Scholar Publishing Group
International Journal of Big Data Intelligent Technology
https://doi.org/10.38007/IJBDIT.2025.060110
ISSN 2790-0932 Vol. 6, Issue 1: 101-108

Research on Intelligent 3D Reconstruction System Integrating Transformer and Adaptive Point Cloud Registration

Chuying Lu[1, a*]

[1]*University of Michigan, Ann Arbor 48109, MI, United State*

[a]*chuyinglu2226@gmail.com*

[*]*Corresponding author*

Keywords: 3D object detection, point cloud and image fusion, Transformer architecture, adaptive point cloud registration, intelligent 3D reconstruction system

Abstract: With the rapid development of intelligent vehicle technology, 3D object detection and tracking play a crucial role in the field of autonomous driving. This article deeply studies the point cloud and image fusion 3D object detection and tracking algorithm in the field of autonomous driving. Laser radar and surround view camera are used as perception sensors, and based on deep learning theory and methods, the focus is on overcoming the difficulties of multi-modal data fusion in 3D detection and tracking under complex traffic conditions. This article proposes a series of innovative algorithms, including MaskSensing algorithm based on image instance segmentation, DeformFusion algorithm based on Transformer architecture, MixFusion algorithm with hybrid fusion strategy, and DeepTrack3D algorithm. These algorithms have achieved significant results on datasets such as nuScenes, effectively improving the accuracy and robustness of 3D object detection and tracking. In the future, further research is needed in areas such as temporal fusion, interactive fusion, and unsupervised learning to enhance the performance of autonomous driving technology.

1. Introduction

With the surge in car ownership and advancements in AI technology, smart cars have garnered significant attention. The American Society of Automotive Engineers categorizes autonomous driving into six levels, aiming ultimately for Level 5 full autonomy. Currently, achieving Level 4 highly autonomous driving holds practical importance and is a key research area for both industry and academia. The intelligent automotive system comprises modules like environmental perception, decision planning, and motion control. Among these, the environmental perception system is crucial

for autonomous driving. Traditional 2D perception technology is inadequate for autonomous driving tasks, which require 3D spatial awareness of objects' distance, size, direction, and speed. 3D object detection and tracking are vital in autonomous driving. By spatially calibrating and temporally aligning sensors, point cloud and image coordinates can be correlated, facilitating multi-sensor multimodal fusion perception. This article aims to integrate point cloud and image information to enhance 3D object detection accuracy, and based on these detections, achieve object tracking, record historical trajectories, and provide comprehensive perception data for downstream modules.This research is of great significance for the development and application of intelligent vehicles. This article uses laser radar and surround view cameras, based on deep learning theory and methods, to focus on overcoming the difficulties of multimodal data fusion in three-dimensional detection and tracking under complex traffic conditions, so that the perception system has excellent robustness, accuracy, and adaptability in urban and unstructured scenes.

2. Correlation Theory

In recent years, a large amount of research has fully demonstrated the outstanding performance of Transformer architecture in multiple fields. In the cutting-edge field of molecular generation, researchers have proposed a hybrid Transformer architecture with a quantified self attention mechanism, aimed at further optimizing the process of molecular generation, improving efficiency and accuracy. In the field of mechanical fault diagnosis, especially in the diagnosis of bearing faults, researchers have combined Swin Transformer deep learning with acoustic emission technology to create a new fault diagnosis method, providing strong support for early detection and accurate positioning of mechanical faults. In terms of remote sensing image processing, the TransRefine method proposed by researchers cleverly utilizes the characteristics of Transformers, significantly enhancing the effect of feature refinement, thereby achieving zero sample scene classification of remote sensing images. Researchers have also made significant progress in the fields of point cloud processing and 3D reconstruction. Their various algorithms, such as unsupervised domain adaptive multimodal feature fusion method for partial point cloud registration and Transformer based end-to-end cloud registration algorithm, not only improve the efficiency of point cloud processing, but also enhance the accuracy of 3D reconstruction. There are also studies that combine advanced technologies such as adaptive neighborhood feature loading rate and adaptive segmentation, providing new solutions for fast registration of MEMS LiDAR point clouds. In terms of understanding and reconstructing 3D scenes, the OpenOcc method proposed by researchers achieves open vocabulary 3D scene reconstruction through occupancy representation, providing a new perspective and tool for understanding and interacting with 3D scenes. The intelligent situational awareness method based on fractal dimension and multidimensional reconstruction boosts accuracy and real-time performance. Edge-guided methods yield precise 3D reconstructions from multi-view sketches and RGB images. Combining improved SFM and deep learning reconstructs and segments road potholes, aiding road maintenance and safety. These studies propel technological advancements and pave the way for intelligent transportation and smart cities.

