Fuzzy Systems and Soft Computing
ISSN : 1819-4362

OBJECT DETECTION FOR MEDICAL IMAGE ANALYSIS USING REAL-TIME DETECTION TRANSFORMER MODEL

Vishal Rajendrabhai Patel, Research Scholar, Department of Computer Science, Sankalchand Patel University, Visnagar, Gujarat, India. Email: vishal.uttargujarat@gmail.com
Dr. Ronak B. Patel, Associate Professor, Shrimad Rajchanddra Institute of Management and Computer Application, Uka Tarsadia University, Bardoli, Gujarat, India.
Email: ronakcjp@gmail.com

Abstract- Deep learning has become a game-changing method for resolving challenging object detection and pattern recognition problems. The use of a novel detection framework based on the Real-Time Detection Transformer model for the analysis of complex picture data is the main emphasis of this paper, especially in fields like the identification of Kidney stone. To detect early-stage stone, a major cause of vision loss worldwide, precise and effective image processing is necessary. Based on a Transformer-based architecture, the suggested Real-Time Detection Transformer model performs very well in processing complicated and high-dimensional visual input with improved accuracy and robustness. In comparison to models like YOLOv5, YOLOv8, SSD, and DETR, Real-Time Detection Transformer performs better in terms of precision, recall, mAP50, and mAP50-95 metrics, especially when it comes to recognizing small-scale objects and closely spaced targets.

Keywords- *Kidney stone, Real-Time Detection Transformer, object detection, deep learning, automatic diagnosis.*

I. INTRODUCTION

One of the many urological disorders is kidney stones, which, if left untreated, can result in kidney failure, urinary tract obstructions, infections, and other problems. Early diagnosis is essential to safeguard those who suffer from these dangerous conditions and increase the likelihood of patient success stories. Because CT scans may identify kidney stones, even the smallest ones, they are frequently used to diagnose kidney stones. However, manually interpreting CT scans is time-consuming and can result in evaluation errors that lead to unneeded or misleading therapies. A machine learning system designed to diagnose kidney stones from CT scan images is used in this study. It is anticipated that the technology will lessen human error and the burden of radiologists.

In recent years, the rapid development of computer vision and deep learning technology has provided strong support for automatic diagnosis and stone detection in medical image analysis. Convolution Neural Network (CNN) has shown excellent performance in various medical image tasks, especially in image classification and object detection. However, traditional object detection models may face challenges such as small stones and low contrast when processing medical images. Therefore, it is of great significance to select deep-learning models suitable for medical image object detection [3].

Real-Time Detection Transformer is a new target detection model that combines the advantages of the Transformer structure and the traditional detection head and performs well in image target detection tasks. Real-Time Detection Transformer has a detection mechanism that does not require non-maximum suppression (NMS), which can better adapt to small and dense target detection [4]. This feature is particularly important in the detection task of Kidney Stone, because the tiny stone in kidney images are usually densely distributed, and conventional detection models may be difficult to effectively capture and locate [5].

To enhance the precision and reliability of stone detection, this research employs the Real-Time Detection Transformer model for the automatic identification of kidney stones. Through the optimization of the network architecture and the modification of the loss function, the model is capable of accurately identifying stones of various sizes in medical images without the need for

post-processing. Furthermore, taking into account the variations in shape and texture of stones in medical images, the model incorporates a simulated-scale feature extraction mechanism to bolster its ability to recognize targets of differing scales.

This research will involve training and validating the model utilizing the publicly available Kidney Stone dataset, while assessing its performance through various indicators, including detection accuracy, recall rate, and mean average precision (mAP). The results of this experiment will be compared against established target detection models to demonstrate the advantages of the Real-Time Detection Transformer in the domain of medical image target detection. Future enhancements to the model will focus on optimizing the attention mechanism, implementing adaptive weight adjustments, and exploring additional strategies to enhance the robustness and stability of the detection outcomes.

This research focuses on the development and execution of an automated detection system for kidney stones utilizing a Real-Time Detection Transformer. By creating a robust and precise deep learning model, the study offers an automated approach for the early screening and diagnosis of retinopathy. This initiative facilitates the timely identification and prevention of chronic kidney diseases, thereby enhancing patient quality of life and optimizing the efficiency of healthcare services.

