applied sciences

MDPI

Article

UG-Net: An Unsupervised-Guided Framework for Railway Foreign Object Detection

Zhuowen Tian and Jinbai Zou *

School of Railway Transportation, Shanghai Institute of Technology, Shanghai 201400, China; zhuowen0307@163.com
* Correspondence: zoujb@sit.edu.cn

Abstract

Foreign object intrusion severely threatens railway safety. Existing methods struggle with open-set categories, high annotation costs, and poor label-efficient generalization. To address these issues, we propose UG-Net, an unsupervised-guided label-efficient detection framework. The core idea is a two-stage strategy: first, a masked autoencoder (MAE) learns "normality" priors from unlabeled data and generates a spatial attention mask via a deep feature difference strategy; then, this mask is fused as a fourth channel into a lightweight YOLOv8n detector. This approach effectively alleviates reliance on manual annotations. On a self-constructed railway dataset, UG-Net achieved 94.56% mAP@0.5 using only 200 labeled samples, significantly outperforming the YOLOv8n baseline (86.91%). The framework provides a label-efficient solution for industrial anomaly detection.

Keywords: railway foreign object detection; unsupervised learning; masked autoencoder (MAE); YOLOv8; transfer learning; deep feature difference

1. Introduction

1.1. Background Introduction

However, the intrusion of foreign objects into the railway clearance gauge remains a prominent hazard threatening operational safety. These foreign objects encompass a wide variety of items: metal debris such as bolt fragments and aluminum cans left on the rail head; irregular ballast stones invading rail gaps; wind-blown lightweight materials like fabrics, kites, and plastic bags; and pedestrians or animals crossing the tracks. Despite significant differences in morphology and scale, all pose potential safety risks, ranging from minor incidents such as the jamming of braking systems to catastrophic derailments that can trigger large-scale service suspensions and mass casualties.

According to the Global Railway Safety Indicators Report published by the International Union of Railways (UIC) in 2023, collisions between trains and obstacles resulted in 1107 operational interruptions globally between 2015 and 2023. This accounts for 54% of all accidents caused by external factors, with an average annual growth rate of 7%. In terms of accident nature and impact, collisions caused by the direct intrusion of humans and animals are the primary cause of severe safety incidents. Meanwhile, small-target foreign objects with pixel dimensions smaller than 32 × 32—such as bolt fragments, fine gravel, and metal debris—have become the main source of "latent faults" due to their high concealment. These objects are frequently missed during routine inspections. Once trapped in the gap between the wheel and the rail, they not only accelerate wheel-rail

Check for updates

Received: 29 November 2025
Revised: 28 December 2025
Accepted: 31 December 2025
Published: 9 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

wear and shorten equipment lifespan but can also easily trigger operational disruptions, significantly impacting railway transport efficiency.

1.2. Challenges and Bottlenecks of Existing Methods

To mitigate the risks of foreign object intrusion, traditional solutions have been employed but face significant limitations. Manual inspection suffers from short daily coverage distances and high miss rates for small targets. Infrared detection often fails to identify materials with low thermal conductivity, such as plastics and fabrics. LiDAR systems, while effective, are hindered by high unit costs, making large-scale deployment economically unfeasible. Given these inherent limitations, automated foreign object detection based on computer vision and deep learning has become the inevitable trend for railway safety inspection, owing to its low cost and high real-time performance.

Although existing automated detection methods have made some progress, they still face three core challenges in real-world railway scenarios, which hinder their practical application.

The Dilemma of label-efficient Annotation Railway foreign objects exhibit "open-set" characteristics, with categories being nearly infinite—ranging from 0.5 cm bolt fragments to 1 m woven bags, and from crushed stones and aluminum cans to animals and humans. However, in actual inspection scenarios, the proportion of anomalous samples is extremely low; in natural scenes, frames containing anomalies account for less than 0.5%. Consequently, the cost of annotating effective samples is prohibitively high. It is estimated that annotating 500 effective images covering 8 categories of foreign objects requires 3 professional annotators working continuously for 10 days. If rare categories (e.g., waste tire fragments, metal wires) are included, the cost grows exponentially. This contradiction between "infinite categories" and "scarce samples" makes traditional supervised learning methods expensive and unable to cover all object types. When trained on large amounts of easily collected but repetitive data, these models are prone to overfitting and suffer from poor generalization capabilities.

Clarification on Terminology: It is important to note that in the context of generic computer vision, "few-shot learning" typically refers to the N-way K-shot strategy (e.g., learning from 1 to 5 samples). However, in this study, we use the term in the context of industrial anomaly detection to describe a "label-efficient" or "low-data" regime. specifically, we refer to a scenario where the model is fine-tuned using a drastically reduced dataset (e.g., 200 samples, representing approx. 10% of the standard requirement) compared to the thousands of annotations typically needed for fully supervised industrial detectors. To avoid ambiguity, we will predominantly use the terms "label-efficient" or "low-data training" throughout the manuscript, reserving "few-shot" strictly for the fine-tuning experimental setup.

1.2.1. The Intrinsic Difficulty of Detection Tasks

Foreign object detection in railway scenes must simultaneously address three technical hurdles: feature dilution of small targets, class confusion, and the coexistence of static and dynamic targets.

Feature Dilution: Small object miss rates are severe. From a drone inspection perspective, key objects like bolts and fine gravel often have pixel dimensions smaller than 32×32. For instance, a rail head bolt may occupy only 16×24 pixels. After downsampling by a CNN backbone, feature information loss exceeds 60%, resulting in an mAP generally lower than 45% for such targets in traditional detectors.

Class Confusion: Distinguishing anomalies from the background is difficult. For example, the grayscale mean difference between "anomalous stones" and "normal ballast"

is only 8%, and local texture overlap reaches 75%. Traditional methods relying solely on local textures struggle to differentiate them, leading to false alarm rates exceeding 40%.

Static-Dynamic Coexistence: The scene requires unified modeling for both static objects, which require precise "spatial anomaly" localization, and quasi-dynamic objects, which require capturing "temporal changes." Existing single-frame detectors fail to address both simultaneously.

1.2.2. Limitations of Existing Research

Current mainstream methods, such as YOLOv8s, YOLOv10s, and Vision Transformer-based detectors, predominantly focus on the detection of large targets like "pedestrians and animals". In public datasets like the Railway CV Model, target sizes in test samples are mostly concentrated above 64 × 64 pixels, with insufficient attention paid to small targets (<32 × 32 pixels). Furthermore, these methods typically rely on pre-training with general-purpose datasets like COCO and fail to incorporate "normality priors" specific to railway scenes (e.g., "no protrusions on rail heads," "ballast is distributed between sleepers"). Consequently, they struggle to quickly adapt to the specificities of the railway domain during label-efficient fine-tuning. While existing research focuses on large targets (>128 × 128 pixels) that are feature-rich and easy to detect, this study focuses on concealed, static, small targets: bolt fragments on rail heads, anomalous stones in the ballast, and metal wires at rail gaps—objects characterized by feature dilution and high confusion with the background.

1.3. UG-Net

To overcome the aforementioned challenges, this paper proposes a core philosophy: the key to solving railway foreign object detection lies in shifting from the traditional classification strategy of learning infinite variations of "what constitutes an anomaly" to a novel modeling perspective focused on deeply understanding "what constitutes normality." Normality patterns in railway scenes exhibit high consistency, characterized by the parallel linearity of rails, the distribution texture of ballast, and the unobstructed nature within the clearance gauge. While traditional methods require massive annotations to exhaustively cover diverse foreign objects, the proposed approach aims to enable the model to first learn these normality rules in an unsupervised manner. This strategy transforms anomaly discrimination into a metric of deviation between input features and normality rules, thereby converting the problem into a localization task that can be solved with minimal annotations.

