OC-DETR: DETR with Orthogonal Channel Attention and CSPO-Fusion for component detection of TEDS

1st Chaowei Song
Huazhong University of Science and Technology
School of Artificial Intelligence and Automation
Wuhan, China
cwsong@hust.edu.cn

2nd Gang Peng*
Huazhong University of Science and Technology
School of Artificial Intelligence and Automation
Wuhan, China
penggang@hust.edu.cn

3rd Chaoze Wang
Huazhong University of Science and Technology
School of Artificial Intelligence and Automation
Wuhan, China
wangcz@hust.edu.cn

4th Mingjun Cong
Huazhong University of Science and Technology
School of Artificial Intelligence and Automation
Wuhan, China
1791169896@qq.com

5th Cong Li
Wuhan Lisai Technology Co.
Wuhan, China
licong2582@163.com

6th Xinbin Xiong
Beijing Railway Engineering Electromechanical Technology Research Institute Co.
Beijing, China
3268407065@qq.com

***Abstract*—In recent years, the evolution of automated maintenance technology for high-speed railways has positioned the Train of EMU failures Detection System (TEDS) as a critical technology for ensuring the safe operation of high-speed trains. TEDS employs line-scan cameras along railway tracks to capture undercarriage and side-view images of Electric Multiple Units (EMUs). However, the diverse shapes of EMU components, complex component backgrounds, and a large number of small components have intensified the challenges of designing component detection algorithms. To address these limitations, this paper introduces a novel real-time component detection framework known as OC-DETR (DETR with Orthogonal Channel Attention and CSPO-Fusion). The framework incorporates a specialized feature fusion module named CSPO-Fusion, within the Cross-Scale Feature Fusion Module (CCFM) to adeptly handle the small target features of EMU components. Additionally, the framework integrates the orthogonal channel attention into the block of the backbone, effectively reducing the redundant information caused by traditional convolution and improving the accuracy of component detection. To further enhance performance across diverse scales, we propose an IoU-aware size-dependent weighted loss function which increases detection precision for small, medium, and large components. In this paper, the TEDS installed at the throat of the train is used to collect the data of EMUs, and the TEDS dataset including CRH380 and CR400AF is constructed. Extensive experimental validation using this dataset demonstrates that the accuracy and speed of the proposed method have improved compared with the current mainstream target detector, and the detection mAP50 of this method is increased by 2.2% compared with the RT-DETR benchmark target detection algorithm, which effectively verifies the effectiveness of this algorithm.**

Keywords—TEDS, Orthogonal Channel Attention, CSPO-Fusion, component detection, lightweight network

I. INTRODUCTION

AS China's railway network continues to expand, with operational speeds reaching up to 350 km/h, the slightest damage to even the smallest components can have a profound impact on the safe operation of the railway. Consequently, routine inspections of EMUs have become crucial for ensuring operational safety. Among various maintenance methods for EMUs, TEDS serves as an essential monitoring tool for monitoring the operational status of EMUs. It achieves real-time image detection and fault identification for critical components on the sides and undersides of EMUs. Through centralized analysis and multi-level application management facilitated by a comprehensive railway network, TEDS plays a significant role in safeguarding the operational safety of EMUs[1], The actual scene diagram of TEDS system is shown in Fig.1. When the high-speed EMU enters the station at a low speed and passes through the line scanning cameras at a constant speed, the camera scans the EMU from multiple angles and splices the line scanning images to obtain the image information of the high-speed EMU. However, because interested objects are usually characterized by small sizes, dim features, low contrast, various types of shapes, and insufficient information, causing extra difficulties in component detection. The images of the high-speed EMU underbody and side captured by the TEDS line scan camera are shown in Fig. 2.

Fig. 1. **The actual scene of the TEDS.** When the EMUs enters the station at a constant speed, the line scan camera deployed at the throat will collect the image information of the EMUs and determine its health status.

In this study, our motivation is to design a component detector with high accuracy that has the potential to be applied to real-time processing on TEDS in the future. The key to alleviate the above problems lies in the feature of small component enhancement and fusion. In terms of feature enhancement, this paper adds the orthogonal channel attention [2] module in the feature extraction network ResNet18[3]. The channel attention generated by Gram-Schmidt orthogonalization is convoluted to the input features along the channel direction, which can effectively remove the redundant

*Corresponding author: Gang Peng

information brought by the traditional convolution, and then more effectively enhance the efficiency of the feature extraction network. At the same time, in the process of CCFM, this paper introduced the low-level feature information of the backbone and uses SPD Conv[4] instead of traditional convolution in the feature fusion to avoid the loss of small target features caused by the traditional convolution operation. In terms of feature fusion, this paper considers the characteristics of images captured by TEDS device, designing a feature fusion module named CSPO-Fusion. Through the three branches in the CSPO-Fusion module, we can effectively learn the global to local feature representation, and then improve the detection accuracy of the components.

