2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Detection of EMU Components Based on Optical Flow Attention Prior and Multi-modal RGBD RTDETR*

Mingjun Cong[1], Gang Peng[1], *Member, IEEE*, Yongchang Tang[1], Chaowei Song[1], Chaoze Wang[1]

***Abstract*— To address challenges in high-speed train inspection such as complex backgrounds, diverse component types, and compact dimensions, this paper proposes a defect detection method called RTDETR-FAMC (RTDETR with Optical Flow Attention and Multimodal CSwin Transformer). The approach integrates RGB images and depth data through a dual-branch CSwin Transformer backbone network that fully utilizes both visual and depth information. At the same time, the improved Sea-RAFT optical flow estimation is combined to generate dynamic spatial prior attention for standard images and test images, so as to guide the network to focus on target regions. A Mask Feature Fusion (MFF) module achieves channel-space attention synergy optimization, while HWD wavelet transform downsampling and CSP-PAC multi-scale feature fusion modules enhance detection accuracy. Experimental results based on a self-built high-speed rail EMU fine-grained scanning dataset (containing 3,881 high-resolution images) demonstrate significant accuracy improvements compared to mainstream detection algorithms. Compared with YOLO series and standard RTDETR methods, the proposed approach achieves at least 3% improvement in mAP50 metric, validating its effectiveness as a reliable technical solution for intelligent EMU inspection.**

I. Introduction

In recent years, China's rail transportation sector has achieved remarkable progress, setting new records in both operational mileage and passenger/freight throughput. As of the first half of 2024, the country's total railway network reached 162,000 kilometers, with high-speed rail (HSR) infrastructure accounting for 47,000 kilometers, maintaining its position as the world's longest HSR network. During this period, China's railways transported 1.95 billion passengers and handled 2.08 billion tons of freight, setting new historical benchmarks. With the continuous development of China's HSR construction, maintenance requirements for EMU trains have become increasingly stringent. Current inspection methods for HSR trains primarily rely on manual approaches, but the complexity of maintenance tasks and high labor intensity often lead to operator fatigue, resulting in prolonged inspection durations, low efficiency, and significant safety risks. Among various maintenance solutions, the Train Electronic Data System (TEDS) [1] stands out as a critical tool for real-time monitoring. Through centralized analysis and multi-level application management supported by railway networks, TEDS enables visual inspection and fault detection of critical components on EMU trains' sides and undercarriage. To enhance inspection efficiency and safety, mobile robots installed beneath train chassis can capture high-resolution images using robotic arms, generating precise scanning datasets. AI-powered systems assist in accurately identifying potential faults, significantly reducing inspection time while minimizing manual workload, thereby ensuring safe and stable high-speed rail operations.

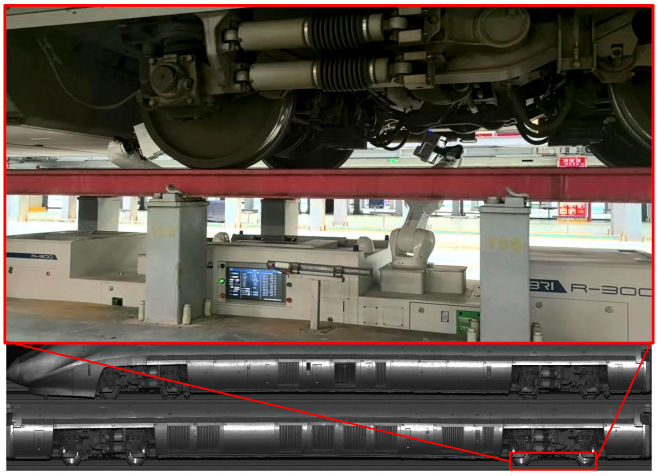

Fig. 1. On site photos of the inspection robots collecting chassis data of train sets.

As shown in Fig. 1, during maintenance of high-speed rail EMU trains in the garage, inspection robots move under the tracks to scan the lower sections. By parking at specific locations on the train and using robotic arms to photograph the chassis from different angles, a set of high-resolution depth-information-rich scanning images can be obtained. This facilitates subsequent identification and fault diagnosis of train components. Since the relative positions of parts are fixed when the EMU leaves the factory, the spatial structures captured through data collection at identical relative positions within the same type of carriage remain similar. The scanned data is presented as follows in Fig. 2.

This study aims to develop a high-precision component detector that utilizes scanned aligned depth maps and RGB images. To address the complex backgrounds, diverse component types, and varied defect categories in EMU chassis parts, while fully leveraging the features of both RGB and depth maps along with prior relationships between camera positions under identical shooting angles, we propose the **RTDETR-FAMC** defect detection network based on optical flow attention with standard priors and RGB-D multimodal dataset. The network integrates a dual-branch CSwin Trans-

*This research was supported by Hubei Province Unveiling Science and Technology Project (2021BEC008). The corresponding author is Gang Peng, e-mail address: penggang@hust.edu.cn

[1] Mingjun Cong, Gang Peng, Yongchang Tang, Chaowei Song and Chaoze Wang are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430000, China.