Guided by this philosophy, we designed the UG-Net (unsupervised-guided net), an innovative two-stage detection framework:

In the first stage, a masked autoencoder (MAE) is utilized and trained on a large scale of unlabeled normal railway images and pseudo-anomaly images, enabling it to master the capability of reconstructing "normal" scenes. When an image containing a foreign object is input, the MAE's reconstruction naturally "erases" the object that is inconsistent with the learned normality. Subsequently, a deep feature difference method is proposed. It utilizes a pre-trained VGG network to compare the features of the original and reconstructed images at the semantic level, generating a high signal-to-noise ratio spatial attention mask that precisely localizes anomalies.

In the second stage, this clean attention mask serves as a fourth channel fused with the original RGB image, providing a powerful spatial prior for a lightweight YOLOv8n detector. This mask significantly simplifies the detection task by narrowing the search space from the global image down to a few high-suspicion regions. Benefiting from this, the detector no longer requires massive data to learn how to search for targets within

complex backgrounds; instead, it can learn to precisely identify and localize objects within the masked regions using only a minimal number of labeled samples.

1.4. Contributions

The main contributions of this paper can be summarized as follows:

To explicitly distinguish this work from prior arts and highlight its methodological novelty, we emphasize three key contributions: First, unlike traditional supervised methods that rely heavily on exhaustive annotations, we introduce an unsupervised-guided strategy. By leveraging a masked autoencoder to reconstruct "normal" railway scenes, the model learns to identify anomalies through deviation rather than classification, drastically reducing the dependency on labeled data. Second, we propose a deep feature difference strategy. Distinct from the pixel-level background subtraction used in conventional surveillance, this method operates in the semantic feature space, making it robust to environmental noise and illumination changes common in railway environments. Third, we design an efficient channel fusion mechanism. The spatial attention mask generated by the feature difference module is seamlessly integrated as a fourth channel into a lightweight YOLOv8n detector. Combined with a novel weight transfer strategy, this approach improves convergence speed and detection recall without altering the core efficient architecture of the detector.

1.4.1. Methodological Contribution

We propose an unsupervised-guided label-efficient detection framework, UG-Net, which innovatively combines unsupervised normality modeling with label-efficient supervised detection. The core contribution lies in the proposal of a deep feature difference method. This method utilizes a pre-trained VGG network to compare the input and reconstructed output of the masked autoencoder (MAE) at the semantic level, thereby generating a high-quality spatial attention mask. By abandoning the limitations of traditional pixel-level comparisons, this method enables precise anomaly localization within the feature space, providing a novel strategy for addressing the problem of label scarcity in industrial scenarios.

1.4.2. Architectural Contribution

We designed an efficient attention fusion and weight transfer strategy. The generated binary mask acts as an explicit spatial attention mechanism and is fused with the original RGB image to form a four-channel input. To adapt to this structure, the first convolutional layer of the lightweight YOLOv8n detector was modified. Furthermore, we employed an efficient weight transfer strategy: fully retaining the pre-trained weights for the RGB channels while performing optimized initialization for the newly added attention channel. This architectural design allows powerful spatial prior knowledge to be seamlessly injected into a standard detector, significantly simplifying the learning task and accelerating model convergence.

1.4.3. Practical Contribution

We constructed and validated a label-efficient railway foreign object detection benchmark. The effectiveness of the framework was verified on our self-constructed Rail-AD (Railway Anomaly Detection) dataset. This dataset contains 2000 images, blending real-world anomalies with pseudo-anomalies generated via CutPaste to simulate diverse realistic scenarios. Crucially, we demonstrated that using only 200 (10%) labeled samples for training, UG-Net's performance surpasses traditional supervised models trained on the same limited data scale. This provides a verified solution and a dataset benchmark aimed at reducing annotation costs for anomaly detection research in railway and other industrial fields.

2. Related Work

Railway foreign object intrusion detection is a core component for ensuring the operational safety of rail transport. Traditional methods often suffer from significant performance degradation when facing common challenges such as nearly infinite object categories, high costs for full-scale target annotation, and poor generalization in label-efficient scenarios. In recent years, the research focus has gradually shifted from fully supervised learning relying on massive labeled data to more practical technical routes, such as self-supervised/semi-supervised learning, lightweight model design, and spatiotemporal information fusion. This section will systematically review representative works in these directions and clarify the background and innovative value of the technical route proposed in this study.

2.1. Detection Methods Based on Traditional Sensors

Early railway foreign object detection centered on physical sensors such as Light Detection and Ranging (LiDAR), infrared, and ultrasonic sensors, with research focusing on single-sensor performance optimization and multi-sensor fusion.

LiDAR achieves precise positioning through 3D point cloud modeling. Some studies utilize the RANSAC algorithm to segment ground point clouds for identifying small-sized obstacles [1], or optimize reflective materials to enhance signal response [2]. Infrared sensors leverage the advantage of thermal radiation detection, making them suitable for dynamic target detection in low-light environments such as nighttime and foggy weather [3]. Ultrasonic sensors, known for their low cost, are often fused with millimeter-wave radar for close-range early warning to distinguish between metal and non-metal foreign objects [4]. To improve environmental adaptability, multi-sensor fusion schemes (such as the fusion of LiDAR and infrared) employ Kalman filtering for spatiotemporal data alignment, effectively reducing false alarm rates [5].

However, methods based on physical sensors possess inherent limitations: First, environmental robustness is poor; heavy rain and dense fog can severely interfere with LiDAR and infrared signals, leading to a drastic decline in detection performance. Second, equipment procurement and maintenance costs are prohibitively high. Most critically, these methods lack semantic understanding capabilities and cannot distinguish between dangerous foreign objects and benign objects, resulting in persistently high false alarm rates. These defects have prompted research to turn towards vision-based solutions.

2.2. Detection Methods Based on Classical Computer Vision

With the development of image processing technology, foreign object detection has shifted towards vision-based pattern analysis schemes, distinguishing foreign objects from the background by extracting hand-crafted features such as motion and texture.

The mainstream techniques for dynamic foreign object detection include optical flow and background subtraction. Optical flow identifies anomalous trajectories by tracking pixel motion vectors but suffers from significant errors under illumination changes [6]. Background subtraction detects newly appeared objects by comparing the current frame with a spatiotemporally aligned "clean" background reference frame; this approach solves the background drift problem associated with moving cameras and has been optimized for real-time detection on high-speed trains [7,8]. For static foreign objects, researchers utilize features such as texture and shape combined with classifiers for recognition [9], or employ one-class classification algorithms (such as One-Class SVM) to determine whether the entire image is anomalous [10]. Although the latter can identify anomalous images, it fails to provide the specific location of the foreign object.

Recent advancements have also focused on low-cost, camera-based solutions that balance performance with deployment feasibility. For instance, Martínez Núñez et al. [11] proposed a

system combining background subtraction with deep learning for real-time surveillance, effectively detecting people and objects on tracks using a single conventional camera. Similarly, Niu et al. [12] introduced "MSL-YOLO," a lightweight detector specifically tailored for railway environments, which addresses the challenges of small object detection under real-time constraints. However, these methods typically rely on either pixel-level background modeling, which is sensitive to light variations, or standard supervised learning, which requires extensive datasets. In contrast, our UG-Net employs a semantic-level reconstruction approach, addressing both the limitations of pixel-based subtraction and the high annotation costs associated with fully supervised detectors.

Such methods rely heavily on hand-crafted features and thus possess inherent limitations: First, environmental adaptability is poor, as illumination changes severely affect feature stability; second, it is difficult to effectively describe the incomplete features of small-sized foreign objects (such as bolt fragments); finally, the high computational complexity of multi-frame optical flow tracking and background modeling makes it difficult to meet real-time requirements in high-speed scenarios. These defects have driven research towards deep learning, which is characterized by automatic feature learning.