(a) (b)

Fig. 2. (a) shows the image of the EMU underfloor, the red box and yellow box represent the locking wire and hexagon bolt. (b) shows the image of the EMU side, the blue box and green box represent the cross bolt and the side door component.

The main contributions of this article are listed as follows:

1)To address the redundancy inherent in features extracted using traditional convolution and to uncover the latent features within strongly correlated channel slices, this paper proposes an orthogonal channel attention residual module. Integrated into the feature extraction network, this module harnesses the orthogonality of channels to more efficiently distill salient image features.

2)To enhance the identification accuracy of EMU components, this paper introduced an innovative module known as CSPO-Fusion. The module is designed to seamlessly integrate multi-scale features at an optimal level within our computational framework. By employing this architecture within the CCFM for the fusion of low-level feature information, the detection efficacy of the component is significantly amplified.

3)To further improve detection performance across various scales of components, this paper introduced a size-dependent weighted loss function integrated with an IoU-aware query selection mechanism. This unified approach ensures better feature learning, significantly enhancing detection accuracy across small, medium, and large objects.

4)This study constructs a dataset for EMU component detection using TEDS, covering various perspectives of the vehicle's underside and sides for the CRH380 and CR400AF models. Experimental analysis based on this dataset demonstrates that the proposed method outperforms other popular object detection models in terms of accuracy, thereby highlighting its superiority and showcasing its potential for application in component detection within the TEDS system.

II. Related Work

At present, object detection methods based on deep learning are mainly divided into two-stage target detection models represented by the fast RCNN series and single-stage target detection models represented by the YOLO and DETR series[5]. As a pioneering work, R-CNN[6] is a two-stage detector, that generates candidate regions in the image by a selective search algorithm, extracts features by using pre-trained AlexNet[7], and then classifies them by SVM. Fast R-CNN[8] improved R-CNN through unified feature extraction and multitask loss function, realized end-to-end training, and significantly improved detection accuracy. Faster R-CNN[9] further introduced the region proposal network (RPN) to share the weight with the backbone network, realizing the rapid and accurate generation of candidate regions. The Yolo series is a representative of the single-stage detector. By directly predicting the target frame on the convolution feature, it simplifies the detection process, realizes rapid detection, and is suitable for industrial real-time applications. YOLOv2[10] introduces anchor frame to improve detection accuracy, YOLOv3[11] realizes multi-scale detection head, YOLOv4 introduces feature fusion neck based on FPN and PAN, while YOLOv5[12], YOLOv6[13], YOLOv7[14], YOLOv8[15] and YOLOv9[16] are further optimized in performance and speed. In addition, the object detection method based on transformer represents the latest progress in this field. As the first end-to-end transformer target detection model, DETR uses self-attention to encode the relationship between image features and targets and regards target detection as a set prediction problem. The advantage of DETR is that it eliminates the dependence on traditional manual design modules such as NMS, but it also faces the challenges of small object detection and slow training convergence. In order to solve these problems, Deformable DETR[18] enhances the detection ability of small objects by introducing deformable convolution. In 2023, the RT-DETR[19] proposed by Baidu team combined the global attention characteristics of lightweight CNN and transformer. By optimizing the model structure and training strategy, it not only surpassed the current most advanced real-time detector in accuracy and speed, achieved significant improvement in detection accuracy and reasoning speed, but also remained stable, and made full use of the advantages of the end-to-end detection process, paving the way for the application of transformer model in the field of real-time object detection.

III. Proposed Method

A. Overview of the Proposed Method

OC-DETR is mainly composed of three parts: backbone feature extraction network, efficient hybrid encoder and decoder with auxiliary prediction heads, as illustrated in Fig. 3. In the structural schematic, to enhance small component detection capabilities, the S_2 layer is connected to the feature fusion module via SPD Conv. The multi-scale feature fusion is realized by an efficient hybrid encoder. The encoder converts multi-scale features into image feature sequences through intra scale feature interaction (AFI) and CCFM. In order to start the decoding process, IOU uses the perceptual query selection mechanism to select a fixed query as the image feature number decoder for the initial object query. Finally, the decoder with auxiliary prediction header iteratively optimizes the object query to generate boxes and confidence scores.

