Deep Learning-Based Pedestrian Detection and Analysis with YOLOv5

Xuchen Cui

Guangdong University of Technology, Guangdong, 510006, China
3120002648@mail2.gdut.edu.cn

Abstract. Fueled by the swift advancements in artificial intelligence, computer vision technology has found extensive applications across various domains. This article will focus on how to use the You Only Look Once version 5 (YOLOv5) to enhance the accuracy and efficiency of pedestrian detection. It begins by providing an overview of the development background and current research status of pedestrian detection, which can as method take into scene study. Subsequently, detailed insights into the network architecture of YOLOv5 and its constituent structures are presented. Finally, the study employs YOLOv5 to train the model and comprehensively analyzes and discusses the training outcomes. The experimentation is conducted using the roboflow dataset. The findings demonstrate that YOLOv5 exhibits robust performance in pedestrian identification across diverse scenarios, showcasing its effectiveness even in complex environments and occlusion scenarios. This research contributes to advancing the capabilities of pedestrian detection, thereby enhancing security monitoring and intelligent transportation systems.

Keywords: Pedestrian Detection, YOLOv5, Computer Vision.

1 Introduction

Pedestrian detection mainly uses computer vision and deep learning technology to identify whether there are pedestrians in a picture or video frame, and if there are pedestrians, it can mark out and provide location information. This technology has been widely used in recent years in applications such as assisted driving and intelligent video surveillance, which can distinguish pedestrians in different situations in the real environment. Because there are many different characteristics among human beings, and there are many other factors such as occlusion and Angle of view, there are many differences in the reality of pedestrians, which makes pedestrian detection a difficult point in computer vision research [1].

Pedestrian detection technology has developed from the traditional human-assisted feature detection to the feature detection based on deep learning. In early pedestrian detection techniques, the only means of identifying people within an image was to manually extract features from the window and subsequently input those features into a classifier for recognition. The traditional features are Haar-like (Haar), Histogram of

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_85

Oriented Gradient (HOG), Harris and so on [2, 3]. Ddlal proposed a method to detect pedestrians through the edge features extracted by HOG. Because this method needs to calculate through the HOG of the manually extracted image feature area, it takes a lot of time and energy, so the accuracy and efficiency are relatively low [2]. The effectiveness and precision of pedestrian detection algorithms have reached new heights through convolutional neural networks because of the growth of deep learning. Compared with traditional pedestrian detection, modern pedestrian detection based on deep learning can independently extract target features from pictures or videos for learning. Currently, contemporary research in object detection technology predominantly emphasizes one-stage and two-stage detection methodologies. The You Only Look Once (YOLO) method has emerged as a prominent area of investigation within this technological domain [4-7]. Single-stage Target Detection (SSD) algorithm, which can directly output the location of the target and only need one detection to get the output result [8]. The SSD algorithm is fast and efficient, but its accuracy is a little insufficient. The two-phase target detection algorithm is exemplified by the regional convolutional neural network, where the candidate region of the target serves as the reference point. This approach involves extracting the candidate region of the target, followed by feature extraction using Convolutional Neural Network (CNN) [9], and ultimately conducting the detection process. Although the speed is not as fast as the phase target detection algorithm, the accuracy is improved.

Initially, You Only Look Once version 5 (YOLOv5) robust feature extraction capability is harnessed to effectively identify pedestrians across diverse scenarios. Subsequently, meticulous optimization of model architecture and hyperparameters is undertaken to enhance both detection speed and accuracy. A thorough evaluation process is then conducted to scrutinize and juxtapose the performance of model against alternative methodologies. Additionally, data augmentation techniques are employed. Specifically, Mosaic data augmentation is applied at the input end of YOLOv5, randomly amalgamating four images through random scaling and distribution. This augmentation significantly enriches the detection dataset, particularly augmenting small targets, thereby fortifying the network's robustness. Experimental findings underscore the critical role of the learning rate, which dictates the weight update step size during training, and the necessity for an appropriate learning rate to expedite training and mitigate overfitting. Furthermore, the choice of anchor profoundly influences model training reverse updates. Given that YOLOv5's anchor is based on the coco dataset, adjustments to the anchor frame are imperative when utilizing other datasets.