Fig. 2. A display of the dataset captured by the robot. The 1st and 3rd rows are depth maps, and the 2nd and 4th rows are the corresponding RGB images

former [2] backbone that combines RGB images and depth data, enhancing feature perception through multimodal data features and cross-shaped window attention from CSwin Transformer. Additionally, our method incorporates an improved Sea-RAFT [3] optical flow estimation to generate dynamic spatial prior attention, while designing a Mask Feature Fusion (MFF) module to achieve channel-space attention synergy optimization. This approach integrates prior information from standard maps for component localization. The introduction of HWD [4] wavelet transform down- sampling and CSP-PAC multi-scale feature fusion modules enhances feature interpretability, avoiding the need for traditional convolutional networks to stack multiple layers for multiscale perception. These innovations significantly improve detection accuracy and efficiency.

The main contributions of this article are as follows:

- In order to solve the problem of insufficient utilization of multimodal features, the RGB-D dual-branch CSwin Transformer architecture is used to jointly analyze visual texture and depth geometric information, so as to enhance the discrimination ability of moving train chassis component positions.
- To address the issue of spatial prior loss, this study proposes a standard graph-test graph attention mapping mechanism based on Sea-RAFT optical flow. By leveraging the fixed-viewpoint feature characteristics to generate dynamic spatial prior masks, the network is guided to identify potential defect regions. Additionally, an MFF module is designed to integrate optical flow attention with channel attention, thereby enhancing the system's capability to utilize both feature textures and spatial priors.
- In order to solve the problem of low detection efficiency and improve detection accuracy, HWD-CSP-PAC neck network is designed. Wavelet transform is used to explicitly separate high frequency and low frequency features, and parallel hole convolution is used to capture multi-scale context, so as to enhance the perception ability of target regions.

II. Related Work

Current deep learning-based object detection methods are primarily categorized into two types: two-stage detectors like the Fast R-CNN series, and single-stage models such as the YOLO and DETR [5] families. As pioneering research, R-CNN [6] adopted a two-stage detection process: first generating candidate regions through selective search algorithms, then extracting features via pre-trained AlexNet [7], and finally classifying them using support vector machines. The Fast R-CNN [8] achieved end-to-end training by improving unified feature extraction and multi-task loss functions, significantly enhancing detection accuracy. Faster R-CNN [9] introduced Region Proposal Networks (RPN) that share weights with the main network, enabling rapid and accurate candidate region generation. The YOLO series, as representatives of single- stage detectors, simplified the detection process by directly predicting object boxes on convolutional features, making it particularly suitable for industrial real-time applications. YOLOv2 [10] enhanced detection accuracy through anchor box mechanisms, YOLOv3 [11] implemented multi-scale detection head designs, YOLOv4 proposed feature fusion neck structures based on FPN and PAN, while YOLOv5 [12], YOLOv6 [13], YOLOv7 [14], YOLOv8 [15], and YOLOv9 [16] further optimized performance and speed.Recently, YOLO's version has been updated rapidly, with YOLOv10 [17], YOLOv11 [18], YOLOv12 [19] appearing one after another. Among them, the C2PSA module proposed by YOLOV11 is also used in this article.

In recent years, significant progress has been made in multi-modal object detection, focusing on cross-modal fea-

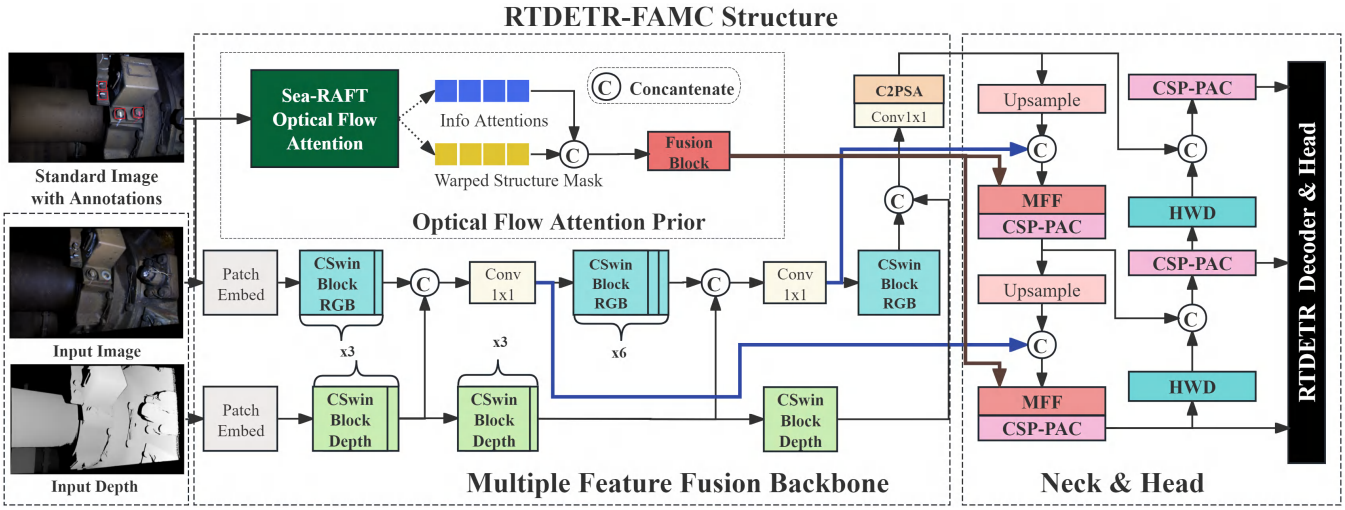

Fig. 3. RTDETR-FAMC Structure.