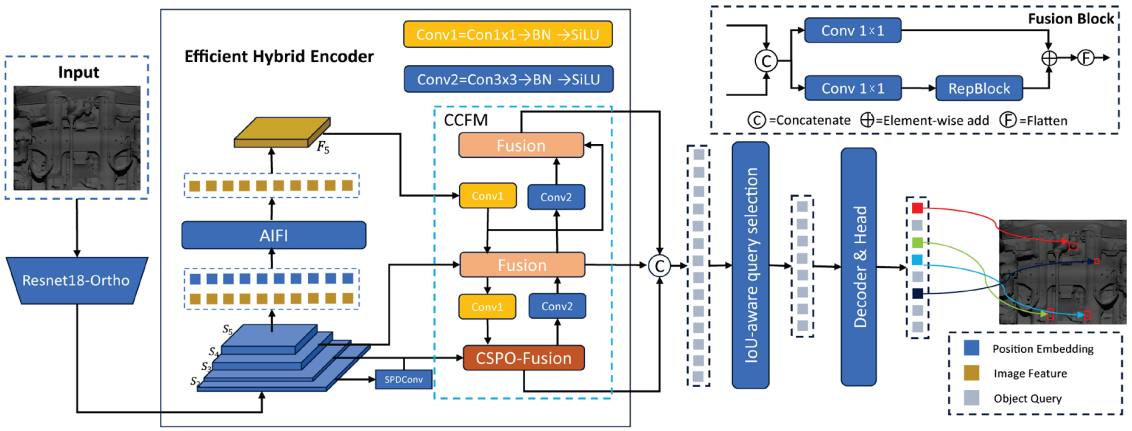

Fig. 3. **The overall framework of OC-DETR proposed in this paper.** The method extracts the features through the backbone Resnet18 with orthogonal attention and sends them to the Efficient Hybrid Encoder for multi-scale feature fusion and sequence conversion. In the fusion process, this article uses SPD convolution to process low-level features to avoid information loss. Through sufficient multi-scale feature fusion, it can help the network effectively understand contextual information. Finally, the component box is output by decoding the fused features.

B. Orthogonal Block in Backbone for Feature Enhancement

The feature map based on traditional convolution has a large amount of redundant information and has the hidden features of strongly correlated channel slices. Therefore, in order to improve the efficiency of feature extraction, it is necessary to reduce redundant features. When the filters are orthogonal, they extract information from the orthogonal subspace of the feature space, focusing on different characteristics. Because the gradient information flows backward along the network, the convolution kernel in front of the squeeze filter can adapt to its unique mapping. Thus the network can extract a richer representation of each feature graph, and then the excitation can be built on this basis.

Based on the above reasons, this paper introduced the orthogonal attention mechanism in the search network to reduce the redundant information of the traditional convolution and improve the efficiency of feature extraction. The residual structure within this mechanism is depicted in Fig. 4. The residual block is crafted in a two-phase approach. The initial phase commences with the initialization of a random filter, which is initialized to the size of the input characteristic matrix$(H \times W \times C)$. After that, the $H \times W \times C$ filter is Schmidt orthogonalized to obtain a filter orthogonal along the channel direction.

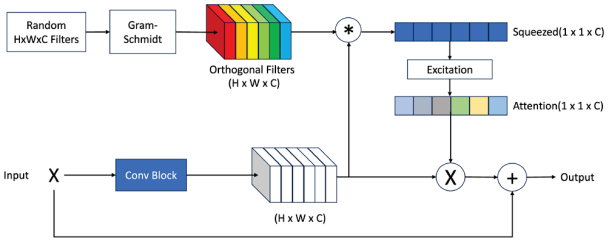

Fig. 4. **Residual block with orthogonal channel attention.** Compared to traditional attention mechanisms, orthogonal attention mechanisms can effectively reduce redundant information caused by traditional convolution.

The filter is orthogonalized through the Schmidt process to ensure the optimal transformation of linearly independent vector sets. Following this, a convolution operation is applied to the input, culminating in the extraction of a one-dimensional vector. This vector, when multiplied element-wise with the feature map, accentuates the salient features within the data. The resultant product is then integrated with the residual component through addition, culminating in the final output. This sequence of operations refines the feature representation, enhancing the module's capability to capture and emphasize the most informative aspects of the input. Due to its orthogonality, compared with the channel attention generated by the sequence and exception in SENet[20] and the frequency channel attention in FcaNet[21], the orthogonal channel attention can effectively reduce the representation of redundant information in the backbone. This enhanced feature extraction capability empowers the network to more effectively capture a richer and more nuanced representation of the feature landscape, thereby facilitating a more profound understanding of the underlying data.