The significance of this research lies in the pivotal role of pedestrian detection in domains such as autonomous driving and smart home technology, where it holds immense potential for enhancing traffic safety and public security. Ultimately, this advancement promises to furnish individuals with greater convenience and safety in their daily lives.

2 Methodology

2.1 Dataset Description and Preprocessing

The dataset employed in this endeavor is designated as the Human Detection Data Set, which is an open-source data set on roboflow [10]. The dataset covers a variety of different scenarios and environments, including city streets, indoor Spaces, crowd areas, and more. This diversity ensures that the model can accurately detect and track pedestrians in a variety of situations. The dataset contains images of pedestrians viewed from different angles and distances, as well as pedestrians with different states. It is helpful to improve the adaptability of the model to various pedestrians.

Image preprocessing was not carried out in this project, because YOLOv5 comes with a process of image preprocessing. The pre-processing of YOLOv5 is usually to scale the image in equal proportions first, then center it, and fill the letterbox with the excess. The sample is shown in the Fig. 1.

Fig. 1. The example of human detection data set.

2.2 Proposed Approach

YOLOv5 uses CNN as the basic framework. YOLOv5 is structured with four fundamental layers, namely the input layer, backbone layer, bottleneck layer, and header layer. Initially, the image data undergoes preprocessing within the input layer, followed by the extraction of essential feature information from the image through the backbone layer. The neck network undergoes additional fusion and enhancement of characteristics, while the header network is tasked with producing the prediction outcomes for target detection. Fig. 2 shows the architectural design of YOLOv5.

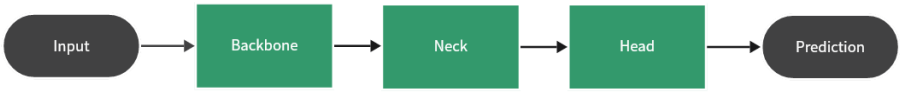

Fig. 2. The architectural design of YOLOv5.

Input. At the input side, the image is preprocessed. In Yolov5, the input data is enhanced by Mosaic technology. The coordinate of image stitching reference point (x, y) is randomly selected, and four pictures are randomly selected. After sizing and scaling based on reference points, the four images are positioned in the four quadrants of the map of a predetermined size. The scaling details of each image are mapped to its image label. Then, using the specified horizontal and vertical coordinates, these images are pieced together into a complete large image. Any coordinates outside the boundary will be adjusted to ensure the image is properly aligned. This approach to data enhancement greatly increases data diversity.

Backbone. The Focus module and Cross Stage Partial Network (CSPNet) are used to form the Backbone. In the YOLOv5 model, the Focus module located between the image input and Backbone processes the image first. It performs a downsampling operation by extracting pixels from the original image at a certain interval to obtain four subimages of similar size that complement each other. Thanks to its unique sampling method, the process retains all original image information and avoids any data loss. With this processing, the information is retained in the channel dimension, and the number of input channels is increased from the original 3 to 12. A convolution operation is then performed on this extended channel image, resulting in a feature map that is effectively double downsampled without any information loss. This process not only ensures information integrity, but also improves the computing efficiency of the network. However, it is not supported and unfriendly to some devices, which is very expensive, and the model will crash if the slices are not properly matched. In object detection problems, using CSPNet as Backbone brings great improvement. Fig. 3 shows the structural components of the Backbone module.

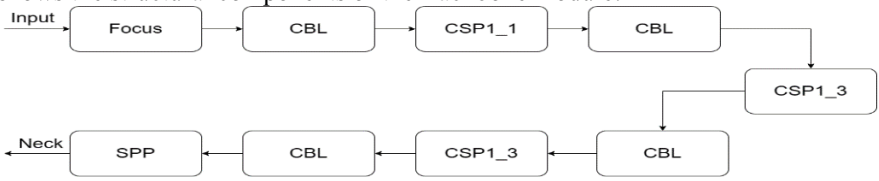

Fig. 3. The structural components of the backbone module.