ture fusion, open-vocabulary detection, and 3D perception. GGSTrack [20] proposed a geometric graph convolutional network (GCN) and spatio-temporal convolution-based multi-object tracking method, significantly improving performance in dense scenes. MQ-Det [21] introduced a multi-modal query mechanism (text + visual examples), achieving a 7.8% accuracy improvement in open-world object detection while resolving ambiguity in text queries. For cross-modal domain adaptation, the illumination-aware feature adaptation framework (IAM+PPA) [22] employed progressive prototype alignment, increasing pedestrian detection mAP by 12.6% in visible-thermal imaging. In 3D object detection, PPF-Det [23] proposed a point-pixel fusion strategy with multi-pixel perception (MPP) and point-voxel-wise triple attention (PVW-TAF), enhancing LiDAR-camera fusion robustness. SuperYOLO [24] leveraged super-resolution-assisted training and pixel-level multi-modal fusion, outperforming YOLOv5 by 10% mAP in remote sensing small object detection.

III. Optical Flow Attention Prior and Multi-modal RGBD RTDETR

A. Overall Network Structure

The proposed RTDETR-FAMC network (RTDETR with Optical Flow Attention and Multimodal CSwin Transformer) , which is shown as Fig. 3, is built upon three key components: test image inputs, depth maps corresponding to test images, and fully annotated standard maps. Given that each camera's viewpoint remains fixed during EMU chassis inspections, the spatial structures of target boxes captured by each camera should maintain similarity across all frames. This framework enables the creation of standardized maps for each camera position, which comprehensively include all object categories within the view while ensuring no missing or damaged elements. For test images, their acquisition locations correspond to identical positions on other trains, with depth maps derived from LiDAR data aligned with the test images.

The backbone network is a multi-feature fusion architecture with dual branches, each processing RGB and depth images respectively. Each branch contains multiple CSwin Transformer Blocks that extract features at different levels, while 1x1 convolution modules integrate these features. Given that RGB input serves as the primary feature source, the network design prioritizes the RGB branch to ensure its dominance in final feature fusion. The Depth branch provides auxiliary information through its CSwin Transformer Block following each Depth branch, enhancing target recognition accuracy. Finally, the output features undergo C2PSA module processing with self-attentional correction. This architecture enables comprehensive image content understanding, thereby improving detection performance.

The system employs a standard image with complete annotations and a test image input to activate another optical flow estimation branch. This branch utilizes the backbone network of Sea-RAFT's optical flow estimation as the standard-to-test optical flow mapping. The mask matrix generated from annotated bounding boxes in the standard image is deformed into attention mask information for the test image through optical flow results. This information is then fused with optical flow outputs and marker attention from Sea-RAFT to generate a dynamic spatial prior attention map, which guides the network to focus on target regions. Following this, the Fusion Block normalizes the attention space to produce a fused optical flow sampling attention tensor, providing essential attention guidance for subsequent feature fusion processes.

The Neck module retains the multi-scale feature fusion architecture from the YOLO series, employing both upsampling and downsampling branches. Specifically, after each upsample, the Mask Feature Fusion Block (MFF) integrates optical flow spatial structure priors with backbone network features. This block combines channel attention mechanisms that leverage optical flow spatial priors with spatiotemporal attention patterns, while integrating features

from the backbone network through a HSPFN module to fine-tune feature maps and enhance target region saliency. During the downsampling phase, the HWD module utilizes Harr wavelet transform to extract high-frequency and low-frequency components for feature fusion. The Neck module employs the CSP-PAC module for cross-stage local connections, which enhances feature representation capabilities through enhanced feature expression.

In the head part, the detection head adopts RTDETR encoder and multi-scale feature map to optimize target positioning through dynamic decoding strategy to ensure detection accuracy.

B. RGB & Depth Dual Branch Architecture Feature Fusion

The CSwin Transformer adopts a hierarchical pyramid architecture with four stages, each implementing downsampling through Merge Blocks. Its core innovation lies in Cross-Shaped Window Attention, which breaks down traditional square window attention into horizontal and vertical strip windows. By alternately processing these two orientations, the network achieves full spatial coverage within a single layer, effectively addressing the challenge of global interaction required by Swin Transformer's multi-layer shift windows.