C. CSPO-Fusion for better Feature Fusion

On the other hand, as the network's depth intensifies, so does its receptive field, potentially resulting in the diminished capture of fine-grained details. Generally, to enhance the detection performance for small targets, a notion might be to incorporate a fusion block that integrates the features from the S_2 detection layer, thereby augmenting the detection capabilities. However, this approach can engender a cascade of issues, including an escalated computational burden and prolonged post-processing times, as a consequence of incorporating the S_2 layer into the framework.

To address the aforementioned challenges, this paper introduced the CSPO-Fusion feature fusion module, an enhancement over the traditional fusion block architecture, specifically designed to integrate the shallow features extracted from the backbone network. This innovative module is tailored to amalgamate these features effectively, thereby optimizing the detection network's ability to process and retain critical micro information. In a departure from the conventional fusion block architecture, this paper pioneers the utilization of SPD Conv for the extraction of the S_2 feature layer, eschewing the standard convolutional approach. This method effectively integrates features laden with small target information with those from the S3 detection layer. Subsequently, we enhance it in conjunction with the branches of the CSPO-Fusion. This advancement not only fortifies the feature integration process but also augments the network's proficiency in capturing subtle details indicative of the small

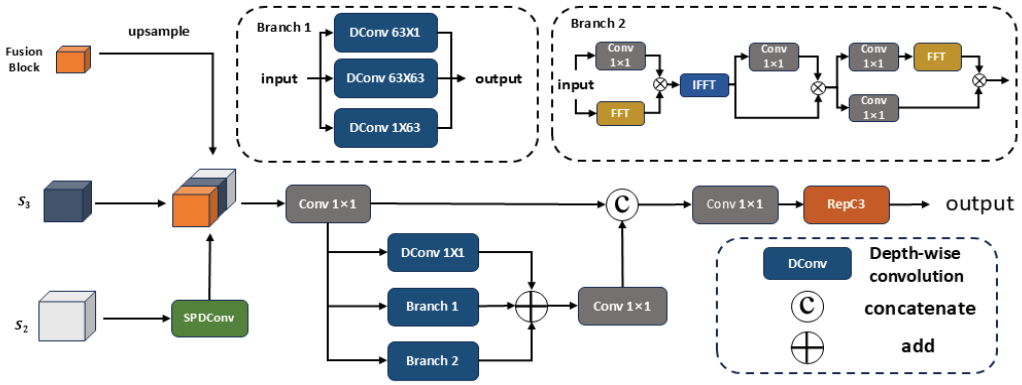

Fig. 5. **Structure of CSPO-Fusion**. The CSPO-Fusion integrates three levels of feature information: the previous layer Fusion Block, the feature extraction layer S_3, and the shallow layer information S_2. After concatenating the information from these three levels, the processed features from different branches are fused to obtain global to local information.

components. The module is constructed with a trifecta of branches. This design is designed to capture feature representations in a hierarchical manner, transitioning from global to local scales. This stratified learning approach is adept at distilling comprehensive feature insights, thereby significantly bolstering the detection performance. The framework of CSPO-Fusion is shown in Fig. 5.

Given the input feature $X \in R^{C\times H\times W}$, the feature map post 1×1 convolutional processing is directed into the local, large, and global branches. This orchestrated flow is designed to enrich the multi-scale feature representation, capturing a spectrum of spatial hierarchies. Ultimately, the feature map is subjected to a final 1×1 convolutional layer, culminating in the output feature graph, which encapsulates the distilled essence of the input across various scales.

Large Branch The large branch uses the depth convolution of $K \times K$ to achieve a large receptive field. At the same time, the branch also uses two depth volume integrators ($K \times 1$ and $1 \times K$) in parallel to obtain bar context information.