Neck. The neck module is a crucial component within target detection algorithms, serving to integrate feature maps from various levels in order to generate feature maps that are abundant in multi-scale information. This process ultimately contributes to improving the precision of target detection. In YOLOv5, Feature Pyramid Network (FPN) is adopted as the feature pyramid structure. Through up-sampling and down-

sampling operations, this structure realizes the fusion of feature mappings at different levels, and then forms a multi-scale feature pyramid.

The upper portion of the model primarily focuses on integrating features across various levels by employing up-sampling and merging coarse-grained feature maps. Initially, the feature map from the final layer is up-sampled to generate a more intricate feature map. Subsequently, the processed feature map from the preceding layer is combined with the feature map from the previous layer to enhance the feature expression. Conversely, the lower segment of the model utilizes convolution layers to merge feature maps from different levels. Initially, the feature mapping from the lowest level undergoes convolution to enhance the feature representation. The resulting convolution feature map is then merged with the feature map from the preceding layer to further enrich the feature representation. This iterative process continues until the highest level is attained. Fig. 4 shows the structural components of the Neck module.

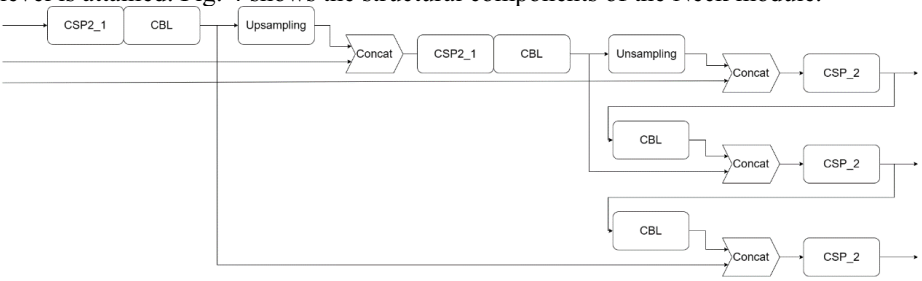

Fig. 4. The structural components of the Neck module.

Loss Function. The bounding box loss in the YOLOv5 model is calculated through the utilization of the Complete Intersection over Union (CIOU), which serves as an enhanced version of the Conventional IOU loss function. The standard IOU metric measures the degree of overlap between two boxes, with values ranging from 0 (no overlap) to 1 (complete overlap). In target detection, IOU helps identify positive and negative samples and is used to evaluate the accuracy of prediction boxes. The CIOU loss function not only considers the intersection area of the bounding box, but also adds a penalty factor that considers the disparity in dimensions between the anticipated bounding box and the actual bounding box and the distance between the center point. This improved loss function design aims to improve the model's understanding of the position and shape of the boundary box, thereby enhancing the model's positioning accuracy. Here is the formula for the CIOU loss function:

$$CIOU\_\text{Loss}=1\text{-CIOU}=(1-(\text{IOU}-\frac{Distance\_2^2}{Distance\_c^2}-\frac{V^2}{1-IOU+v})) \tag{1}$$

The symbol C denotes the smallest enclosing rectangle, with Distance_c indicating the diagonal length of this rectangle. Distance_2 signifies the Euclidean distance between two central points. The IOU metric evaluates the degree of overlap between the predicted bounding boxes and the ground truth bounding boxes. Parameter V is employed to quantify the aspect ratio of the box in terms of its length to width ratio.

3 Results and Discussion

This chapter will analyze and discuss the results of training. In this training, YOLOv5 is used for 100 rounds of training, and a series of results such as confusion matrix, Precision-confidence curve and training detection results were obtained. These results will be analyzed in detail.

3.1 Confusion Matrix Analysis

The confusion matrix is a tabular frequently employed in machine learning classification tasks. It shows how each category predicted by the model compares to the actual category. Fig. 5 shows the confusion matrix obtained after this training. Table 1 shows the structure of the confusion matrix.