Following the completion of backbone network feature extraction, we introduce a C2PSA module. The C2PSA (Channel-to-Pixel Self-Attention) module integrates convolutional and self-attention mechanisms to enhance features through a dual-branch architecture that synergistically extracts both local details and global contextual information, which is first used in YOLOv11.

After concatenating the features from two branches in the channel dimension, the original number of channels is restored through a 1×1 convolution to achieve efficient feature fusion. This design retains the local perception advantage of traditional convolutions while incorporating the global modeling capability of self-attention mechanisms. By implementing a channel partitioning strategy, it effectively controls computational complexity, making this approach adaptable for various dense prediction tasks.

C. Spatial Structural Prior Based on Sea-RAFT Optical Flow Attention

Sea-Raft is an enhanced optical flow estimation method designed to improve efficiency, accuracy, and generalization capabilities. Its core improvements include the introduction of the Mixed Laplacian Loss (MoL), which maximizes the likelihood of true optical flow by predicting parameters of the Laplacian distribution. In terms of model architecture, Sea-Raft simplifies the original RAFT design by adopting a standard ResNet [25] as the backbone network and replacing ConvGRU with ConvNeXt [26] blocks. Sea-Raft demonstrates outstanding performance across multiple benchmarks including Spring [27], Sintel [28], and KITTI [29].

Fig. 4 shows some optical flow results obtained by sea raft on standard and test charts. Since the camera positions of the standard image and the test image are the same, the obtained

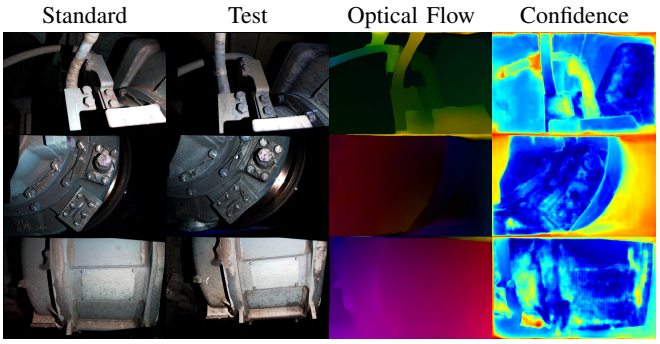

Fig. 4. Flow and confidence between the standard image and the test image by Sea-RAFT.

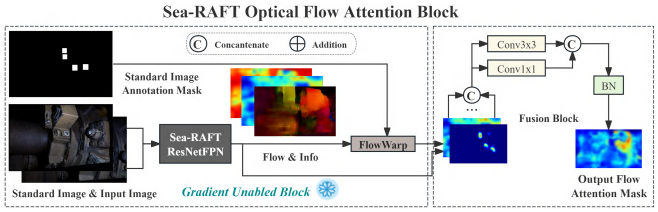

Fig. 5. Sea-RAFT optical flow attention block. The weight of Sea-RAFT backbone is frozen during training.

image has the same structure. To reduce computational load, we applied two downsampling stages to the input standard image I_1 and test image I_2 (both $\mathbb{R}^{3\times640\times640}$), as shown in Fig. 5. Before performing optical flow estimation on Sea-raft, the input underwent two downsampling processes. During the iteration phase, the RNN unit only operated once to obtain preliminary optical flow estimates. For the output stage, we not only captured optical flow information but also integrated uncertain estimates from the network outputs, including both position and scale uncertainties. Using annotated bounding boxes from the standard image, we generated multi-scale target box position masks and performed deformation integration through the network's optical flow output to preliminarily predict target box positions in test images. The final output typically consists of $\mathbb{R}^{B\times(6+N)\times160\times160}$. Simultaneously, we applied Gaussian blurring to this mask and combined it with other network outputs before feeding it into a Fusion Block. This dual-branch convolutional architecture employs 1×1 and 3×3 convolutions with BN modules to generate the final attention space prior mask tensor for optical flow sampling. Since the optical flow estimation prior ResNetFPN module requires extensive data training, its gradients are disabled during network training, serving solely as an initial optical flow prior mask.

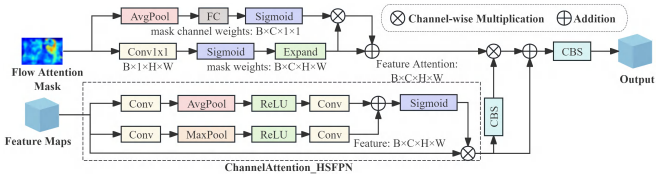

Fig. 6. Mask Feature Fusion Block.

Then, the Mask Feature Fusion (MFF) module is added to the neck feature fusion section. As an efficient dual-mode feature fusion module, this module combines channel attention and spatial attention mechanisms to adaptively fuse optical flow prior mask information with visual features (Feature Map), which significantly improves the feature discrimination capability in target detection tasks.