2.3. Vision-Based Detection via Supervised Deep Learning

To overcome the aforementioned limitations, research has entered the phase of "end-to-end supervised learning" centered on convolutional neural networks (CNNs). The YOLO series has become a mainstream benchmark due to its efficient single-stage architecture. Researchers have enhanced its performance by introducing deformable convolutions [13,14] or optimizing network structures [15]. Meanwhile, to meet the real-time requirements of on-board or UAV inspections, solutions based on lightweight backbone networks like MobileNetV3 integrated with attention mechanisms have been proposed, achieving a balance between speed and precision on embedded devices [16]. Furthermore, researchers have specifically improved YOLO's detection capabilities for small-sized foreign objects by adding extra detection heads and optimizing loss functions [17].

However, all supervised learning methods suffer from fundamental limitations: First, models cannot generalize to identify foreign object categories unseen in the training set; second, the cost of constructing large-scale annotated datasets is prohibitively high; finally, model performance significantly degrades in complex scenarios such as turnouts and curved tracks. These issues have driven research towards unsupervised and semi-supervised anomaly detection.

2.4. Self-Supervised Anomaly Detection

To break through the limitations of supervised learning, research has shifted towards the unsupervised/semi-supervised strategy of "learning normality to identify anomalies," which is theoretically capable of detecting any unknown foreign objects. This direction is dominated by reconstruction-based models and synthetic anomaly techniques.

Self-supervised learning (SSL) is one of the core technologies. Reconstruction methods, represented by masked autoencoders (MAEs), provide powerful visual representations for downstream detection tasks by pre-training on a large number of normal images via masked reconstruction, significantly improving the accuracy of label-efficient fine-tuning [18,19]. Meanwhile, the synthetic anomaly strategy represented by CutPaste provides supervision signals by pasting pseudo-anomaly samples onto normal images; it has shown excellent performance on industrial datasets, offering a new approach for detection with limited annotations [20]. Generative adversarial networks (GANs) have also been employed for this task, where a generator learns the normal data distribution to localize anomalies based on the difference between the input and generated images [21]. While this approach has

been successfully applied to railway scenarios, it is difficult to meet real-time requirements due to its slow inference speed [22].

Furthermore, emerging architectures such as state space models (SSMs), represented by Mamba, have demonstrated great potential in handling railway surveillance video streams characterized by strong spatial structures and temporal correlations, owing to their efficient long-sequence modeling capabilities [23,24]. However, their specific application in this domain remains a direction to be explored.

Nevertheless, unsupervised/semi-supervised methods still face core challenges: First, a balance must be struck between reconstruction fidelity and anomaly sensitivity; overly strong reconstruction capability tends to "repair" anomalies, leading to missed detections, whereas overly weak capability results in false alarms caused by normal environmental changes. Second, high-precision pixel-level localization is often accompanied by substantial computational costs, making it difficult to achieve real-time detection in high-speed scenarios.

To provide a comprehensive overview of the research landscape, Table 1 systematically compares the characteristics, strengths, and limitations of these mainstream detection strategies. As shown in the table, our proposed UG-Net uniquely balances low annotation costs with high robustness, effectively addressing the gaps left by traditional and fully supervised methods.

Table 1. Comparison of different railway object detection strategies.

Method Type	Representative Work	Core Strategy	Annotation Need	Robustness to Light
Traditional Sensor	LiDAR/Radar [1,5]	Physical Echo/ToF	Low	Low (Rain/Fog)
Classical Vision	Background Subtraction [7]	Pixel Difference	None (for motion)	Low (Shadows)
Supervised DL	YOLO/MSL-YOLO [12]	End-to-End CNN	High (Full)	High
Proposed	UG-Net	Unsupervised + label-efficient	Low (10%)	High

3. Method

To address the critical challenges of open-set object categories and prohibitive annotation costs in railway foreign object detection, this paper proposes a novel unsupervised-guided label-efficient detection framework, named UG-Net (unsupervised-guided network).

The workflow consists of three main stages, as illustrated in Figure 1: (I) unsupervised normality modeling, where a masked autoencoder (MAE) learns to reconstruct normal railway scenes; (II) spatial attention generation, where a deep feature difference module compares the semantic features of the original and reconstructed images to generate a high-quality spatial attention mask; and (III) attention-guided label-efficient detection, where this mask is fused as a fourth channel into a modified YOLOv8n detector to guide precise object localization.

3.1. Unsupervised Normality Modeling Based on MAE

The goal of this stage is to construct a deep model capable of profoundly understanding "normal" railway scenes without relying on any manual annotations. We selected the masked autoencoder (MAE) as the core model, which is an efficient self-supervised learning strategy.

We employed a standard ViT-Large backbone (patch size 16 × 16, 24 transformer blocks, 1024 embedding dimension) as the encoder to capture global semantic dependencies.

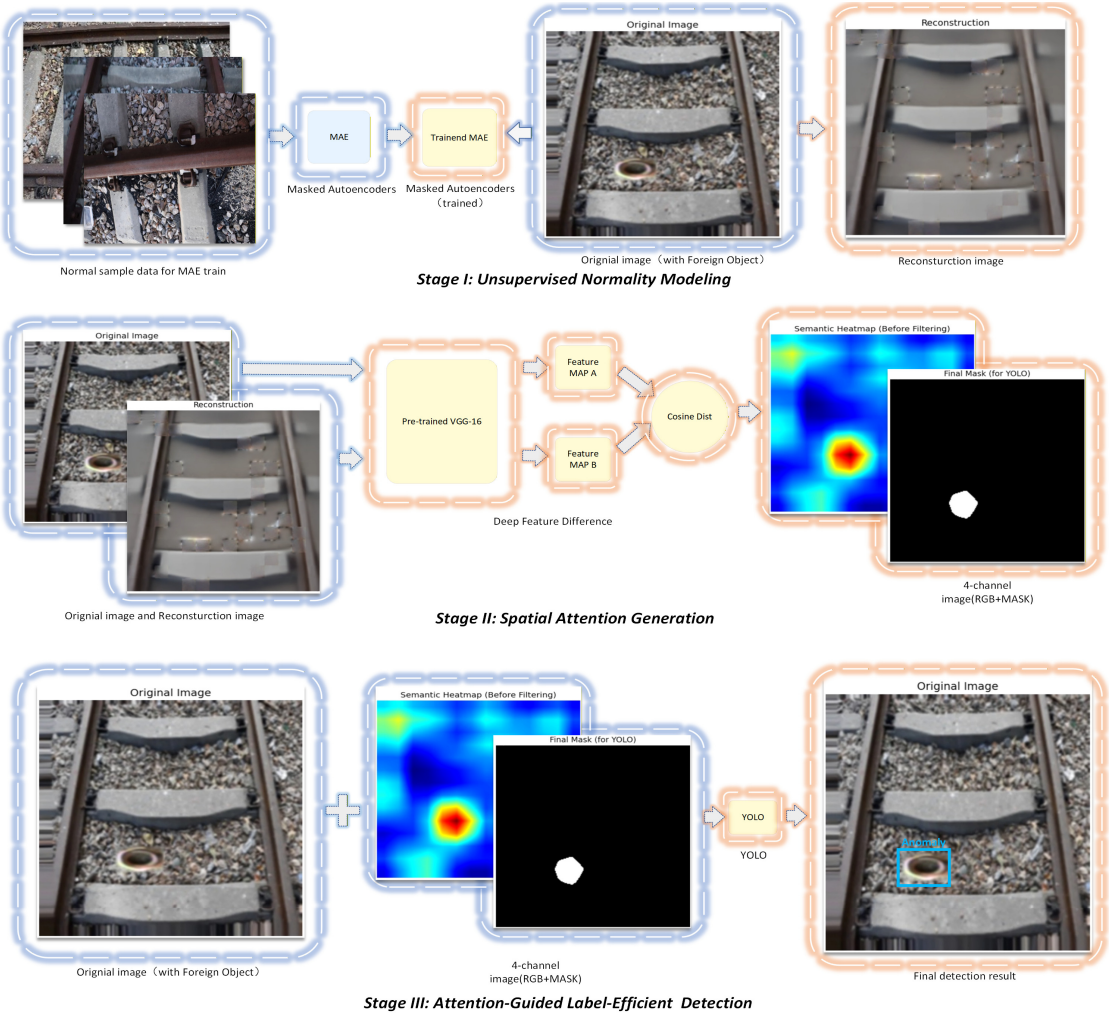


Figure 1. The overall pipeline of the proposed UG-Net framework.