Global Branch The branch is composed of dual-domain channel attention module(DCAM) and frequency-based spatial attention module(FSAM). Given the branch input characteristic $X_{Global} \in R^{C\times H\times W}$, the DCAM module first processes the frequency channel attention mechanism (FCA) of X_{Global}, as follows:

$$X_{FCA} = \mathrm{IF}\left(F(X_{Global}) \otimes W_{1\times1}^{FCA}\left(GAP(X_{Global})\right)\right) \quad (1)$$

where F and IF respectively represent the fast Fourier transformer and its inverse transform, X_{FCA}, $W_{1\times1}^{FCA}$ 和 GAP respectively represent the output of FCA, 1 × 1 convolution layer and global average pooling, and ⊗ represents element by element multiplication through Fourier transform, the global features of the image are effectively refined according to the spectral convolution theorem. After being globally modulated, the obtained features are further fed to the spatial channel attention module (SCA):

$$X_{DCAM} = X_{FCA} \otimes W_{1\times1}^{FCA}\left(GAP(X_{Global})\right) \quad (2)$$

X_{DCAM} is the output of DCAM. DCAM only enhances the dual domain feature on the coarse granularity of channel mode. Then, we apply the frequency based attention module in the spatial dimension to refine the spectrum at the fine-grained level, which is formally expressed as:

$$X_{FSAM} = \mathrm{IF}\left(F\left(W_{1\times1}^{1}(X_{DCAM})\right) \otimes W_{1\times1}^{2}(X_{DCAM})\right) \quad (3)$$

Where X_{FSAM} is the result of FSAM. By doing so, the model notices the information frequency component.

Local Branch Local branches contain more information and play an important role in component detection. Therefore, it is necessary to design a local branch that uses a 1 × 1 depth convolution layer for local signal modulation.

D. IoU-Aware Size-Dependent Weighted Loss Function

To further enhance the detection performance across varying object scales in the TEDS system, we propose an improved loss function that integrates a size-dependent weighted mechanism with IoU-aware query selection. This design addresses two critical challenges: balancing the contribution of objects of different sizes during training and ensuring consistency between classification accuracy and localization precision.

Objects of different sizes contribute unequally to the training process of detection models. Large objects often contain richer feature information, which can aid in improving feature learning for both large and small objects. To harness this potential, we introduce a size-dependent weighting term based on the logarithm of the object area:

$$W_i = \log(h_i \times w_i + \epsilon) \quad (4)$$

where $h_i$and w_i are the height and width of the box, and ϵ is a small value for numerical stability. W_i is normalized across the batch to balance gradient contributions:

$$\alpha_i = \frac{W_i}{\sum_{k\in B_{batch}} W_k} \quad (5)$$

The normalized weight α_i ensures that large objects, which are generally fewer in number, contribute more effectively to the optimization process without overshadowing smaller objects.

To further refine the classification performance, we incorporate an IoU-aware query selection mechanism. This mechanism adjusts the classification scores by combining the predicted class probabilities with the IoU scores between predicted and ground truth bounding boxes:

$$\widehat{c_i} = c_i \cdot IoU\left(b_i, b_i^{gt}\right) \quad (6)$$

where c_i is the predicted class probability, b_i and b_i^{gt} represent the predicted box b_i and its corresponding ground truth box.