3. Method

3.1. MaskSensing 3D Object Detection Algorithm Framework

The MaskSensing algorithm is a 3D object detection algorithm that combines image instance segmentation and point cloud processing, aiming to improve the detection performance of distant and small target objects through the fusion of multimodal information. This algorithm takes raw point clouds, instance masks, semantic categories, camera intrinsic parameters, and extrinsic

parameters between LiDAR and camera as inputs. In terms of processing flow, the algorithm first projects the point cloud onto the corresponding image by calculating the transformation matrix and calculates the depth. Then, it performs foreground densification on the point cloud based on the instance mask, while adding category information to enhance the semantic features of the point cloud. In order to overcome the sparsity and disorder of point clouds, the algorithm introduces dynamic voxel geometric feature enhancement encoding, which regularizes the point cloud through dynamic voxel encoding and adds geometric encoding on the basis of voxel geometry to balance point cloud information and improve detection performance. In terms of feature enhancement, the MaskSensing algorithm adopts various strategies. Densifying the foreground point cloud and adding category information improve detection of small targets. Dynamic voxel encoding and geometric feature enhancement enhance feature extraction, stabilizing and accuracy detection results. Evaluated on the nuScenes dataset, MaskSensing outperforms state-of-the-art methods, especially in detecting small or rare objects.The data results are shown in Figure 1