In the specific workflow, we first collected a large-scale, unlabeled, and data-augmented dataset of normal railway images and generated 5000 pseudo-anomaly images using the CutPaste method. During the training process, transfer learning was employed based on the initial weights of ViT-Large. We froze the training for the first 20 epochs and utilized a learning rate schedule that first slowly increased (warm-up) and then slowly decreased. This strategy ensures that the model learns the normality priors of the railway tracks without destroying the powerful weights pre-trained on ImageNet-1k. Each image input to the MAE is randomly masked with a high ratio of image patches. The model's task is to recover the complete original image based solely on the few visible patches. This challenging "cloze test" forces the model to learn deep structural and semantic knowledge, such as image context, texture distribution, the "continuity of rails," and the "specific texture patterns of ballast".

To enable the MAE network to not only understand the "normality" of railway scenes but also proactively identify and distinguish "anomalies," we designed a composite loss function $\mathcal{L}_{total}$ consisting of three parts. This function is composed of a weighted sum of normal reconstruction loss $(\mathcal{L}_{normal})$, abnormal suppression loss $L_{abnormal}$, and feature contrast loss $(\mathcal{L}_{contrast})$.

3.1.1. Normal Reconstruction Loss

This component corresponds to the standard masked autoencoder (MAE) loss. Its objective is to enable the model to learn how to accurately reconstruct the "normal"

background from masked images. The L_1 loss (mean absolute error) is computed exclusively on those image patches that are masked and belong to the normal background.

The mathematical expression is defined as follows:

$$\mathcal{L}_{normal} = \frac{1}{|M_{norm}|} \sum_{i \in M_{norm}} \|P_i - \hat{P}_i\|_1 \quad (1)$$

In Equation (1), P_i represents the i-th original image patch, and $\hat{P}_i$ is the model's reconstruction result for the i-th image patch. M_{norm} denotes the set of all masked normal image patches in the dataset, and $|M_{norm}|$ is the count of image patches in this set.

3.1.2. Abnormal Suppression Loss

This component is one of the core innovations of the framework. The design intention is to force the model to choose to ignore anomalies even if it recognizes them, and instead reconstruct the normal background that should have existed in that location. To achieve this, we aimed to maximize the dissimilarity between the model's reconstruction of these regions and the original input containing the pseudo-anomaly. The Structural Similarity Index (SSIM) was used to measure this dissimilarity.

The mathematical expression is

$$\mathcal{L}_{abnormal} = 1 - SSIM(P_j, \hat{P}_j) \quad (2)$$

In Equation (2), P_j and $\hat{P}_j$ represent the original and reconstructed versions of the j-th image patch, respectively. j belongs to the set $V_{abnormal}$, which represents all visible anomalous image patches. The range of the SSIM function is $[-1, 1]$, where a value closer to 1 indicates higher similarity between the two images. Therefore, using $1 - SSIM$ as the loss effectively forces the model's reconstruction result to diverge from the pseudo-anomaly input.

3.1.3. Feature Contrast Loss

The objective of this loss is to increase the distance between foreign object features and normal rail texture features in the feature vector space. This reflects the idea of contrastive learning, enabling the model to learn deep features with greater discriminative power.

The specific process is as follows: Feature maps are extracted from the deep layers of the backbone network. Based on the ground truth, all feature vectors in the feature maps are divided into a "normal feature set" (F_{norm}) and an "abnormal feature set" (F_{abn}). The average of all abnormal features is calculated to obtain an "anomaly prototype" anchor vector (v_{anchor}). We aimed to minimize the cosine similarity between this anomaly anchor and all normal feature vectors (i.e., making the vector directions as opposite as possible, with an angle close to 180°).

$$\mathbf{v}_{anchor} = mean\left(\{\mathbf{v}_j | \mathbf{v}_j \in F_{abn}\}\right) \quad (3)$$

$$\mathcal{L}_{contrast} = 1 + \frac{1}{|F_{norm}|} \sum_{v_i \in F_{norm}} \frac{\boldsymbol{v}_{anchor} \cdot \boldsymbol{v}_i}{\|\boldsymbol{v}_{anchor}\| \|\boldsymbol{v}_i\|} \quad (4)$$

In Equation (4), the meaning of v_{anchor} is given by Equation (3). The range of cosine similarity is $[-1, 1]$. When two vectors are completely opposite, the similarity is -1, and $\mathcal{L}_{contrast}$ reaches its minimum value of 0. When two vectors are identical, the similarity is 1, resulting in the maximum loss.

3.1.4. Total Loss Function

Finally, these three loss terms are combined via a weighted sum to balance the model's focus on different objectives, enabling it to better reconstruct railway normality while avoiding the reconstruction of anomalies.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{normal} + \lambda_2 \mathcal{L}_{abnormal} + \lambda_3 \mathcal{L}_{contrast} \quad (5)$$

In Equation (5), λ_1 is set to 0.6, λ_2 is set to 0.2, and λ_3 is set to 0.2. The reconstruction effects of the railway tracks in the MAE task are shown in Figures 2 and 3. As seen in Figure 2, there is a green aluminum can on the rail head in the upper part of the image, which was not masked; the model successfully chose to reconstruct the entire rail, ignoring the anomalous parts. Similarly, in Figure 3, there is a black wrench in the center of the track, and the model likewise successfully reconstructed the normal track.

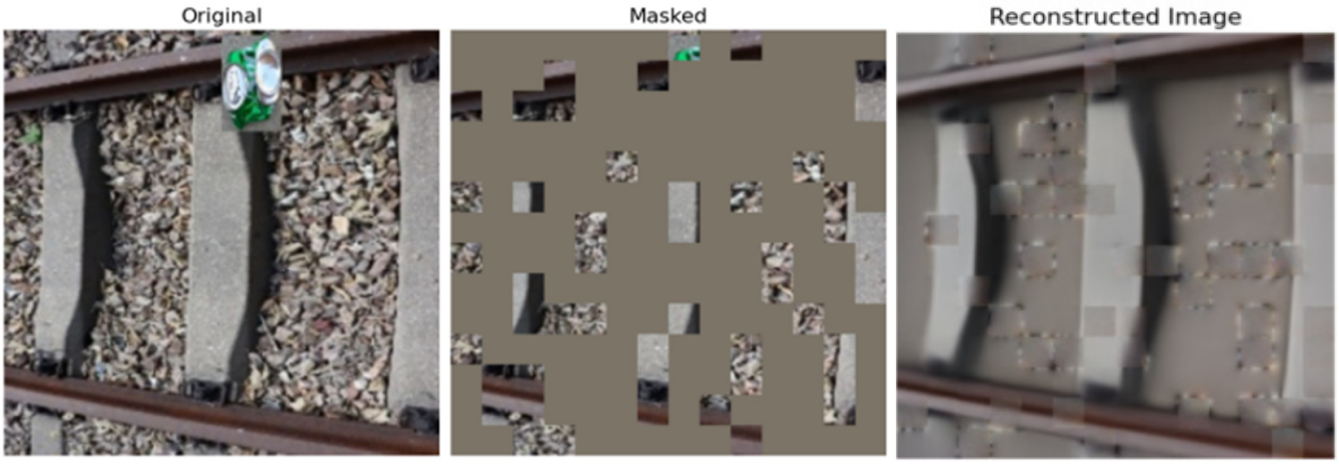


Figure 2. Schematic diagram of rail mask reconstruction (1).