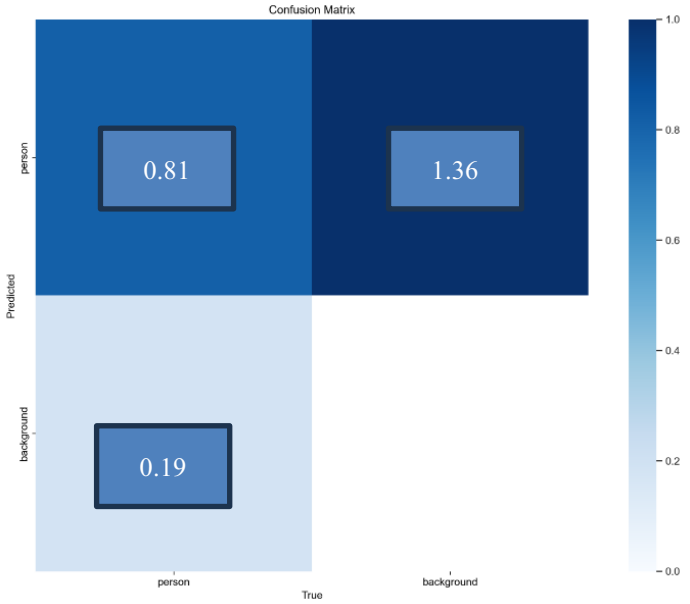

Fig. 5. Confusion matrix.

Table 1. The structure of the confusion matrix.

Confusion Matrix		TRUE	
		Negative	Positive
Prediction	Negative	TN	FN
	Positive	FP	TP

True positive (TP) means that the predicted result and the actual data point uniformly to the positive class, i.e. when the sample is positive, it is correctly predicted to be positive. True negative (TN) means that both the predicted result and the actual data show that the sample is a negative class, i.e. when the sample represents a nega-

tive class, it is accurately classified as such. False positive (FP) is when a sample that is actually a negative class is incorrectly predicted to be a positive class, i.e. the prediction shows it to be a positive class when in fact it is a negative class. False negative (FN) describe the situation where a sample is incorrectly predicted to be negative when it should be positive, i.e., the prediction results classify the sample as negative despite the fact that it belongs to the positive class. Based on the four fundamental values of TP, TN, FP, and FN, different performance measures like accuracy, precision, recall, and F1 score are calculable. The formulas for these metrics are as follows.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$Pr\,e\,cision = \frac{TP}{TP+FP} \tag{3}$$

$$Re\,c\,all = \frac{TP}{TP+FN} \tag{4}$$

$$F1\ score = 2 * \frac{Pr\,ecision * Re\,call}{Pr\,ecision + Re\,call} \tag{5}$$

Accuracy is defined as the proportion of accurately predicted samples to the total number of samples, usually expressed as a percentage. Nevertheless, accuracy may not offer a precise evaluation in cases where there is an unequal distribution of classes. Precision refers to the accuracy in identifying positive samples, i.e. the proportion of samples that are actually positive out of all samples that are identified as positive by the model. This metric shows the model's ability to reduce false positives. Recall assesses the model's ability to identify all truly positive samples. F1 score is a comprehensive index that can reflect the overall effectiveness of the model.

According to the confusion matrix, it can be known that since Yolov5 pedestrian detection only detects pedestrians and does not detect other objects, it can be seen that there is no counterexample that is correctly predicted. So, in particular, the confusion matrix has a TP of 0.

Based on the confusion matrix provided, the parameters just mentioned can be calculated. However, according to these parameters, it can be found that some do not achieve good results. And even the results of different parameters seem to contradict each other. The accuracy, precision, recall, and F1 score were determined to be 0.405, 0.448, 0.81, and 0.577. Notably, the accuracy, precision, and F1 score were observed to be relatively low. This is due to reason. Since there is no counterexample of correct judgment, that is, no objects other than people are detected, but sometimes some objects in the background will be misjudged as adults, resulting in the column where True is the background, all samples are gathered into FP. As a result, when calculating precision and accuracy, the denominator will be too large and the result will be too low, thus affecting the value of F1 score. This shows that none of these three values are suitable for evaluating the model. On the contrary, the recall metric remains unaffected by variations in FP and TN, as it is not contingent on these parameters in

its computation. Consequently, recall provides an accurate representation of the model's performance. The relatively high recall value of 0.81 suggests that the model effectively identifies pedestrians and exhibits robust recognition capabilities for positive samples.