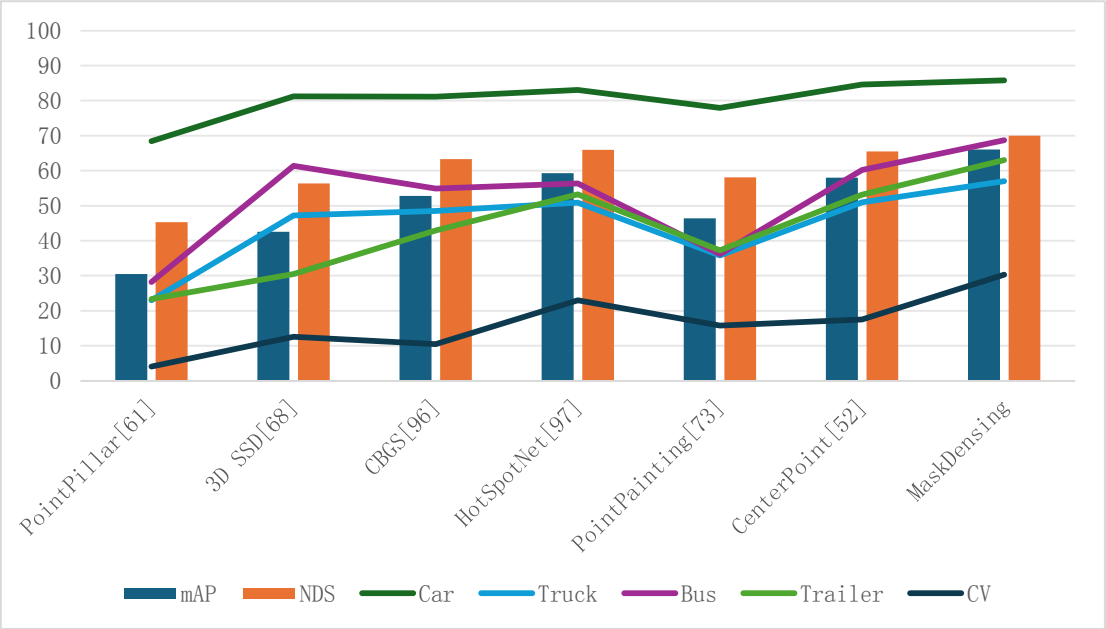

Figure 1. Performance Comparison of 3D Object Detection Methods

Compared to the baseline algorithm CenterPoint, MaskSensing showed an 8.1% increase in mAP and a 4.5% increase in NDS, demonstrating the superiority of the multimodal fusion algorithm. In order to comprehensively verify the effectiveness and rationality of each component of the MaskSensing algorithm, detailed ablation experiments were also conducted in the experiment. The ablation experiment results show that the introduction of each component has a positive impact on the algorithm performance, proving the rationality and effectiveness of the algorithm design. In particular, the semantic enhancement and foreground densification of the original point cloud, as well as the introduction of dynamic voxel encoding and geometric feature enhancement in image instance segmentation results, contribute the most to the improvement of algorithm performance. The MaskSensing algorithm effectively improves the accuracy and stability of 3D object detection by integrating multimodal information from image instance segmentation and point cloud processing. This algorithm performs well in processing flow, feature enhancement, and experimental evaluation, providing strong technical support for application scenarios such as autonomous driving.

3.2. DeformFusion 3D Object Detection Algorithm

The DeformFusion 3D object detection algorithm cleverly integrates a deformable attention module and a hotspot decoder, achieving efficient fusion of point clouds and image features, significantly improving detection accuracy and convergence speed. This algorithm is inspired by Deformable DETR and proposes a deformable space constrained feature aggregation module, which uses the initial point cloud prediction as a reference point to adaptively fuse image features. The algorithm framework combines point cloud and image branches. The point cloud branch uses dynamic voxel encoding and a 3D network to create BEV feature maps, introducing a hotspot decoder for faster convergence and higher accuracy. The image branch fuses features adaptively using Deformable DETR. Both branches learn object-environment associations via Transformers, enhancing detection capabilities.The result is shown in Figure 2

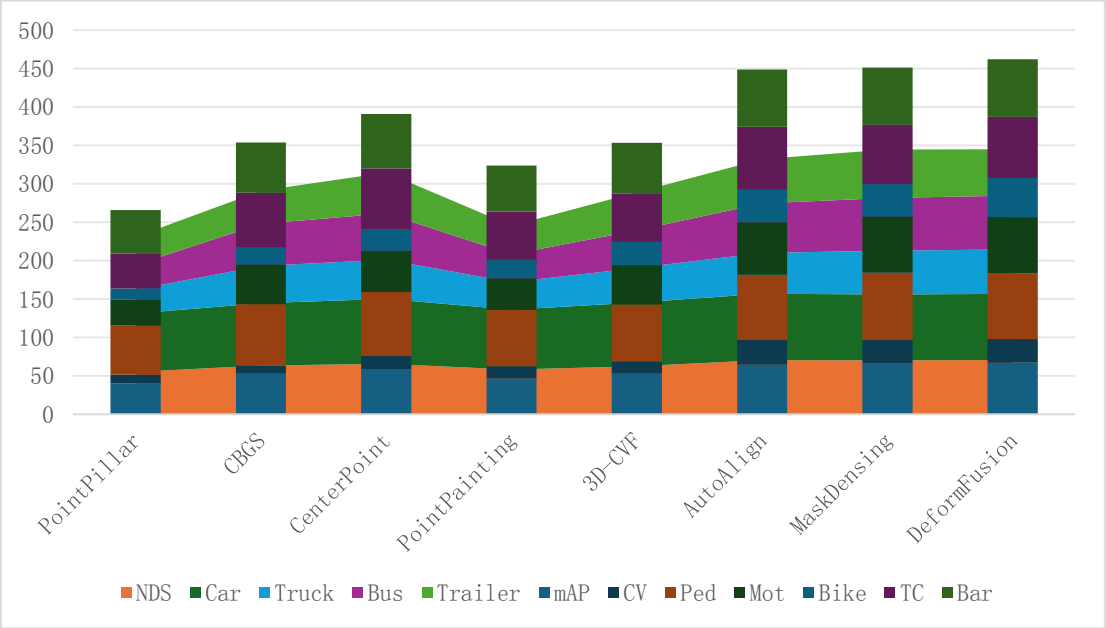

Figure 2. Comparison Table of 3D Object Detection Algorithm Performance

This section validates the effectiveness and rationality of the DeformFusion algorithm component through ablation experiments using a validation set. DeformFusion is a 3D object detection algorithm that combines point clouds and image features at the feature level, demonstrating its potential in optimizing initial predictions and outputting final detection results.The HotSpotDecoder in DeformFusion's pure point cloud branch initializes queries with hotspots, boosting mAP and NDS by 1.0% and 0.8% respectively. DeformFusion's deformable space feature aggregation module effectively fuses point clouds, image features, and LiDAR queries, improving mAP and NDS by 0.8% and 0.6% over visual methods. Experiments show DeformFusion outperforms input-level fusion, validating Transformer's advantages and feature-level fusion's effectiveness in 3D object detection.