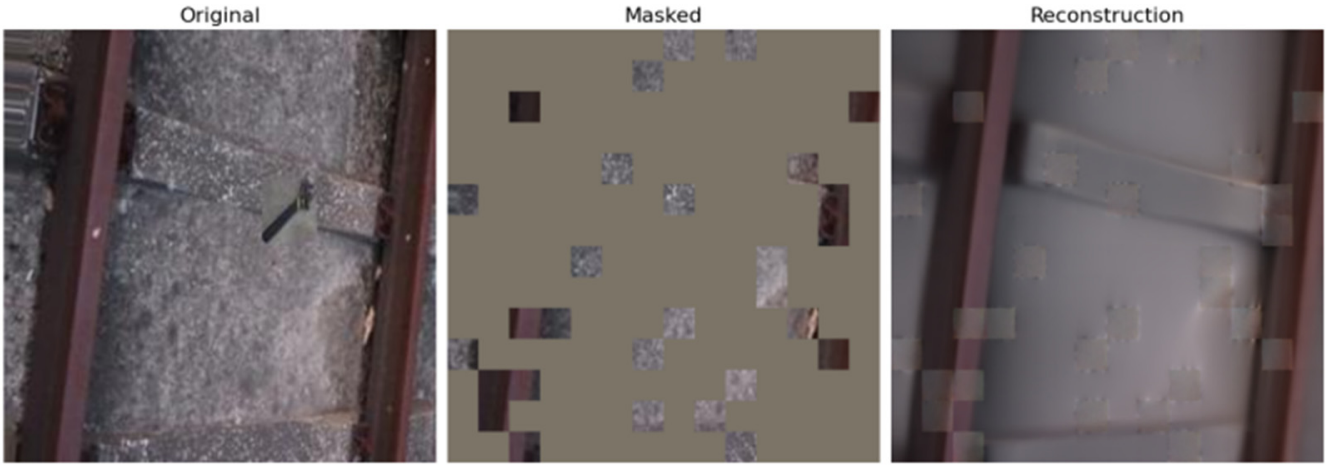


Figure 3. Schematic diagram of rail mask reconstruction (2).

Through self-supervised training on massive normal samples, we ultimately obtained a powerful "normality" model. This model is not merely a simple tool for image compression and reconstruction but an expert proficient in the "visual grammar" of normal railway scenes, laying a solid foundation for subsequent anomaly localization.

3.2. Attention Mask Generation Based on Deep Feature Difference

After obtaining the "normality" model, the goal of this stage was to leverage it to generate a precise anomaly attention mask for any input image.

It is worth noting that while semantic segmentation models could theoretically generate precise masks, they strictly require pixel-level annotations (ground truth masks) for training, which are prohibitively expensive and labor-intensive to acquire in industrial settings. Crucially, our proposed deep feature difference strategy generates these spatial attention masks in a completely unsupervised manner. This effectively eliminates the need for manual pixel-wise segmentation annotations, thereby overcoming a major bottleneck in traditional supervised defect detection pipelines.

First, when an "anomalous" image containing a foreign object is input into the trained MAE, since the foreign object does not conform to the learned "normality" distribution, the model tends to "repair" the foreign object area using normal background textures during reconstruction. To precisely measure the discrepancy between the original and reconstructed images, we abandoned pixel-level comparison, which is susceptible to background texture interference, and instead proposed a deep feature difference method.

Theoretically, traditional pixel-level comparisons (e.g., L_1 or L_2 distance) assume strict spatial alignment and pixel-value consistency. However, in outdoor railway environments, high-frequency noise caused by lighting changes, shadows, or slight camera vibrations can lead to significant pixel-wise differences even in normal regions, resulting in false positives. In contrast, deep features extracted by CNNs (like VGG) capture high-level semantic abstractions (e.g., shapes, textures) and possess translation invariance [25]. By computing differences in this semantic space, our method effectively filters out low-level environmental noise while remaining highly sensitive to structural anomalies. Quantitative experiments presented in Section 4 further demonstrate that this semantic-level comparison significantly outperforms pixel-level (L_1) and structural (SSIM) metrics in terms of noise robustness and detection accuracy.

This method utilizes a VGG16 network pre-trained on ImageNet as a feature extractor to compare the deep feature maps of the two images at the semantic level. Comparison in the feature space effectively ignores minor reconstruction errors in the background texture while drastically amplifying the semantic concept differences caused by the "erasure of the foreign object," thereby generating a "semantic heatmap" with an extremely high signal-to-noise ratio. Subsequently, a peak relative thresholding method is applied to the semantic heatmap to convert it into a clean binary (black and white) mask containing only the core anomaly region. This mask serves as a powerful, explicit spatial attention prior in the next stage.

As shown in Figures 4 and 5, the location information of the foreign object is ultimately transformed into a clean binary mask containing only the core anomaly region, waiting to be passed to the YOLO detector.

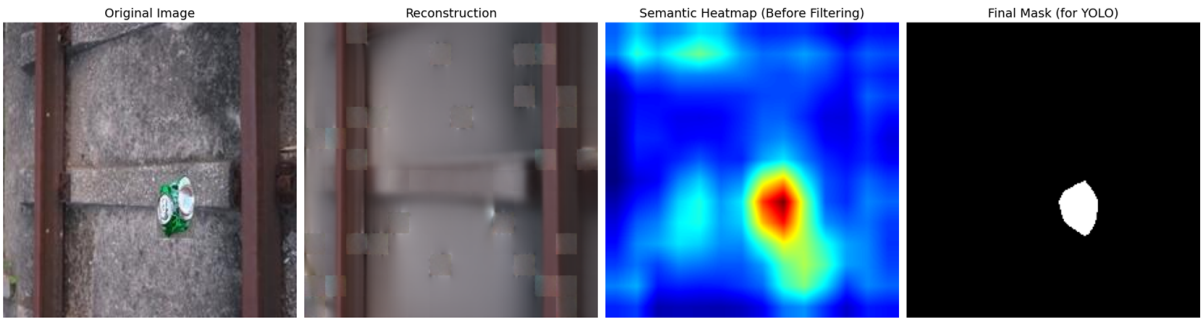


Figure 4. Schematic diagram of binary mask generation (1).

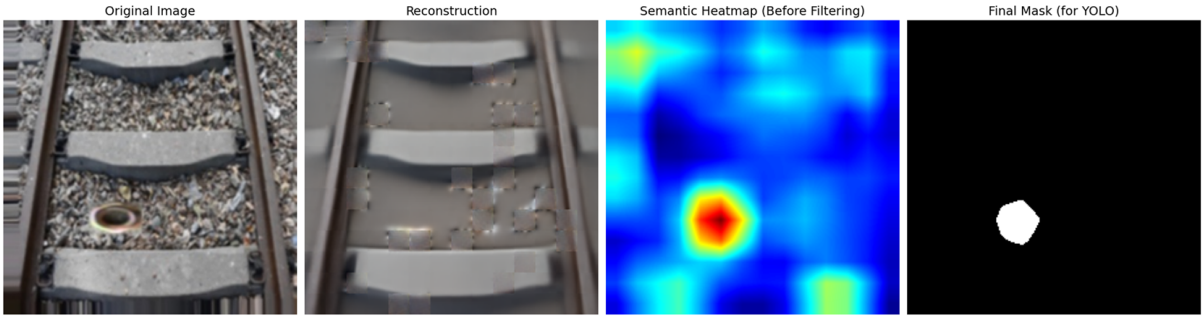


Figure 5. Schematic diagram of binary mask generation (2).

3.3. Attention-Guided Label-Efficient Foreign Object Detection

The objective of this stage is to utilize the attention mask generated in the previous stage to guide a standard detector to learn precise foreign object recognition and localization using an extremely small number of labeled samples.

(1) Input Fusion via Four-Channel Concatenation: The binary mask generated in the second stage is treated as the fourth channel and concatenated with the original three-channel RGB image to form a four-channel input tensor (RGB + Mask).

(2) Detector Architecture and Weight Transfer Strategy: We selected the lightweight YOLOv8n as the base detector. To accommodate the four-channel input, the first convolutional layer of its network structure was modified. Simultaneously, to maximize the utilization of prior knowledge and accelerate convergence, an efficient weight transfer strategy was adopted: the pre-trained weights for the three RGB channels learned on the COCO dataset were fully retained, while the weights for the newly added fourth channel underwent optimized initialization.

