of multi-object data association in complex environments.

For image segmentation, a given image data is treated as a collection of many fuzzy elements belonging to different clusters. The membership degree of each pixel is blurred to each cluster, and the image segmentation algorithm based on OTSU-IFCM is used. The algorithm first utilizes the OTSU algorithm (Otsu algorithm is an adaptive threshold segmentation algorithm based on image gray histogram. It determines an optimal threshold by maximizing the inter-class variance and minimizing the intra-class variance, which realizes the purpose of automatic image segmentation. The basic principle of Otsu algorithm is to select the optimal threshold by the maximum inter-class variance. In the process of image segmentation, we want the difference between the two categories after segmentation to be as large as possible, that is, to maximize the inter-class variance. Otsu algorithm calculates the inter-class variance under different thresholds and selects the threshold with the largest inter-class variance as the best threshold. To find the optimal segmentation threshold of image foreground and background, an image segmentation algorithm based on OTSU-IFCM is used. The algorithm firstly uses Otsu algorithm (Otsu algorithm is an adaptive threshold segmentation algorithm based on image gray histogram. It determines an optimal threshold by maximizing the inter-class variance and minimizing the intra-class variance, which realizes the purpose of automatic image segmentation. The basic principle of Otsu algorithm is to select the optimal threshold by the maximum inter-class variance. In the process of image segmentation, we want the difference between the two categories after segmentation to be as large as possible, that is, to maximize the inter-class variance. Otsu algorithm calculates the inter-class variance under different thresholds and selects the threshold with the largest inter-class variance as the best threshold. The optimal segmentation threshold of foreground and background of image is found. By studying the positioning process of moving targets in video over time, the detection-based online tracking method is adopted in the target tracking task of real scene, considering the real-time and flexibility of the algorithm.

Then the IFCM algorithm is used, which is a variant of the intuitive fuzzy C-means clustering algorithm. The basic idea is to treat a given image data as a set of fuzzy elements belonging to different clusters. By blurring the membership degree of each pixel to each cluster, and then iteratively adjusting the cluster center and membership degree according to the update formula, the image segmentation is finally realized.

Multi-target tracking technology (MOT) is designed to detect and obtain the position of the target that is interested or wants to track in the video frame, and assign an ID to each target, and keep each target ID unchanged during the process of target movement. After consulting a lot of data, we were able to reconstruct the feature extraction network of YOLOv5 using the improved EfficientNetv2, and the association fusion network solved the competition problem between multi-task learning in JDE algorithm, greatly reducing the number of parameters and calculation amount of multi-target tracking algorithm, and better meeting the real-time requirements of mobile devices. At the same time, a fast multi-object basis algorithm based on semi-suppressed fuzzy clustering proposed by Zhang Junhui et al is used to greatly improve the real-time and reliability of multi-object data association in complex environments. The specific realization flow chart is as follows:

The specific implementation flow chart is as follows:

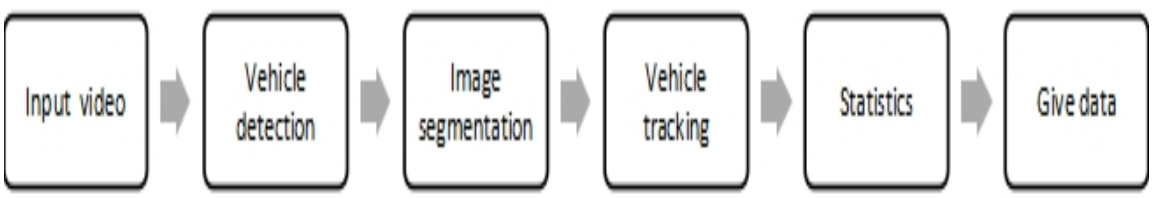


**Figure 5.** Implementation flowchart

## 5. Experimental Results

In the area of target detection, there are four types of samples: true positive(TP), false positive(FP), true positive(TP), false positive(FP), in order to quantify the performance comparison of the proposed network with other networks, this paper uses typical evaluation metrics such as precision (P) and recall (R), the formula is as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\%$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\%$$

The average precision (AP) is the area below the P-R curve, which is the average of the Precision values of the P-R curve. For P-R curve, a definite integral is used for the calculation:

$$\mathrm{mAP} = \int_0^1 p(r)\mathrm{d}r$$

In the actual calculation of target detection performance test, P-R curve should be smoothed. For each point on the P-R curve, the Precision value is calculated as follows:

$$P_{\mathrm{smooth}}(r) = \max_{r' \geqslant r} P(r')$$

MAP is an average value of a plurality of classes mAP. Because the method in this paper only detects pavement objects, the meaning of mAP and AP in this paper is the same. MAP is the most important index to measure the performance of target detection.

In this experiment, we collected a lot of videos on the network and took some videos for testing from different angles, and a large number of videos were recognized well under the model and algorithm, and the accuracy was very high. The accuracy and efficiency of traffic flow statistics are fully improved, and the algorithm mentioned in this paper can have a good detection effect on traffic flow statistics. The YOLOv5 algorithm can be applied to a large number of object detection projects to improve the accuracy and efficiency of the project.

## 6. Conclusion

In this paper, the YOLOV5 algorithm is used to improve the accuracy and efficiency of traffic flow statistics, and improve the accuracy of actual detection. In the future, we will consider adapting the model algorithm to the traffic camera to truly achieve the statistics of the actual application of traffic flow.

## Acknowledgments

College Students' Innovation and Entrepreneurship Training Plan of University of Science and Technology Liaoning in 2024.

## References

[1] Zhang Yingwei. Research on traffic flow and violation detection technology in intelligent transportation[D].Liaoning University of Science and Technology, 2023.DOI: 10. 26923/ d. cnki.gasgc.2023.000474.

[2] Research and application of multi-target tracking algorithm based on YOLOv5 and DeepSORT[D]. WANG Jialin Shandong University, 2021.

[3] Liu Bei. Research on traffic flow detection algorithm based on object detection and tracking[D]. Inner Mongolia: Inner Mongolia University,2022.

[4] Editorial Department of Chinese Journal of Highway and Transportation. Review of China Automotive Engineering Research Progress in 2017. China soumnal of Highway and Transpont,2017, 30(6):1-197. Editorial Department of Chinese Journal of Highway Engineering, Review of Chinese Automotive Engineering Academic Research, 2017 [J]. Chinese Journal of Highway, 2017,30 (6): 1-197.

[5] Lin Shu, Liu Mingying, Tao Zhiying, Underwater Treasure Detection Based on Attention Mechanism and Improved YOLOv5 y!.Transactions of the chinese society of Agricutural Engineerng, 2021, 37(18):307-314.