3.2 Precision-Confidence Curve

The Precision-Confidence Curve (PCC) serves as a commonly utilized visualization tool for illustrating the efficacy of object detection algorithms. By showing the precision corresponding to different confidence levels, this curve helps evaluate the performance of the detector and AIDS in setting the appropriate decision threshold. In PCC, the x-axis represents confidence levels while the y-axis represents precision values. By analyzing the curve shape and position of the detector, its performance and reliability can be evaluated. Fig. 6 shows the PCC.

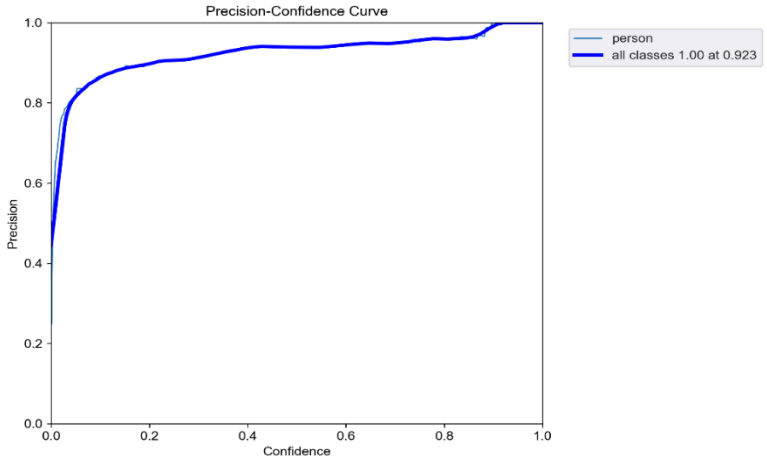

Fig. 6. The precision-confidence curve.

In the PCC diagram, the curve curves upward and left in the low confidence level region, which reflects that the detector can maintain a high recall rate while maintaining a low false positive rate, indicating that the detector has a high target recognition precision in this region, and thus shows a better performance.

3.3 Training Detection Results

Fig. 7 and Fig. 8 shows the training results.

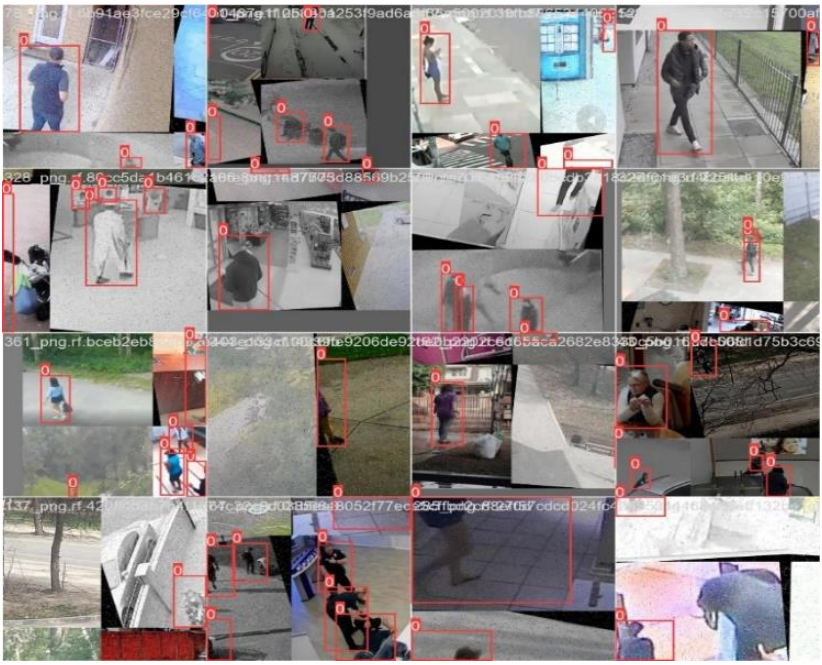

Fig. 7. Training set training results.