To accommodate the 4-channel input (RGB + Attention Mask), we modified the first convolutional layer (Conv1) of the YOLOv8n backbone. The original filter shape was (c_out, 3, k, k). We expanded this to (c_out, 4, k, k). The weights for the first 3 channels were initialized from the COCO pre-trained model, while the weights for the 4th channel were initialized with a truncated normal distribution (mean = 0, std = 0.02) to allow the network to gradually learn the importance of the attention prior. The spatial mask generated by the MAE is resized to match the input resolution of the detector, ensuring seamless tensor concatenation.

(3) Label-Efficient Fine-Tuning: Ultimately, this modified four-channel detector requires fine-tuning on only a minimal scale (e.g., 200 images) foreign object dataset with bounding box annotations to achieve or even surpass the performance of traditional fully supervised methods. While "label-efficient learning" typically refers to N-way K-shot tasks (1–10 samples) in generic computer vision, in this industrial context, we use the term to describe a "label-efficient" regime. Training with only 200 samples (10% of the dataset) represents a significant reduction compared to the thousands of annotated images usually required for robust industrial detectors.

4. Experiments

4.1. Experimental Setup

All experiments in this paper were conducted based on PyTorch 1.9.0 and CUDA 11.1. The unsupervised pre-training stage was performed on an NVIDIA A10 GPU (NVIDIA Corporation, Santa Clara, CA, USA), while the supervised fine-tuning of the detector was completed on an NVIDIA RTX 4060 Ti GPU (GIGABYTE Technology Co., Ltd., Taipei City, Taiwan). The experimental equipment and specific hyperparameter settings are detailed in Table 2.

Table 2. Experimental environment and hyperparameters.

Parameter	MAE Stage (Pre-Training)	YOLO Stage (Detection)
CPU	Intel i5-13490f	Intel i5-13490f
GPU	NVIDIA A10 (24G)	RTX4060ti (8G)
Framework	torch = 1.9.0 + cu111 timm = 0.3.2 numpy = 1.21.5	Torch = 1.9.0 + cu111 numpy = 1.26.4
Hyperparameters	batchsize = 64 warmup = 20 epoch = 200 blr = 1.5×10^{-4} maskratio = 0.75	batchsize = 8 epoch = 200 numwork = 0

4.2. Data Acquisition and Processing

To effectively validate the performance of the proposed framework in label-scarce scenarios, we constructed a dataset specifically for railway foreign object detection, named Rail-AD (Railway Anomaly Detection). The design of this dataset aims to realistically overcome the challenge of having massive amounts of railway scene data but few effective labeled samples.

4.3. Dataset Composition

The Rail-AD dataset contains a total of 2000 images with a unified resolution of 224 × 224. Its composition features mixed diversity to comprehensively evaluate model generalization:

Real Anomaly Samples: These images were captured via field photography or extracted from real surveillance footage, covering various foreign objects appearing in actual railway environments, such as stones, metal parts, plastic bags, and small tools. This portion ensures scene authenticity.

Synthetic Pseudo-Anomaly Samples: To significantly expand the diversity of foreign object samples without increasing annotation costs, the CutPaste data synthesis strategy was employed. This strategy seamlessly pastes various foreign object image patches collected from the web onto collected normal railway background images using techniques like Poisson blending and brightness matching, thereby rapidly generating a large number of pseudo-anomaly samples with varying shapes, sizes, and positions.

4.3.1. Dataset Splitting and Annotation

The entire dataset was split and annotated following the principles of label-efficient learning. The 2000 images were strictly divided into training, validation, and test sets in a 70%/15%/15% ratio (training: 1400 images, validation: 300 images, test: 300 images). Regarding annotation, to highlight the label efficiency of our framework, we randomly selected only 200 images from the 1400 training images for fine-grained bounding box annotation. Conversely, the validation set (300 images) and test set (300 images) used for model tuning and final performance evaluation were fully annotated to ensure the objectivity and rigor of the evaluation process.

4.3.2. Data Augmentation

To further enhance the model's robustness in complex environments and suppress overfitting under label-efficient conditions, rich online data augmentation techniques were applied during the YOLO detector training stage. When the model reads each batch of images from the data loader, a series of random transformations are applied, primarily including: geometric transformations (random horizontal flip, small-angle rotation, ran-

dom scaling, and translation); color and optical transformations (random adjustments to brightness, contrast, saturation, and hue); and noise injection (simulating real sensor noise by randomly adding Gaussian noise). Additionally, mosaic augmentation was utilized to stitch four different training images into one, allowing the model to learn multiple targets in different backgrounds within a single forward pass, which is particularly beneficial for improving the detection capability for small targets.

Through these processes, we constructed a high-quality dataset that reflects real-world challenges and effectively validates the performance of the label-efficient learning framework.

4.4. Evaluation Metrics

This paper adopts precision (P), recall (R), mean average precision (mAP), floating point operations (FLOPs), and model parameter count as evaluation metrics to comprehensively assess the performance of the model. An improvement in mean average precision indicates enhanced detection accuracy of the refined model, while a reduction in FLOPs demonstrates the lightweight characteristics of the improved model. The calculation formulas for these metrics are presented in Equations (6)–(8):

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{8}$$

where TP represents the number of positive samples correctly identified; FP denotes the number of negative samples identified as positive; FN indicates the number of positive samples misclassified; n is the number of categories; and AP represents the Average Precision for a specific target class.

In the context of foreign object detection:

True Positive (TP): A foreign object is correctly detected with an intersection over union (IoU) > 0.5.

False Positive (FP): The model detects a foreign object where there is none (background misclassification).

False Negative (FN): A real foreign object is missed by the detector. Note that True Negative (TN) is generally not used in object detection metrics (mAP) as the number of background bounding boxes is infinite.

4.5. Experimental Results and Analysis

To verify the effectiveness of the proposed four-channel model on the foreign object detection task, a series of comparative experiments was conducted. The standard YOLOv8n model served as the 3-channel baseline, and its performance was compared with the proposed 4-channel UG-Net. The primary evaluation metrics were mAP@0.50 and mAP@0.50:0.95.

All quantitative metrics reported in this section were evaluated on the independent test set (300 images), which was strictly isolated from the training process.

Experiments were conducted on the training subset containing only 200 images. The experimental results are shown in Figure 6. The left side of the figure illustrates the detection results of YOLOv8n under label-efficient conditions, where false detections, missed detections, and low confidence scores are observable. In contrast, the right side

shows the results of the proposed UG-Net architecture, which demonstrates high detection precision and high confidence.



Figure 6. Visualization of experimental detection results.

As shown in Figure 7, on this small-scale dataset, UG-Net achieved an mAP@0.50 of 94.56%, significantly outperforming the baseline model's 86.91%. Notably, the performance achieved by UG-Net using only 200 training images approximates the results that the baseline model could only achieve when trained on the full set of 1400 images. This result indicates that the proposed method possesses higher data efficiency and can more effectively learn target features in label-efficient scenarios.

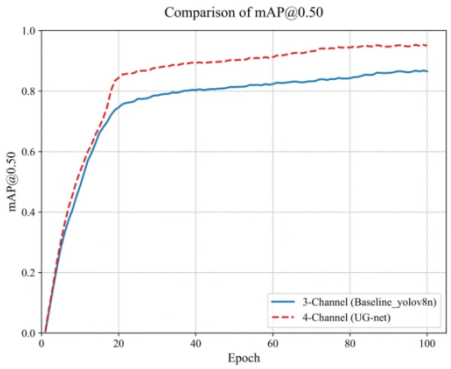


Figure 7. Comparison of mAP@50 curves.