II. METHOD

To achieve automatic detection of kidney stones, this research employs the Real-Time Detection Transformer deep learning model to facilitate the identification and localization of stones through an end-to-end target detection framework. The Real-Time Detection Transformer is utilized for target detection. To enhance the optimization of model parameters, this study incorporates a comprehensive loss function, wherein the classification loss is represented by cross-entropy loss, while the localization loss is defined by both L1 loss and generalized IoU (GIoU) loss. The comprehensive loss function is defined as:

$$L = \lambda_{cls} \cdot L_{cls} + \lambda_{L1} \cdot L_{L1} + \lambda_{GIoU} \cdot L_{GIoU}$$

Among them, λcls, λL1 and λGIoU are the weight hyper parameters of the corresponding losses. The cross parts: feature extraction network, position encoding module, and target detection head, forming an NMS-free target detection architecture. This section will derive the target detection process of the model and the definition of the loss function in detail. The model architecture is shown in Figure1.

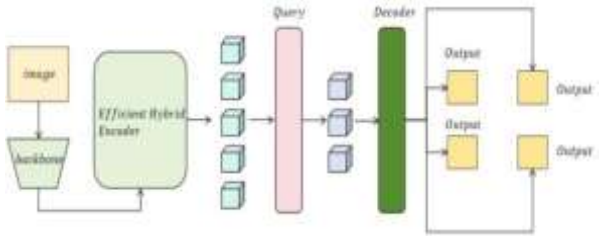


Figure 1 Model architecture diagram

First, in the image input stage, given an input image *I* with a size of *H*W*C*, a multi-scale feature presentation is extracted through the backbone network (such as ResNet or CNN module), which is recorded as a feature map $F \square R^{H \square W \square C}$, where *D* represents the feature dimension. The feature map is input to the position encoding module, and the position encoding vector is obtained through the position embedding layer to ensure that the Transformer module can capture the spatial position information of the target.

In the object detection head, Real-Time Detection Transformer uses a multi-head self-attention mechanism between query embedding and feature maps to calculate the attention score through dot

product operations. The core formula of the attention mechanism is:

$$Attention(Q,K,V) = soft\max(\frac{QKT}{\sqrt{d_k}})V$$

Among them *Q, K, V* represents the query, key and value matrices respectively, and $d_k$is the scaling factor of the feature dimension. Through this formula, the model can distinguish the stone area from the global background and capture potential stone targets.

The detection results output by the model include the category distribution and position bounding box prediction of the target. In order to optimize the model parameters, this study

Introduces a comprehensive loss function, in which the classification loss uses the cross-entropy loss, and the position loss uses the L1 loss and the generalized IoU (GIoU) loss.

In order to further improve the detection performance, Real-Time Detection Transformer completes the target assignment through a dynamic matching algorithm during the prediction process.

Through the optimization of the above target detection mechanism and comprehensive loss function, this study achieved efficient automatic detection of Kidney Stone lesions. The model fully utilized the attention mechanism and NMS-free detection strategy of Real-Time Detection Transformer, improved the recognition ability of small lesions and dense targets, and ensured the accuracy and robustness of the detection results.

III. EXPERIMENT

A. Datasets

This study selected the dataset, a publicly available dataset of Kidney Stone, as the experimental data source. This dataset is a widely used annotated dataset for medical image analysis, specifically for the automatic diagnosis of Kidney Stone. This Dataset contains a collection of CT scan images of human kidneys, categorized into four distinct classes: Simple and Stone. Normal images are 3200 and stone images are 800. Lastly, this dataset has been developed to be used solely for medical imaging research and development for the algorithms that segment and classify kidney diseases. This data set is suitable for the training and verification of deep learning models.

Through the training of this dataset, researchers can evaluate the performance of various deep learning models in actual clinical scenarios and further promote the development of early screening and diagnosis technologies for Kidney Stone. Through comparative experiments with the model, this dataset can effectively verify the generalization ability and accuracy of the model, especially its application effect in early diagnosis of the disease.