As shown in Fig. 6, this module employs a channel attention HSPFN module for input feature maps, utilizing dual-branch feature extraction with average and max pooling operations to perform weighted fusion of fine-grained features. The input optical flow prior mask undergoes dual-branch feature aggregation, focusing on both channel attention and spatial attention while refining the input features through weighted processing. For the optical flow attention mask, adaptive alignment is applied first to ensure spatial dimension compatibility with the feature map. In the secondary MFF modules at the neck layer, the optical flow prior masks are resized at ratios of 2, 4, 4, and 2 respectively. To align with input feature map dimensions, the optical flow attention mask is split into spatial and channel weights, followed by adaptive adjustment via a residual module. This design preserves structural information in the mask, automatically focuses on target key regions through attention mechanisms, effectively suppresses background interference, and fully utilizes optical flow prior information to enhance recognition accuracy.

D. Modules in the Neck

The HWD wavelet transform module adopts Haar wavelet decomposition. This explicit frequency-domain decomposition enhances feature interpretability and avoids the need for stacking multiple layers in traditional convolutional networks to achieve multiscale perception. The module only requires lightweight 11 convolutions for channel compression during feature fusion, resulting in substantially fewer parameters compared to conventional multi-scale modules.

For head feature extraction, the CSP-PAC module is utilized, which is a refined bottleneck structure integrating Cross Stage Partial (CSP) architecture with Parallel Atrous Convolution (PAC) [30]. This design efficiently extracts multi-scale features while maintaining computational efficiency, which is shown as Fig. 7.

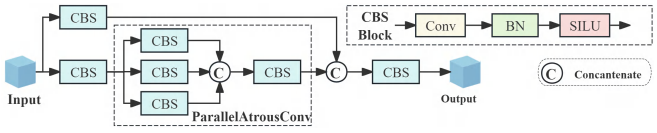

Fig. 7. CSP-PAC Block.

The module first divides the input channels into two parts through two 1×1 convolutions. One part is processed by an innovative ParallelAtrousConv sub-module, while the other retains the original features as shortcut connections. The ParallelAtrousConv sub-module employs three different dilation rates (default ratio=$[1, 2, 3]$) for parallel 3×3 hollow convolutions to capture features from different receptive fields. Standard convolution focuses on local details, large dilation rate convolutions capture wide-area context, and finally, a 1×1 convolution merges multi-scale features while adjusting the number of channels.

This design enhances the network's perception of multi-scale targets, making it suitable for visual tasks involving significant scale variations. The feature concatenation operation at the module's end combines processed multi-scale features with original features, then integrates information through a final 1×1 convolution. This approach retains the gradient flow optimization characteristics of the CSP architecture while enhancing multi-scale representation capabilities via parallel hollow convolutions.

IV. Experimental Analysis

A. Construction of the Dataset

In the dataset construction phase, inspection robots were deployed to capture imaging data from two high-speed train models: CRH380 and CRH400AF. The curated dataset comprises 3,044 high-resolution images (1920×1200 pixels each), providing rich visual resources for experimental analysis and training, which is shown as Fig. 8.

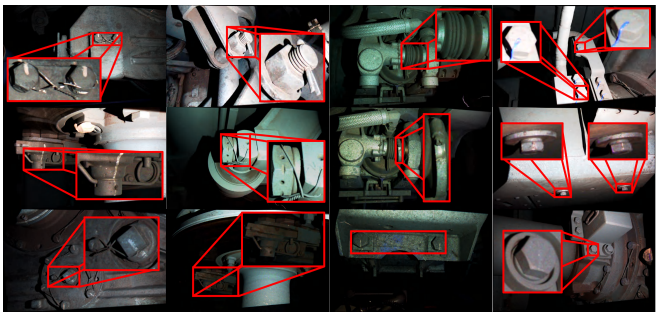

Fig. 8. Dataset components and their presentation.

These robot-acquired images feature 28 distinct camera positions capturing 34 types of components (including bolts, tail pins, clamps, throat clamps, dust covers, anti-loosening wires, brake calipers, axles, stone-sweeping machines, etc.), all accompanied by corresponding depth maps. During model labeling, all images from each camera position are organized into 28 separate folders, with each containing one standard image and test images. Standard images must contain complete annotation information. For example, if a camera position has four bolts, only images with full bolt annotations should be selected as standard reference images.

The network's input consists of three-channel test image inputs, corresponding depth maps for single-channel tests, and fully annotated 3D standard images. The annotated bounding boxes are converted into single-channel masks, resulting in an input dimension of $\mathbb{R}^{B\times 8\times 640\times 640}$. For data with identical viewing angles, a joint cross-training verification mechanism using standard images and test images can be employed. This method alternates between using specific images from the same viewing angle as standard images and others as test images, effectively leveraging existing dataset information to achieve optimal training performance, which is shown as Fig. 10.