Figure 8 further reveals that the four-channel model exhibits significant advantages throughout the entire training process. On the stricter mAP@0.50:0.95 metric, the proposed four-channel model ultimately achieved a performance of 0.79756, whereas the three-channel baseline only reached 0.64869. This represents an absolute performance improvement of approximately 14.9 percentage points. Furthermore, the curves in Figures 7 and 8 demonstrate that the four-channel model not only achieves higher final accuracy but also has a convergence speed far superior to that of the baseline model. This proves that introducing information from the fourth channel significantly improves the model's learning efficiency.

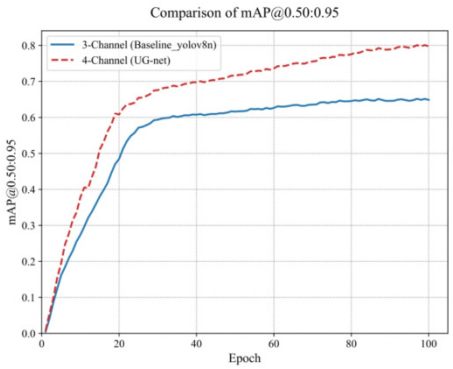


Figure 8. Comparison of mAP@50:95 curves.

4.6. Comparison with Mainstream Algorithms

Although recent self-supervised anomaly detection methods (e.g., PatchCore, PaDiM) have shown promise in industrial inspection, they typically output pixel-level anomaly scores or heatmaps. However, automated railway maintenance robots strictly require precise bounding box coordinates for mechanical manipulation. Therefore, direct quantitative comparison (mAP) with these methods is methodologically infeasible. Consequently, we limit our quantitative comparison to standard object detectors (e.g., YOLO series, Faster R-CNN) that align with the required engineering output format.

We compared UG-Net with existing mainstream models under identical experimental environments, training parameters, and training datasets. The relevant parameters are recorded in Table 3.

Table 3. Comparison with mainstream models.

Model	P/%	R/%	mAP%@0.50	mAP@0.50:0.95%	Params (M)
RCNN	60.12	78.3	75.8	61.2	41.3
ssd	82.6	63.2	70.1	48.7	24.0
Yolov5	87.3	91	77.2	61	7.2
Yolov7	88.2	94	80.23	62.3	6.2
Yolov8n	90.1	92	86.91	64.87	3.2
UG-Net	94	95.1	94.56	79.76	3.2 (+0.04)

As shown in Table 3, under conditions of limited training data, the proposed network architecture surpasses mainstream models such as Faster-RCNN, SSD, and YOLOv7 in multiple metrics, including Precision, Recall, mAP@50, and mAP@50:95. Moreover, since the proposed architecture is built upon YOLOv8n during the detection phase, there is no significant difference in overall parameter count. The four-channel model only adds a small number of parameters in the first convolutional layer to adapt to the fourth input channel (the mask channel). This demonstrates that the significant improvement in performance does not stem from increased model complexity, but rather from the effective utilization of the additional channel information.

Statistical Analysis of Stability

To mitigate the influence of random initialization and verify the robustness of the proposed method, we conducted 5 independent runs for both the Baseline (YOLOv8n) and UG-Net using different random seeds. The statistical results (mean ± standard deviation) are summarized in Table 4.

Table 4. Statistical analysis of mAP@0.50 over 5 independent runs.

Method	Run 1	Run 2	Run 3	Run 4	Run 5	Mean ± Std (%)
Baseline (3-Channel)	86.91	86.54	87.12	86.80	86.65	86.80 ± 0.23
UG-Net (4-Channel)	94.56	94.32	94.81	94.45	94.60	94.55 ± 0.19

Analysis of Difference Strategies: To validate the effectiveness of the proposed deep feature difference strategy, we compared it with traditional pixel-level difference (L1distance) and structural similarity (SSIM) methods. As shown in Table 5, pixel-level methods suffer from a high false alarm rate (42.3%) due to their sensitivity to environmental noise such as lighting changes and shadows in outdoor railway scenes. In contrast, the proposed VGG-16-based semantic difference achieved the lowest false alarm rate (4.1%) and the highest detection accuracy (94.56%), validating that high-level semantic features are more robust for identifying structural anomalies than raw pixel intensities.

Table 5. Comparison of difference measurement strategies.

Difference Strategy	Metric Space	Robustness to Light/Noise	False Alarm Rate (%)	mAP@0.50 (%)
Pixel-level (L_1)	Raw Pixel Intensity	Low	42.3	68.45
SSIM	Structural Window	Medium	28.5	76.20
Deep Feature (ResNet-50)	Semantic Feature	High	8.4	91.12
Deep Feature (VGG-16)	Semantic Feature	Very High	4.1	94.56

4.7. Ablation Study

To rigorously quantify the contribution of each key component in the proposed UG-Net framework, we conducted a comprehensive ablation study on the Rail-AD test set. Specifically, we investigated two critical aspects: (1) the impact of different loss function components ($L_{abnormal}$ and $L_{contrast}$) on the unsupervised normality modeling and (2) the influence of different thresholding strategies on the quality of mask generation. The quantitative results are summarized in Table 6.

Table 6. Comprehensive ablation study of loss components and thresholding strategies.

(**A**) Contribution of Loss Components.

Model Variant	L_{normal}	$L_{abnormal}$	$L_{contrast}$	Recall (%)	mAP@0.50 (%)
Baseline (Standard MAE)	√	×	×	74.5	81.15
+ Abnormal Suppression	√	√	×	91.2	89.40
+ Feature Contrast	√	×	√	78.4	88.75
UG-Net (Full Method)	√	√	√	95.1	94.56

(**B**) Impact of Thresholding Strategies.

Strategy	Mechanism	mAP@0.50 (%)	Observation
Fixed Threshold ($\tau = 0.3$)	Global Constant	85.20	High False Positive Rate (Noise)
Fixed Threshold ($\tau = 0.7$)	Global Constant	81.45	High False Negative Rate (Misses)
Otsu's Method	Variance-based	89.60	Unstable on ballast textures
Peak Relative (Ours)	Adaptive ($0.6 \times Max$)	94.56	Robust to contrast variations

Analysis of Ablation Results: As shown in Table 6A, the baseline MAE (trained only with reconstruction loss) achieves a relatively low recall (74.5%) because it tends to "repair" anomalies. Introducing the abnormal suppression loss ($L_{abnormal}$) significantly boosts recall to 91.2% by forcing the model to reconstruct the background instead of the foreign object. Meanwhile, the feature contrast loss ($L_{contrast}$) improves feature discriminability, further lifting the overall mAP to 94.56% when combined. Regarding the mask generation in Table 6B, fixed thresholds ($\tau = 0.3$ or 0.7) fail to adapt to varying illumination, leading to either excessive noise or missed detections. Although Otsu's method improves performance (89.60%), it occasionally misclassifies complex ballast textures as anomalies. The proposed Peak Relative Thresholding strategy achieves the best performance (94.56%) by adaptively focusing on the most salient regions in the semantic heatmap, validating its superiority for generating high-quality attention priors.

5. Conclusions and Future Work

5.1. Conclusions

This paper addresses the core challenges of high annotation costs and poor label-efficient generalization in railway foreign object detection by proposing an unsupervised-guided label-efficient detection framework (UG-Net). This framework successfully combines unsupervised "normality" modeling with efficient supervised detection through an effective two-stage framework.

In the first stage, a masked autoencoder (MAE) is utilized to learn the "normality" priors of railway scenes from a large amount of unlabeled data. We innovatively propose a deep feature difference strategy, which uses a pre-trained VGG network to compare the original image with the MAE's reconstructed output in the semantic space. This process generates a spatial attention mask with an extremely high signal-to-noise ratio that precisely identifies anomaly locations.

In the second stage, this mask serves as an explicit spatial prior and is fused with the original RGB image to form a four-channel (RGB + Mask) input, which is then fed into a lightweight YOLOv8n detector. Benefiting from the strong guidance provided by this attention channel, the detector achieves efficient fine-tuning using only a minimal number of labeled samples.