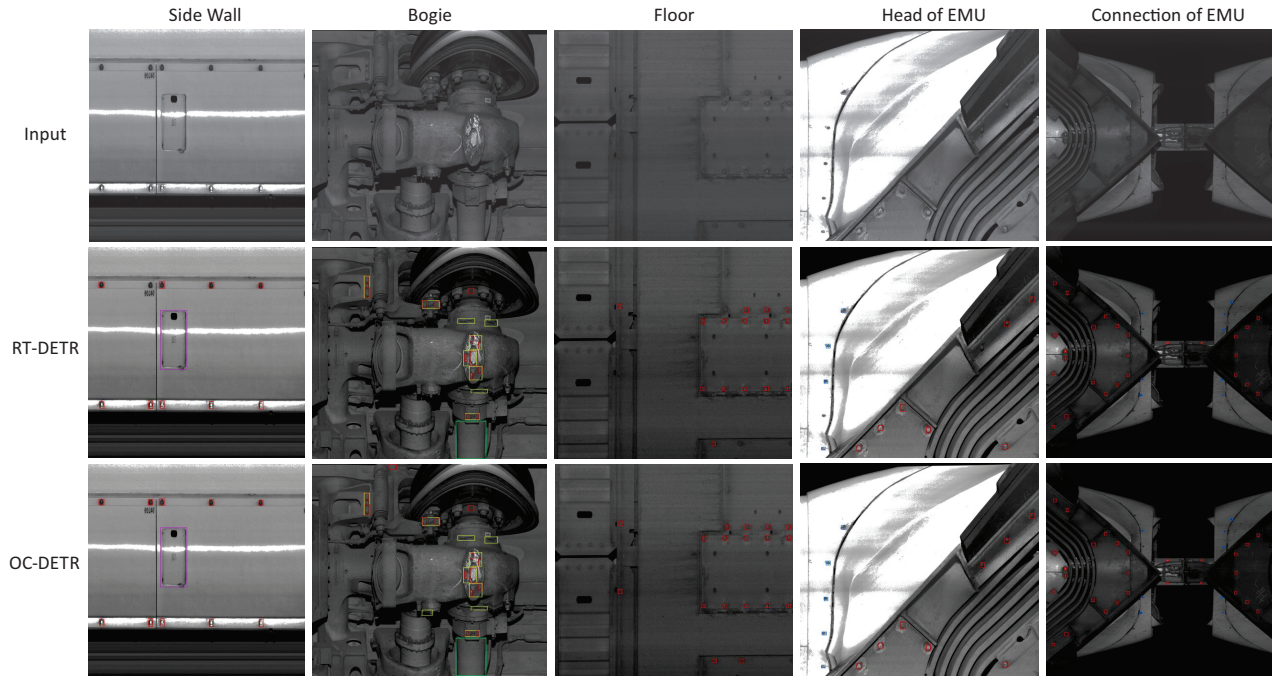

Fig. 6. **Comparison of detection results.** between RT-DETR and OC-DETR based on TEDS images. The visualization effect of the algorithm proposed in this article and the benchmark algorithm is shown in the figure. It can be found that the algorithm proposed in this article can better identify components and reduce the occurrence of missed detections.

The overall loss function integrates both mechanisms into the classification and localization branches, formulating a unified objective:

$$L(\hat{y}, y) = \sum_{i \in B_{batch}} \alpha_i \cdot \left[L_{cls}(\hat{c_i}, c_i) + L_{box}(\hat{b_i}, b_i^{gt}) \right] \quad (7)$$

where L_{cls} is the classification loss incorporating IoU scores, and L_{box} is the localization loss for bounding box regression.

The combined loss not only prioritizes large objects but also enforces consistency between the predicted classification scores and spatial alignment. The proposed loss function effectively balances the contributions of objects of various sizes while enhancing the alignment between classification and localization.

IV. Experiments

A. Experimental Dataset Description

We utilized TEDS devices to collect imagery information of high-speed trainsets, encompassing two models: CRH380 and CRH400AF. The curated dataset, after selection and annotation, comprises a total of 3881 high-resolution images, each with a dimension of 2048 x 2000 pixels, providing a wealth of visual information for our experimental analysis and training endeavors. It includes seven distinct types of components: hexagonal bolts, locking wires, side doors, cross bolts, cylinder locks, axle, and rubber sleeves. Fig. 7 illustrates the cropped images of the seven different components within the dataset. Additionally, we have tabulated the distribution of the pixel area occupied by the annotated targets in the dataset, as shown in Fig. 8, which displays the size distribution of all targets in the dataset, with target areas smaller than 50x50 pixels accounting for approximately 50%. Within the dataset, the same components are not fixed in position and are distributed across various locations, as depicted, covering imagery information from multiple perspectives, including the inner and outer sides of the undercarriage wheels and the sides of the trainset. Furthermore, the dataset is partitioned into training, validation, and testing sets at an 8:1:1 ratio for model training and comprehensive performance evaluation.

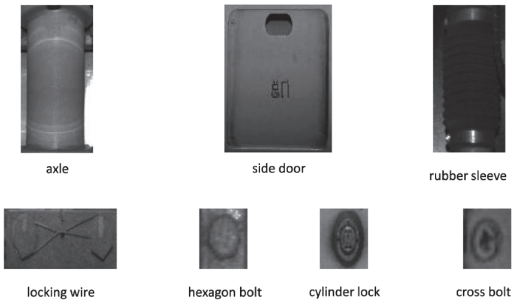

Fig. 7. Schematic diagram of different components in dataset

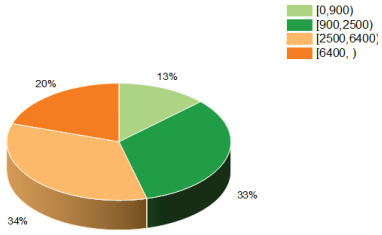

Fig. 8. Distribution of pixel area proportion of components

B. Model Training and Evaluation Metrics

The proposed model was implemented in PyTorch and deployed on a workstation with an NVIDIA 4090 GPU. Adam

TABLE I. EXPERIMENTAL RESULTS FOR COMPONENT OBJECT DETECTION ALGORITHM COMPARISON.