TABLE I

RTDETR-FAMC AND OTHER METHOD COMPARISON RESULTS

Method	Backbone	Parameters	mAP50	mAP50:95	FPS
YOLOv5-l	CSPDarkNet53	53.2M	0.916	0.735	**109.2**
YOLOv8-l	CSPDarkNet53	43.7M	0.920	0.737	97.4
YOLOv9-e	DarkNet53	55.5M	0.923	0.738	69.7
YOLOv10-x	EfficientRepV3	30.3M	0.916	0.734	90.9
YOLOv11-x	CSPNet	54.3M	0.928	0.740	79.2
YOLOv12-x	R-ELAN	56.5M	0.929	0.739	65.1
RT-DETR-x	HGBNet	64.3M	0.929	0.741	50.2
RTDETR-FAMC(ours)	CSwin-transformer	46.2M	**0.952**	**0.766**	22.9

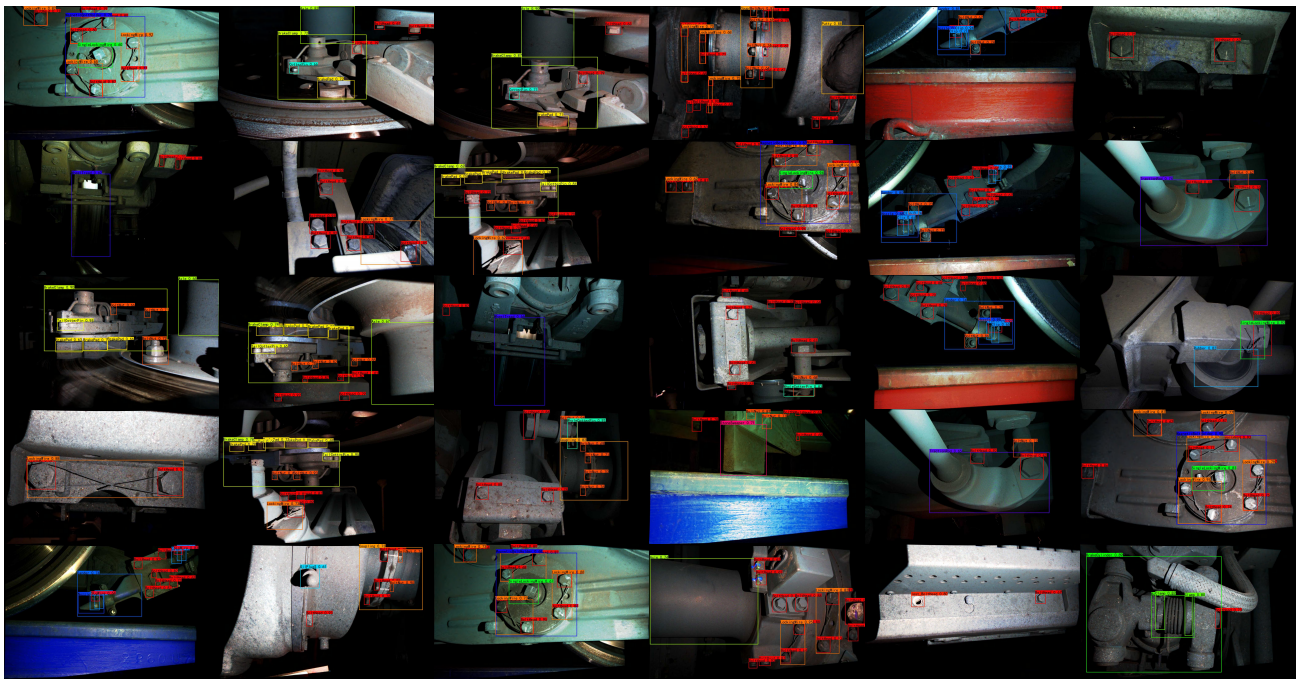

Fig. 9. Example of RTDETR-FAMC test results, which include a representative part of all 34 types of EMU chassis components.

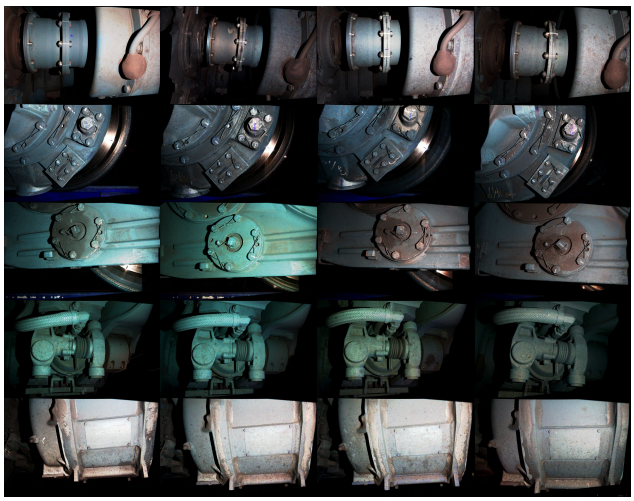

Fig. 10. Standard diagram-test diagram joint cross training verification mechanism schematic diagram. Each line represents the images taken at the same camera location in different carriages, which can be used as standard and test images for each other.