3.3. MixFusion 3D Object Detection Algorithm

The MixFusion algorithm has demonstrated excellent performance in the field of 3D object detection, cleverly combining the advantages of input layer fusion and feature layer fusion. This algorithm takes raw point clouds and panoramic images as inputs, uses MaskRCNN for instance segmentation, and generates virtual point clouds based on instance results and point cloud

projections. The MaskSensing algorithm enhances the density and semantic information of the point cloud, improving detection recall and accuracy. In the feature extraction stage, MixFusion adopts the Transformer architecture to update object queries through self attention and cross attention mechanisms, and utilizes a three-dimensional basic network to transform voxel features into BEV feature maps. In order to optimize feature layer fusion, the algorithm uses DeformFusion's deformable space constrained feature aggregation module to adaptively fuse image features after initial prediction in the first stage, further optimizing the detection results. MixFusion adopts a label assignment strategy based on bipartite graph matching, constructs a cost matrix, and solves it using the Hungarian algorithm to assign truth or background labels to each predicted bounding box. In the design of the loss function, the algorithm comprehensively considers category loss, bounding box geometric loss, and hotspot assisted supervised loss, and achieves high-precision 3D object detection through fine loss calculation and optimization. Therefore, the MixFusion algorithm has broad application prospects in fields such as autonomous driving and robot navigation that require high-precision 3D object detection.

4. Results and Discussion

4.1. 3D Object Tracking Algorithm Based on Point Cloud and Image Fusion

The DeepTrack3D 3D object tracking algorithm is an innovative algorithm that extends the two-dimensional object tracking algorithms SORT and DeepSort to three-dimensional space. The SORT algorithm, with its simplicity, effectiveness, and practicality, establishes the basic paradigm of detection before tracking. It predicts the tracking box of the previous frame through the motion model of the Kalman filter, calculates the intersection and union ratio with the detection box of the current frame, and uses the Hungarian algorithm to calculate the matching relationship to achieve object tracking. The SORT algorithm only matches through intersection and union ratios, resulting in significant switching of identifiers. The DeepSort algorithm introduces image depth epigenetic information based on this, uses Mahalanobis distance to measure the similarity between detection boxes and tracking boxes, and introduces cascade matching, greatly reducing label switching. The DeepTrack3D algorithm utilizes the high-precision 3D object detection results generated by the fusion of point clouds, image input layers, and feature layers to redefine the Kalman filter state variables and observations, and extends the DeepSort algorithm in 3D space. The algorithm framework adopts the paradigm of "tracking through detection" and uses MixFusion algorithm as the detection algorithm to obtain the query features of 3D bounding boxes and corresponding fused objects. Using the Kalman filter algorithm to predict the tracking frame, combining the semantic features of object queries in the previous and subsequent frames with the geometric features calculated from the 3D bounding box to construct a cost matrix, and using the Hungarian algorithm for object matching. Perform lifecycle management based on matching results, and use Kalman filtering algorithm combined with matching detection boxes to update objects for successfully matched tracking boxes, forming a closed loop to achieve object tracking. The DeepTrack3D algorithm improves the robustness of tracking by introducing semantic features of point cloud and image fusion and geometric features of 3D bounding boxes, and adopts a two-stage lifecycle management strategy to effectively reduce the number of identification transformations.

4.2. DeepTrack3D Association Update Strategy

The DeepTrack3D algorithm updates the status of all tracking boxes in real-time in the lifecycle management module, ensuring the accuracy and continuity of tracking. The status of the tracking box is mainly divided into undetermined state, determined state, and deleted state. When the first