Method	Backbone	Params	mAP50	mAP50:95	GFLOPs	FPS(bs=1)
YOLOv5m[12]	CSPDarkNet53	20.87M	0.915	0.710	49.0	31.3
YOLOv7[14]	DarkNet53 (ELAN)	37.22M	0.926	0.722	105.2	28.5
YOLOv8m[15]	CSPDarkNet53	25.90M	0.922	0.732	79.1	32.2
YOLOv9m[16]	DarkNet53 (GELAN)	20.00M	0.920	0.721	76.3	36.1
RT-DETR[19]	Resnet18	**20.00M**	0.921	0.740	58.7	**43.9**
RTMDET[22]	CSPNeXt	24.71M	0.912	0.713	**39.1**	30.3
TOOD[23]	Resnet50	32.20M	0.905	0.688	39.2	33.3
OC-DETR(ours)	Resnet18	20.58M	**0.943**	**0.759**	60.2	40.1

W optimizer was used with initial learning rate 0.0001, momentum 0.937, and weight decay 0.0001 to learn the parameters. The batch size during training was set to 16 and the input image resolution standardized to 1280x1280 pixels. The training process iterated over 100 epochs to ensure thorough model convergence and stability. Mean average precision (mAP) is used as the standard evaluation metric, which can be divided into mAP50, mAP50:95, and so on, according to the different IOUs. Here, mAP50 and mAP50:95 are used as the main evaluation metrics.

C. Experimental Results

In the experimental study, the mean Average Precision (mAP) emerges as the pivotal metric for performance, offering the reflection of detection accuracy cross varying Intersection over Union (IOU) thresholds. Specifically, mAP50 and mAP50:95 are deployed as the primary benchmarks for evaluation, enabling a comprehensive analysis of the model's precision at both moderate and extensive IOU ranges. This approach ensures a detailed appraisal of the detection system's capability to identify objects under diverse spatial overlap conditions.

Fig. 6 presents the schematic of the detection results by OC-DETR. The figure shows the detection effects of different parts of the train, including the side wall, bogie, floor, head and connections. Through intuitive detection results, it can be seen that the algorithm proposed in this article has a significant improvement in performance compared to the baseline algorithm.

Table I delineates a comparative analysis between our proposed methodology and several prevailing object detection frameworks that dominate the current landscape. The empirical outcomes underscore that our enhanced algorithm while maintaining competitive detection velocities, achieves superior detection accuracy. Additionally, it exhibits efficient GFLOPs (Giga Floating Point Operations) relative to its contemporaries in object detection algorithms, thereby illustrating its robustness and efficacy in real-world applications.

D. Ablation Study

To analyze the importance of each component in OC-DETR, this paper extends an in-depth ablation experimental analysis on the proposed method, encompassing a suite of benchmark models such as RT-DETR benchmark model, the RT-DETR equipped with the CSPO-Fusion, the RT-DETR with orthogonal channel attention, and OC-DETR. In addition, to more intuitively demonstrate the effectiveness of the method proposed in this paper, we borrowed the feature fusion method of PANet that directly leads to the S_2 detection layer.

Table II shows the results of the ablation experiment, these analyses provide a granular understanding of the contributions and impacts of various components within the detection framework, thereby reinforcing the methodological advancements presented in this study.

TABLE II. EFFECT OF EACH COMPONENT. **CF**: CSPO-FUSION, **OA**: ORTHOGONAL CHANNEL ATTENTION, **LW:** IOU-AWARE SIZE-DEPENDENT WEIGHTED LOSS FUNCTION, **S2:** FEATURE FUSION METHOD OF PANET

Baseline	S2	OA	CF	LW	mAP50	mAP50:95	GFLOPs
✓					0.921	0.740	58.7
✓	✓				0.93	0.750	91.4
✓		✓			0.935	0.751	63.6
✓			✓		0.934	0.751	**57.0**
✓				✓	0.925	0.747	58.7
✓		✓	✓		0.941	0.756	60.2
✓		✓	✓	✓	**0.943**	**0.759**	60.2