B. Comparative Test of Different Models

The dataset was divided into training, validation, and test sets in an 8:1:1 ratio for model training and comprehensive performance evaluation. The proposed model was implemented in PyTorch and deployed on a workstation equipped with NVIDIA 4090 GPUs. During training, the Adam optimizer was employed with an initial learning rate of 0.0001, momentum coefficient of 0.937, and weight decay of 0.0001. The batch size was set to 16 during training, and all input images were uniformly resized to 640×640 pixels. The model underwent 100 iterations to ensure sufficient convergence and stability. In this experiment, mean average precision (mAP) was adopted as the primary evaluation metric. This metric was categorized into levels such as mAP50 and mAP50:95 based on different intersection-over-union (IOU) ratios, with mAP50 and mAP50:95 serving as the core evaluation criteria of this study.

The experimental results are shown in Tab. I. For the comparative models, we selected the mainstream object detection networks from recent years: the YOLO series and RTDETR network. The YOLO series includes versions

from v5 to v12, with large-scale YOLO networks chosen to ensure comparable computational load. As evidenced by the table, our proposed method significantly outperforms other networks in both mAP50 and mAP50:95 metrics. In the field of FPS (Frame Per Second), the proposed method achieves near real-time detection despite its relatively low frame rate. During robotic data collection, the system only requires positioning the robotic arm at fixed points for data acquisition and inspection, without demanding high frame rates. Thus, the model developed in this study fully meets the requirements for inspection tasks.

Fig. 9 shows the test results of the network in the data set. It can be seen that each category participating in the training can be accurately identified under different camera positions.

C. Ablation Experiment

To validate the effectiveness of the proposed RTDETR-FAMC model, the following ablation experiments were conducted, which is shown in Tab. II. The backbone network was compared with Swin-transformer [31] and CSwin-transformer. Then replace the CSP-PAC module in the neck with Conv (standard convolution + batch normalization + activation). For other module designs, control variables were employed to experimentally investigate optical flow attention and channel feature extraction branches. The depth branch representation removal refers to eliminating the depth input modalities and 1×1 convolution feature fusion module in the original RTDETR-FAMC network. The FA-MFF removal involves removing both optical flow attention and MFF modules, where the CSP-PAC module is directly connected to the upsampling module.

TABLE II
RTDETR-FAMC ABLATION EXPERIMENT RESULTS

Backbone	Neck	Depth Branch	FA+ MFF	mAP50	mAP50:95
CSwin-T	CSP-PAC	√	√	**0.952**	**0.766**
CSwin-T	Conv	√	√	0.942	0.758
Swin-T	CSP-PAC	√	√	0.949	0.761
CSwin-T	CSP-PAC	√		0.924	0.731
CSwin-T	CSP-PAC		√	0.938	0.754
CSwin-T	CSP-PAC		Non-Boxes	0.923	0.739
CSwin-T	CSP-PAC			0.919	0.712
Sea-RAFT				0.813	

The ablation experiment results are presented in the table above, which displays experimental outcomes from different module modifications on the RTDETR-FAMC network. The comparison between Swin-transformer and CSwin-transformer in the backbone network demonstrates that CSwin-transformer achieves higher accuracy with reduced computational costs. When the depth mode is removed without modifying the backbone network, accuracy decreases. Using the CSP-PAC module instead of the standard Convolutional Module can significantly improve accuracy while keeping other parameters unchanged. Notably, eliminating the optical flow attention FA module and Mask Feature Fusion (MFF) module leads to significant performance degradation when no prior optical flow attention is applied. This loss of prior knowledge reduces the network's sensitivity to component initial positions. Additionally, a baseline experiment with default 7 input channels showed substantial performance deterioration when removing annotated masks and optical flow mapping matrices, confirming that incorporating standard graph annotation information provides crucial prior knowledge for network recognition.

Furthermore, since each camera perspective provides a separate standard image, we can fully utilize optical flow for conventional object detection (as indicated in the last row of the table, where mAP50 and mAP50:95 use fixed mAP values). The implementation involves processing Sea-RAFT network inputs with both standard images and test images. Here, we only perform $2\times$ downsampling on the input. The obtained optical flow is then used to remap the annotated bounding boxes from the standard image into the test image as the output result. While this method generates detectable bounding boxes, it performs poorly in detecting small targets, large targets, objects with significant rotation angles, or images with substantial exposure variations. Both large and small targets may result in misaligned bounding boxes. Particularly for bolt detection, missing components could lead to category errors when directly using optical flow (optical flow registration will identify the missing bolt as a normal bolt for example), which significantly increases false alarm rates and adversely affects subsequent fault diagnosis. The specific detection results are shown in Fig. 11.