match between the tracking box and the detection box is successful, the tracking box is initialized to an undetermined state, and the number of matches is marked as 1. When the number of successful matches between the tracking box and the detection box exceeds the set threshold, the tracking box transitions from an undetermined state to a determined state. If the tracking box in an undetermined state cannot be continuously tracked, it will be converted to a deleted state. For tracking boxes in a certain state, if the number of matching failures exceeds the set threshold, they will also be converted to a deletion state. After in-depth analysis of the reasons for identifier conversion, the DeepTrack3D algorithm found that incorrect associations and early stopping were the main reasons, with early stopping being particularly prominent. Due to inaccurate positioning and feature information of low confidence detection boxes, they may be filtered out by the matching threshold in the early stages, resulting in tracking boxes being unable to match correctly and stopping prematurely. When these tracking boxes are tracked again, it may cause identification changes. DeepTrack3D uses a two-stage association strategy: cascade matching and geometric matching. Cascade matching pairs current detection boxes with previous tracking boxes, prioritizing those with shorter disappearance times. Matched boxes update the Kalman filter. Unmatched boxes enter geometric matching using 3D GIoU to determine further matches. This preserves tracking continuity, updating matched boxes, creating new trajectories for unmatched detections, and retaining unmatched tracking boxes based on lost frame counts, enhancing tracking robustness and accuracy.

4.3. Comparative Analysis of Evaluation Effects

This chapter proposes a 3D object tracking algorithm called DeepTrack3D, which integrates point cloud and image data. We have outlined the basic framework of multi-target tracking algorithms and pointed out that the work in this chapter is a three-dimensional multimodal extension based on the two-dimensional object detection algorithm DeepSORT. We introduced two key algorithms in multi-target tracking: Kalman filter and Hungarian algorithm. The core of the DeepTrack3D algorithm lies in its unique matching strategy. We utilize the MixFusion 3D object detection algorithm, which is based on the Transformer architecture and can conveniently obtain 3D bounding boxes and their corresponding point cloud and image fusion features simultaneously. Based on this characteristic, we propose a cascaded matching strategy that integrates the semantic features of point clouds and images with the geometric features of 3D bounding boxes. In order to maintain the continuity of tracking, we introduced a two-stage geometric matching strategy based on 3D GIoU for lifecycle management. We extend the two-dimensional Kalman filter to three-dimensional states to achieve high-precision three-dimensional target tracking. In terms of evaluation, we used the 3D object tracking evaluation metrics from the nuScenes dataset, including Average Multi Object Tracking Accuracy (AMOTA) and Average Multi Object Tracking Precision (AMOTP). The experimental results show that the DeepTrack3D algorithm outperforms previous advanced algorithms in overall performance and has improved tracking performance in various categories. The ablation experiment further validated the effectiveness of each strategy. We visualized the tracking performance of some typical scenarios and qualitatively verified the effectiveness of the algorithm. In summary, the DeepTrack3D algorithm achieves high-precision 3D object tracking by fusing point clouds and image data, utilizing unique matching strategies and Kalman filtering extensions. The experimental results fully demonstrate the effectiveness and superiority of the algorithm.

5. Conclusion

This article has conducted in-depth research on point cloud and image fusion 3D object detection and tracking algorithms in the field of autonomous driving, and has achieved a series of important results. A MaskSensing algorithm based on image instance segmentation is proposed to address the challenges of point cloud lacking semantic information and detecting distant and small objects. Temporal fusion is a direction worth exploring in the future, which can consider multi frame fusion to achieve inter frame feature association fusion, research multi frame feature preservation, dynamic object feature alignment, and inter frame feature fusion methods. In terms of interactive fusion, future research will focus more on the equal fusion of point clouds and images, restoring the natural fusion of image features and point cloud features from a bird's-eye view through learning, or using point cloud and image projection interactive guidance for fusion, further improving detection performance. In terms of unsupervised learning, given the high cost of annotation in autonomous driving tasks, the future will consider using unsupervised pre trained large models combined with task specific fine-tuning or distillation to improve the model's generalization ability. These research directions will provide new ideas and methods for the development of future autonomous driving technology.

References

[1] Fan, Sunjia, et al. "Defense methods against multi-language and multi-intent LLM attacks." International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2024). Vol. 13403. SPIE, 2024.

[2] Tan, Weiyan, Shujia Wu, and Ke Ma. "Freight Volume Prediction for Logistics Sorting Centers Using an Integrated GCN-BiLSTM-Transformer Model." Advances in Computer and Engineering Technology Research 1.4 (2024): 320-324

[3] Xiang, Y., Li, J., & Ma, K. (2024, October). Stock Price Prediction with Bert-BiLSTM Fusion Model in Bimodal Mode. In Proceeding of the 2024 5th International Conference on Computer Science and Management Technology (pp. 1219-1223).