B. Experimental Results

In this study, we will conduct comparative experiments based on the Real-Time Detection Transformer model with four mainstream target detection models, namely YOLOv5[6], YOLOv8[7], SSD [8] and DETR [9], to evaluate the performance of each model in the automatic detection task of Kidney Stone lesions. As the latest versions of the YOLO series models, YOLOv5 and YOLOv8 are widely used in target detection tasks with their fast-reasoning speed and strong real-time detection capabilities. SSD (Single Shot Multibox Detector) performs target detection through multi-scale feature maps, has a good balance, and shows excellent performance between speed and accuracy. DETR (Detection Transformer) introduces the Transformer structure, which can effectively handle complex target relationships, especially when dealing with dense targets. We will compare the performance of each model in terms of detection accuracy (Precision), recall rate (Recall), average precision (mAP), and other indicators, especially when dealing with Kidney Stone images, to

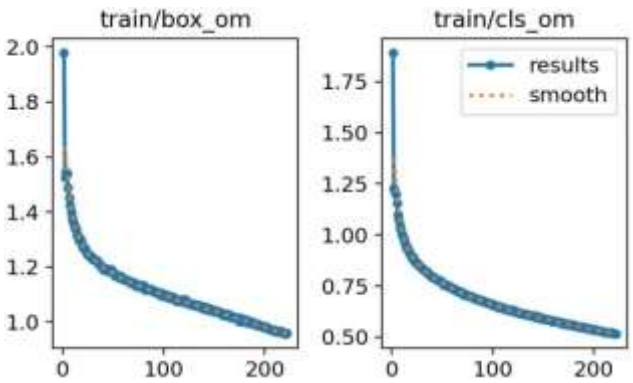
evaluate the recognition ability of each model for small lesions and complex backgrounds. The purpose of analyzing the experimental results is to verify the advantages of Real-Time Detection Transformer over traditional target detection models in Kidney Stone detection, especially its robustness and accuracy in target detection tasks without the need for an NMS detection mechanism. The experimental results are shown in Table 1.

Table 1 Experimental Results

Model	Precision	Recall	mAP50	mAP50-95
SSD	0.81	0.74	0.78	0.61
YOLOv5	0.85	0.79	0.82	0.68
YOLOv8	0.86	0.80	0.83	0.70
DETR	0.85	0.78	0.82	0.66
Real-Time Detection Transformer	0.90	0.84	0.87	0.76

From the experimental results, it can be seen that the Real-Time Detection Transformer model performs well in multiple evaluation indicators, surpassing several other target detection models, especially in small target detection capabilities and robustness in complex scenes. First, the Precision of Real-Time Detection Transformer is 0.90, which is significantly higher than other models, indicating that it can detect Kidney Stone lesions more accurately and reduce the generation of false positive results. In contrast, the precisions of YOLOv5 and YOLOv8 are 0.85 and 0.86 respectively. Although they also perform well, the gap in this indicator shows that Real-Time Detection Transformer has higher reliability in target recognition and positioning. The SSD and DETR models have lower precisions of 0.81 and 0.85 respectively, indicating that they may have certain errors in the process of visual feature extraction and target recognition, especially in the detection of subtle lesions such as Kidney Stone lesions. The performance of Real-Time Detection Transformer is more in line with clinical needs.

Secondly, Real-Time Detection Transformer also has a significant advantage in Recall, with a value of 0.84, which is higher than YOLOv8's 0.80, YOLOv5's 0.79, and DETR's 0.78. A higher recall rate means that Real-Time Detection Transformer can identify more stones. Real-Time Detection Transformer obviously has stronger capture ability. In contrast, although the recall rates of YOLOv5 and YOLOv8 are higher, they are still lower than Real-Time Detection Transformer, which may be related to their limitations in dealing with small targets. The recall rate of the SSD model is 0.74, the lowest among all models, indicating that it is less effective in detecting some small lesions and difficult-to-identify lesion areas, which also reflects the shortcomings of SSD in complex lesion scenes.



In general, Real-Time Detection Transformer not only performs well in all indicators but also its NMS-free design strategy and Transformer-based deep learning architecture are particularly suitable for the automatic detection of Kidney Stone. Real-Time Detection Transformer can effectively cope with the challenges of small stones, which makes it highly practical in clinical applications. Compared with YOLOv5, YOLOv8, SSD, and DETR, RT- DETR has obvious advantages in accuracy, recall, and target positioning accuracy, especially in the automatic detection of Kidney Stone, a complex disease. Real-Time Detection Transformer has shown its unique advantages. These results show that Real-Time Detection Transformer has higher potential in the early screening and diagnosis of the disease, and can provide more reliable and accurate support for the diagnosis of Kidney Stone.

In addition, we also give the loss function decline graph during the experiment, as shown in Figure 2.

Figure 2 Loss function changes with epoch

Finally, we present the ablation experiment of the article. The experimental results are shown in Table 2. During the ablation experiment, we explored the impact of different optimizers on the experimental results.