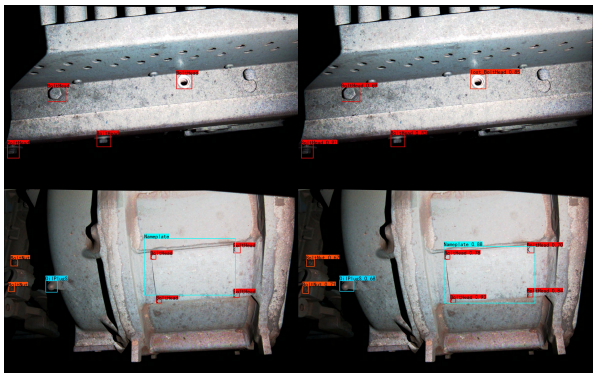

Fig. 11. The left column shows the results using only Sea RAFT, while the right column shows the results using **RTDETR-FAMC**. Using only Sea RAFT will transfer the category of the standard drawing to the test drawing, resulting in identifying the category of missing bolts as normal bolts; Sea RAFT can cause a deviation in detection results for large targets such as Nameplate.

V. CONCLUSION

This paper addresses challenges in high-speed train inspection, including complex backgrounds, diverse component types, and compact dimensions, by proposing a defect detection method called **RTDETR-FAMC** (RTDETR with Optical Flow Attention and Multimodal CSwin Transformer). The approach integrates RGB images and depth data through a dual-branch CSwin Transformer backbone network, combines improved Sea-RAFT optical flow estimation to generate dynamic spatial prior attention, de-

signs a Mask Feature Fusion (MFF) module for channel-space attention synergy optimization, and incorporates HWD wavelet transform downsampling with CSP-PAC multi-scale feature fusion modules, significantly enhancing detection accuracy and efficiency. Experimental results demonstrate that RTDETR-FAMC achieves excellent performance on a self-built high-speed rail EMU fine-grained scanning dataset, with mAP50 and mAP50:95 reaching 0.952 and 0.766 respectively, outperforming mainstream target detection models like the YOLO series and RT-DETR. The proposed method integrates visual texture and depth geometric information via a dual-branch CSwin Transformer backbone network, utilizes enhanced Sea-RAFT optical flow estimation to generate dynamic spatial prior attention guidance for target region focus, and combines MFF module for channel-space attention synergy optimization. The introduction of the HWD-CSP-PAC wavelet transform multi-scale feature fusion module proposed in this paper further improves the detection performance. In summary, RTDETR-FACT performs well in the EMU target detection task, providing reliable technical support for intelligent inspection of high-speed rail EMUs.

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*, vol. 12, pp. 61–65, 2017.

[2] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134.

[3] Y. Wang, L. Lipson, and J. Deng, "Sea-raft: Simple, efficient, accurate raft for optical flow," in *European Conference on Computer Vision*. Springer, 2024, pp. 36–54.

[4] G. Xu, W. Liao, X. Zhang, C. Li, X. He, and X. Wu, "Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation," *Pattern recognition*, vol. 143, p. 109819, 2023.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[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] A. Farhadi, J. Redmon *et al.*, "Yolov3: An incremental improvement," in *Computer vision and pattern recognition*, vol. 1804. Springer Berlin/Heidelberg, Germany, 2018, pp. 1–6.

[12] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, J. Poznanski, L. Yu, P. Rai, R. Ferriday *et al.*, "ultralytics/yolov5: v3. 0," *Zenodo*, 2020.

[13] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie *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] R. Sapkota and M. Karkee, "Ultralytics yolo evolution: An overview of yolo26, yolo11, yolov8 and yolov5 object detectors for computer vision and pattern recognition," *arXiv preprint arXiv:2510.09653*, 2025.

[16] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *European conference on computer vision*. Springer, 2024, pp. 1–21.

[17] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.

[18] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements. arxiv 2024," *arXiv preprint arXiv:2410.17725*, 2024.

[19] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[20] S. Yan, Z. Wang, Y. Huang, Y. Liu, Z. Liu, F. Yang, W. Lu, and D. Li, "Ggstrack: Geometric graph with spatio-temporal convolution for multi-object tracking," *Neurocomputing*, p. 131234, 2025.

[21] Y. Xu, M. Zhang, C. Fu, P. Chen, X. Yang, K. Li, and C. Xu, "Multi-modal queried object detection in the wild," *Advances in Neural Information Processing Systems*, vol. 36, pp. 4452–4469, 2023.

[22] Q. Xie, T.-Y. Cheng, Z. Dai, V. Tran, N. Trigoni, and A. Markham, "Illumination-aware hallucination-based domain adaptation for thermal pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 315–326, 2023.

[23] G. Xie, Z. Chen, M. Gao, M. Hu, and X. Qin, "Ppf-det: point-pixel fusion for multi-modal 3d object detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5598–5611, 2024.

[24] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[25] 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.

[26] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16 133–16 142.

[27] R. Piedrafita, R. Béjar, R. Blasco, A. Marco, and F. J. Zarazaga-Soria, "The digital connectedearth: Open technology for providing location-based services on degraded communication environments," *International Journal of Digital Earth*, vol. 11, no. 8, pp. 761–782, 2018.

[28] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European conference on computer vision*. Springer, 2012, pp. 611–625.

[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[30] Z. Zhang, F. Yu, Q. Ge, B. Zhao, J. Sun, and X. Li, "Parallel asymmetric dilated convolution module," *Computer Systems and Applications*, vol. 30, no. 9, pp. 206–211, 2021.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.