[4] Chen, H., Yang, Y., & Shao, C. (2021). Multi-task learning for data-efficient spatiotemporal modeling of tool surface progression in ultrasonic metal welding. Journal of Manufacturing Systems, 58, 306-315.

[5] Shi C. Research on Deep Learning Algorithms for Predicting DNA-Binding Proteins Based on Sequence Information[C]//2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE). IEEE, 2024: 1566-1570.

[6] Xu Y. Research on UAV Navigation System Based on Behavioral Programming[C]//2024 IEEE 7th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). IEEE, 2024: 419-425.

[7] Hao, Linfeng. "Application of Machine Learning Algorithms in Improving the Performance of Autonomous Vehicles." Scientific Journal of Technology 7.2 (2025): 118-124.

[8] Guo X. Research on systemic financial risk early warning based on integrated classification algorithm[C]//2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE). IEEE, 2024: 1586-1591.

[9] Chen, Junyu. "Research on Intelligent Data Mining Technology Based on Geographic Information System." Journal of Computer Science and Artificial Intelligence 2.2 (2025): 12-16.

[10] Xu, Yue. "Research on Maiustream Web Database Development Techncl ogy." Journal of Computer Science and Artificial Intelligence 2.2 (2025): 29-32.

[11] Li, Bin. "Application of Data Analysis in Climate Policy in Environmental Planning." Frontiers in Science and Engineering 5.2 (2025): 106-112.

[12] Zhu, Zhongqi. "Strategies for Improving Vector Database Performance through Algorithm Optimization." Scientific Journal of Technology 7.2 (2025): 138-144.

[13] Cui, Naizhong. "Optimization Strategies for Traffic Signal and Identification Design." Frontiers in Science and Engineering 5.2 (2025): 92-98.

[14] Wang, Buqin. "Strategies and Practices for Load Test Optimization in Distributed Systems." Scientific Journal of Technology 7.2 (2025): 132-137.

[15] Ding, Maomao. "Design Innovation and User Satisfaction Improvement of AI Video Creation Tools." Scientific Journal of Technology 7.2 (2025): 112-117.

[16] Zhang, Jingtian. "Research on Worker Allocation Optimization Based on Real-Time Data in Cloud Computing." Frontiers in Science and Engineering 5.2 (2025): 119-125.

*[17] Gu, Yiting. "Practical Approaches to Develo**High-performance Web Applications Based on React." Frontiers in Science and Engineering 5.2 (2025): 99-105.*

[18] Ma Z. Strategies for Enhancing Customer Lifetime Value through Data Modeling[J]. European Journal of Business, Economics & Management, 2025, 1(1): 1-7.

[19] Li, X. (2025). Research on Three-dimensional Modeling of Urban Buildings based on CityGM. Scientific Journal of Technology, 7(3), 302-306

[20] Zhang Y. Research on Optimization of Engineering Cost Database Based on Big Data and Intelligent Technology[J]. International Journal of New Developments in Engineering and Society, 2024, 8(5).

[21] Li, X.(2025)"Research on the application of GPS, total station and CAD Technology in architectural Grid." Computer Life (2024),12(3),36-39.

[22] Zhang, Yiru. "Design and Implementation of a Computer Network Log Analysis System Based on Big Data Analytics." Advances in Computer, Signals and Systems,(2024) 8(6),40-46.

[23] Yang, Jinzhu "Integrated Application of LLM Model and Knowledge Graph in Medical Text Mining and Knowledge Extraction."Social Medicine and Health Management (2024), 5(2): 56-62

[24] Zhang, Jinshuo "Research on Real Time Condition Monitoring and Fault Warning System for Construction Machinery under Multi Source Heterogeneous Data Fusion." Journal of Engineering Mechanics and Machinery (2024), 9(2): 139-144

[25] Wang, Yuxin "Research on Intelligent Macro Image Recognition Algorithm of Oil Pipe Failure Based on Deep Learning." Journal of Image Processing Theory and Applications (2025), 8(1): 1-7

[26] Zhao, Fengyi "Development Design and Signal Processing Algorithm Optimization of Traditional Chinese Medicine Pulse Acquisition System Based on CP301 Sensor." Advances in Computer, Signals and Systems (2024), 8(6): 106-111