Table 2 Ablation Experiment Results

LR	**Precision**	**Recall**	**mAP50**	**mAP50-95**
0.025	0.85	0.80	0.82	0.70
0.03	0.84	0.78	0.80	0.68
0.005	0.86	0.82	0.85	0.72
0.02	0.87	0.83	0.86	0.74
0.01	0.90	0.85	0.88	0.76

Through experimental comparisons of different learning rates, the model performance is best when the learning rate is 0.01, and Precision, Recall, mAP50, and mAP50-95 all reach the highest values. Lower learning rates (such as 0.005 and 0.02) perform slightly worse, especially in terms of precision and recall. The experimental results show that an appropriate learning rate can significantly improve the performance of the Real-Time Detection Transformer model in Kidney Stone detection.

IV. CONCLUSION

Through improved performance across precision, recall, and mAP metrics, this study demonstrates the benefits of the Real-Time Detection Transformer model, a Transformer-based automatic detection framework, in advancing state-of-the-art target detection. Comparative tests show that Real-Time Detection Transformer performs better than popular models like YOLOv5, YOLOv8, SSD, and DETR, especially when it comes to managing tiny objects and being resilient against complicated backdrops. These findings support Real-Time diagnosis Transformer's promise as a cutting-edge method for handling high-dimensional image analysis problems, such as kidney stone diagnosis, which necessitates accuracy and effectiveness in deciphering complex visual patterns.

Even though Real-Time Detection Transformer performs well, it may still be further optimized to overcome the difficulties caused by irregular data distributions and extremely complicated visual backdrops. Enhancing the model's multi-scale feature learning capabilities,

improving the caliber and variety of training datasets, and incorporating multimodal data sources can be the main goals of future research. Real-Time Detection Transformer may be used to address a variety of automated image processing problems, including the detection of glaucoma, macular degeneration, and other ocular disorders, in addition to its present use in kidney stone detection. The creation of more thorough diagnostic tools will be made possible by the integration of deep learning with multimodal information, such as physiological measures and medical histories. Additionally, Furthermore, Real-Time Detection Transformer will be positioned as a fundamental framework for general-purpose image analysis as advancements in Transformer-based designs continue to spur innovation in fields including multi-modal data fusion, scalable AI systems, and real-time image processing.

All things considered, this study highlights how artificial intelligence has the potential to revolutionize automated diagnostics and visual identification. Real-Time Detection Transformer is a prime example of how deep learning—specifically, Transformer-based models—can transform challenging target detection and data-driven decision-making tasks, opening the door to clever and reliable solutions in domains ranging from industrial applications to medical imaging.

REFERENCES

[1] Wang J, Luo J, Liu B, et al. Automated diabetic retinopathy grading and lesion detection based on the modified R-FCN object-detection algorithm [J]. IET Computer Vision, 2020, 14(1): 1-8.

[2] Ramesh P V, Ramesh S V, Subramanian T, et al. Customised artificial intelligence toolbox for detecting diabetic retinopathy with confocal true color fundus images using object detection methods [J]. tnoa Journal of Ophthalmic Science and Research, 2023, 61(1): 57-66.

[3] Nur-A-Alam M, Nasir M M K, Ahsan M, et al. A faster RCNN-based diabetic retinopathy detection method using fused features from retina images [J]. IEEE Access, 2023, 11: 124331-124349.

[4] Parthiban K, Kamarasan M. Diabetic retinopathy detection and grading of retinal fundus images using coyote optimization algorithm with deep learning [J]. Multimedia Tools and Applications, 2023, 82(12):18947-18966.

[5] Agarwal S, Bhat A. A survey on recent developments in diabetic retinopathy detection through integration of deep learning [J]. Multimedia Tools and Applications, 2023, 82(11): 17321-17351.

[6] Santos C, Aguiar M, Welfer D, et al. A new method based on deep learning to detect lesions in retinal images using YOLOv5[C]//2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021: 3513-3520.

[7] Rizzieri N, Dall'Asta L, Ozoliņš M. Diabetic Retinopathy Features Segmentation without Coding Experience with Computer Vision Models YOLOv8 and YOLOv9 [J]. Vision, 2024, 8(3): 48.

[8] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multi box 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.

[9] ZhuX, SuW, LuL, et al. Deformable detr: Deformable transformers for end-to-end object detection [J]. arXiv preprint arXiv:2010.04159,2020.