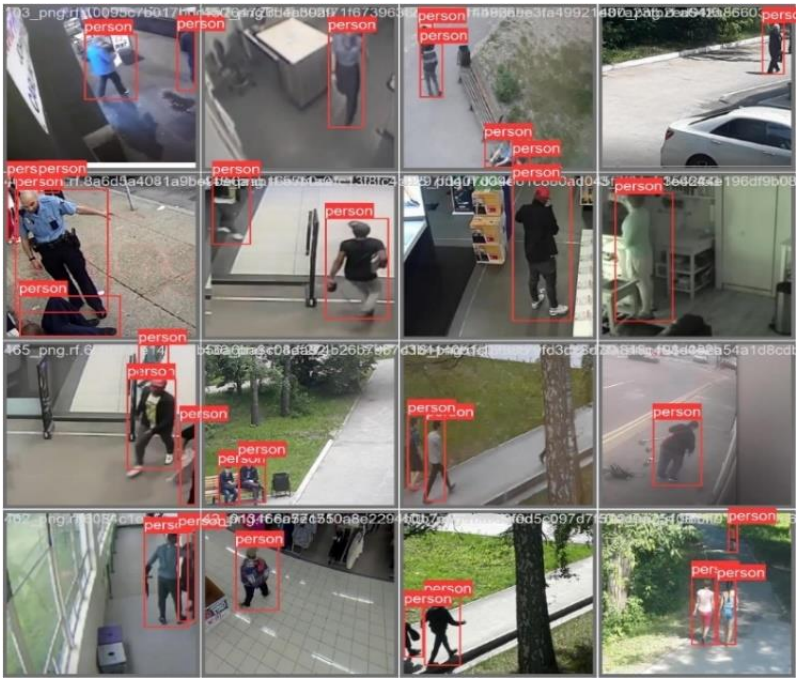

Fig. 8. The result of the validation set.

As illustrated in Fig. 7 and Fig. 8, during the training phase, pedestrians can generally be recognized from various angles and positions, with occasional instances of incorrect identification. During the verification process, the pedestrian can be fully identified. The findings indicate that the model exhibits strong performance and effectively accomplishes the task of pedestrian detection.

4 Conclusion

This study presents the implementation of YOLOv5 for pedestrian detection in real-world scenarios, aiming to enhance the accuracy of pedestrian detection systems for applications in road safety and beyond. YOLOv5 is leveraged to analyze pedestrian presence within detection image data, utilizing CNN for feature extraction and bounding box prediction. The process involves image pre-processing through Mosaic data augmentation, followed by feature extraction and fusion facilitated by the Focus Network and CSP Network in the backbone, and the CSP_2 Network in the neck, culminating in prediction generation. Additionally, model optimization is achieved through the utilization of the CIOU loss function. Experimental findings demonstrate that YOLOv5 achieves high accuracy in pedestrian detection, effectively identifying pedestrians across diverse environments, poses, and angles, with minimal misjudgments. In future endeavors, researchers may explore the utilization of alternative versions of the YOLO model for pedestrian detection training and conduct comparative analyses of their performance.

References

1. Shaozi, L., Shuyuan, C., Songzhi, S., Yundong, W.: Overview of pedestrian detection technology. Acta electronica, 10(04), 814-820 (2012).
2. Dalal, N., & Triggs, B.: Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition, 1, 886-893 (2005).
3. Harris, C., & Stephens, M. A combined corner and edge detector. In Alvey vision conference, 15(50), 10-5244 (1988).
4. Divvala, S., Farhadi, A., Girshick, R., Redmon, J.: You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 779-788 (2016).
5. Farhadi, A., Redmon, J.: YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, 7263-7271 (2017).
6. Farhadi, A., Redmon, J.: Yolov3: An incremental improvement. arXiv preprint: 1804.02767 (2018).
7. Bochkovskiy, A., Liao, H.Y.M., Wang, C.Y.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint:2004.10934 (2020).
8. Anguelov, D., Berg, A.C., Erhan, D., Fu, C.Y., Liu, W., Reed, S., Szegedy, C.: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, pp. 21-37. Springer International Publishing (2016).

9. Darrell, T., Donahue, J., Girshick, R., & Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 580-587 (2014).
10. Roboflow dataset. https://universe.roboflow.com/fyp-qvjss/human-detection-wyz83/dataset/1, last accessed 2023/9/10.