It can be observed that introducing an additional standalone S_2 detection layer for the Fusion Block significantly enhances the detection accuracy of components. Yet, this approach markedly increases the model parameters and computational load. At the same time, the orthogonal channel attention mechanism and CSPO-Fusion module have significantly improved the method proposed in this paper. This is attributed to the channel redundancy compression of the orthogonal channel attention mechanism, which enables the backbone feature extraction network to extract richer representation information. At the same time, using the SPD Conv and CSPO-Fusion modules for small target feature extraction, the S_2 detection layer information is better integrated into the feature pyramid, greatly improving the detection accuracy of small targets such as high-speed train components. Compared with other models, the proposed method has great advantages in both accuracy and computational complexity. Finally, the integration of the IoU-

aware size-dependent weighted loss function further improves the performance of both RT-DETR and OC-DETR. The size-based weighting mechanism enables better learning of features from large objects, while the IoU-aware query selection aligns classification scores with localization precision, resulting in notable improvements in detection. The results demonstrate that the proposed module consistently and synergistically improves the performance of component detection.

V. Conclusion

This paper proposed a new real-time object detection framework called OC-DETR, specifically tailored for component detection in high-speed railway TEDS systems. The images collected by the TEDS system often have a large field of view, complex backgrounds, and significant differences in the size of different components. Therefore, this paper first proposed CSPO-Fusion to better integrate the features of components of different sizes, in order to improve the accuracy of component detection. At the same time, in order to solve the inherent redundancy problem in the features extracted using traditional convolution methods and reveal the potential features in strongly correlated channel slices, this paper replaces the original residual module with an orthogonal residual module, effectively improving the efficiency of feature extraction within the backbone and meeting the strict accuracy requirements of component detection tasks. To further improve detection performance across various object scales, an IoU-aware size-dependent weighted loss function was introduced. The loss function unified approach ensures balanced learning across small, medium, and large components, resulting in improvements in detection accuracy. This method, combined with the real TEDS collection dataset, can quickly and accurately detect various components. At the same time, this paper trained and evaluated other mainstream object detection models to facilitate comparative evaluation of test results. The experimental results show that this method has good performance in TEDS object detection and has great potential for application. This study not only meets the accuracy and real-time requirements of component detection, but also demonstrates the superiority of this method over other models in high-speed train component recognition.

Acknowledgment

This work was supported by Hubei Province Unveiling Science and Technology Project (2021BEC008). The research presented in this article has been significantly facilitated by the data and hardware support graciously provided by Wuhan Lisai Technology Co. and Beijing Railway Engineering Electromechanical Technology Research Institute Co., Ltd.

References

[1] B. Liu, "Research and thought on trouble of moving EMU detection system (TEDS)," *China Railway* , no. 12, pp. 61–65, 2017.

[2] H. Salman, C. Parks, M. Swan et al., "Orthonets: Orthogonal channel attention networks," in *2023 IEEE International Conference on Big Data (BigData)* , IEEE, 2023, pp. 829–837.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , 2016, pp. 770–778.

[4] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* , Cham: Springer Nature Switzerland, 2022, pp. 443–459.

[5] Z. Zou, K. Chen, Z. Shi et al., "Object detection in 20 years: A survey," *Proceedings of the IEEE* , vol. 111, no. 3, pp. 257–276, 2023.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , 2014, pp. 580–587.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems* , vol. 25, 2012.

[8] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* , 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[12] G. Jocher, "YOLOv5 by Ultralytics (Version 7.0)," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3908559

[13] C. Li, L. Li, H. Jiang et al., "YOLOv6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.

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

[15] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO (Version 8.0.0),"2023.[Online].Available: https://github.com/ultralytics/ultralytics

[16] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," arXiv preprint arXiv:2402.13616, 2024.

[17] N. Carion, F. Massa, G. Synnaeve et al., "End-to-end object detection with transformers," in *European Conference on Computer Vision*, Cham: Springer International Publishing, 2020, pp. 213–229.

[18] X. Zhu, W. Su, L. Lu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, 2020.

[19] W. Lv, S. Xu, Y. Zhao et al., "Detrs beat yolos on real-time object detection," arXiv preprint arXiv:2304.08069, 2023.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[21] Z. Qin, P. Zhang, F. Wu et al., "FCANet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3490–3499.

[22] C. Lyu, W. Zhang, H. Huang et al., "RTMDet: An empirical study of designing real-time object detectors," arXiv preprint arXiv:2212.07784, 2022.

[23] C. Feng, Y. Zhong, Y. Gao et al., "TOOD: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 2021, pp. 3490–3499.