Experimental results demonstrate that on the self-constructed Rail-AD dataset, UG-Net achieves an mAP@0.5 of 94.56% using only 200 (10%) labeled training images. This performance significantly outperforms the YOLOv8n baseline model (86.91%) trained on the same data scale, as well as other mainstream models such as Faster-RCNN and YOLOv7. Experiments prove that UG-Net provides a novel, unsupervised, label-efficient, and high-performance solution for solving high-cost industrial anomaly detection problems.

5.2. Limitations and Failure Analysis

While UG-Net demonstrates superior label efficiency, we acknowledge certain limitations and typical failure modes based on our qualitative analysis. First, regarding inference latency, the multi-stage pipeline introduces computational overhead (approx. 45 ms per frame on an RTX 4060Ti). While this satisfies the requirements for offline maintenance vehicle inspection, it presents challenges for ultra-high-speed real-time deployment. Second, we identified two main failure types: (1) False Positives: The model may incorrectly identify newly installed but normal track components as anomalies if they were absent in the unsupervised "normality" training set; (2) False Negatives: Under conditions of severe motion blur or extreme low light, the MAE reconstruction quality degrades, leading to inaccurate attention masks and missed detections of small, low-contrast objects. Finally, although CutPaste synthetic data effectively simulates open-set defects, future work

will aim to collect more real-world samples under diverse weather conditions to further improve robustness.

We acknowledge that the evaluation is limited to a single dataset. This is primarily because existing public anomaly detection benchmarks (e.g., MVTec AD) focus on indoor manufacturing objects with controlled backgrounds, which exhibit a massive domain gap compared to the unstructured, outdoor railway environments addressed in this study. Currently, there is no publicly available large-scale dataset specifically for small-target railway foreign object detection.

5.3. Future Work

Building on the current results and limitations, future research can expand in the following directions:

Optimization of Inference Efficiency: To address the latency issue mentioned above, future work can investigate how to improve the overall inference speed while maintaining high precision through model distillation, pruning, or designing a more compact end-to-end network.

Spatiotemporal Information Fusion: The current framework is mainly based on single-frame static image detection. However, railway inspection is essentially a temporal video stream task. Future work can explore how to introduce temporal information, for example, by utilizing state space models (Mamba) or Transformer architectures to model the dynamic changes of the scene. This helps to further distinguish "quasi-dynamic foreign objects" (such as wind-blown woven bags) from normal background changes.

Multi-modal Sensor Fusion: Railway safety detection is a complex system engineering task. As mentioned in related work, LiDAR and infrared sensors possess distinct environmental adaptation advantages. Future work can explore fusing the visual detection results of this framework with LiDAR point cloud data or infrared thermal imaging data to construct an all-weather, highly robust foreign object detection system.

Author Contributions: Writing—original draft, Z.T.; Writing—review and editing, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: To promote reproducibility and transparency, the Rail-AD dataset constructed in this study, along with the core implementation code, has been publicly released. The dataset and code are available in the author's GitHub repository at: https://github.com/1226xmas/Rail-AD (accessed on 28 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Qu, J.; Li, S.; Li, Y.; Liu, L. Research on Railway Obstacle Detection Method Based on Developed Euclidean Clustering. *Electronics* **2023**, *12*, 1175. [CrossRef]
2. Kim, J.H.; Patil, V.; Chun, J.M.; Park, H.S.; Seo, S.W.; Kim, Y.S. Design of Near Infrared Reflective Effective Pigment for LiDAR Detectable Paint—Addendum. *MRS Adv.* **2020**, *5*, 2535. [CrossRef]
3. Berg, A.; Öfjäll, K.; Ahlberg, J.; Felsberg, M. Detecting Rails and Obstacles Using a Train-Mounted Thermal Camera. In *Image Analysis. SCIA 2015*; Paulsen, R., Pedersen, K., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9127. [CrossRef]
4. Agarwal, V.; Murali, N.V.; Chandramouli, C. A Cost-Effective Ultrasonic Sensor-Based Driver-Assistance System for Congested Traffic Conditions. *IEEE Trans. Intell. Transp. Syst.* **2009**, *10*, 486–498. [CrossRef]
5. Amaral, V.; Marques, F.; Lourenço, A.; Barata, J.; Santana, P. Laser-Based Obstacle Detection at Railway Level Crossings. *J. Sens.* **2016**, *2016*, 1719230. [CrossRef]
6. Adam, A.; Rivlin, E.; Shimshoni, I.; Reinitz, D. Robust Real-Time Unusual Event Detection Using Multiple Fixed-Location Monitors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 555–560. [CrossRef] [PubMed]

7. Mukojima, H.; Deguchi, D.; Kawanishi, Y.; Ide, I.; Murase, H.; Ukai, M.; Nagamine, N.; Nakasone, R. Moving Camera Background-Subtraction for Obstacle Detection on Railway Tracks. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3967–3971. [CrossRef]
8. Nakasone, R.; Nagamine, N.; Ukai, M.; Mukojima, H.; Deguchi, D.; Murase, H. Frontal Obstacle Detection Using Background Subtraction and Frame Registration. *Q. Rep. RTRI* **2017**, *58*, 298–302. [CrossRef] [PubMed]
9. Tian, Q.; Zhuang, Y.; Yao, C. Efficient railway tracks detection and turnouts recognition method using HOG features. *Neural Comput. Appl.* **2013**, *23*, 245–254. [CrossRef]
10. Schölkopf, B.; Williamson, R.C.; Smola, A.J.; Shawe-Taylor, J.; Platt, J. Support Vector Method for Novelty Detection. In Proceedings of the Advances in Neural Information Processing Systems 12 (NIPS), Denver, CO, USA, 29 November–4 December 1999; pp. 582–588.
11. Núñez, M.; Hernández, F.C.L.; Granados, J.J.R. Automatic Surveillance of People and Objects on Railway Tracks. *Int. J. Interact. Multimed. Artif. Intell.* **2025**, *9*, 107–116. [CrossRef]
12. Niu, H.; Feng, D.; Hou, T. Research on foreign object intrusion detection in railway tracks based on MSL-YOLO. *J. Eng. Appl. Sci.* **2025**, *72*, 136. [CrossRef]
13. Ye, T.; Zhang, X.; Zhang, Y.; Liu, J. Railway Traffic Object Detection Using Differential Feature Fusion Convolution Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1701–1711. [CrossRef]
14. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. 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 (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
15. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 51094–51112.
16. Zhang, S.; Chang, Y.; Wang, S.; Li, Y.; Gu, T. An Improved Lightweight YOLOv5 Algorithm for Detecting Railway Catenary Hanging String. *IEEE Access* **2023**, *11*, 114061–114070. [CrossRef]
17. Yan, P.; Jia, L.; Wang, J.; Xin, Y.; Huang, K. High-speed railway foreign object intrusion detection algorithm based on improved YOLOv7. *Radio Eng.* **2024**, *54*, 1099–1109. [CrossRef]
18. Liu, Z.; Li, Z.; Mofor, R.N.; Ning, D. Unsupervised Anomaly Detection in Railway Catenary Condition Monitoring Using Autoencoders. In Proceedings of the 2020 IEEE Industrial Electronics Society Annual Conference (IECON), Singapore, 18–21 October 2020; pp. 3390–3395. [CrossRef]
19. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009. [CrossRef]
20. Li, C.L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9664–9674. [CrossRef]
21. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Schmidt-Erfurth, U.; Langs, G. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In *Information Processing in Medical Imaging (IPMI 2017)*; Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.T., Alberola-López, C., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10265, pp. 146–157. [CrossRef]
22. Lyu, Y.; Han, Z.; Zhong, J.; Li, C.; Liu, Z. A GAN-Based Anomaly Detection Method for Isoelectric Line in High-Speed Railway. In Proceedings of the 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Auckland, New Zealand, 20–23 May 2019; pp. 1–6. [CrossRef]
23. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
24. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024. [CrossRef]
25. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.