Generative Approach for Detecting Small Intrusive Foreign Objects in High-Speed Railway Scenario

Quan Hao, Rui Shi, Jiaze Li, *Student Member, IEEE*, and Liguo Zhang, *Senior Member, IEEE*

***Abstract*—Foreign object intrusion into high-speed railway (HSR) catenary systems poses severe operational hazards, making effective detection crucial for safety. Precise detection of these small intrusive objects is essential. However, the lack of datasets and research on foreign object intrusion in HSR scenario brings two major challenges: limited data and low accuracy for detecting small intrusive objects. To address these challenges, this paper introduces a novel generative method for detecting foreign object intrusion. To address data limitations, we use low-rank adaptation to fine-tune a diffusion model, developing a generation-extraction-integration framework that generates true-to-reality HSR images of small intrusive target objects. Furthermore, to enhance the detection of small objects in HSR scenario, we propose a new detection model called SA-YOLO. Based on the YOLOv9 architecture, this model optimizes the backbone network using the star operation, an element-wise multiplication method, and introduces the A-DyS module to improve upsampling through dynamic sampling and attention mechanism. Extensive experiments demonstrate that in the HSR scenario our method outperforms existing state-of-the-art approaches in terms of both generation quality and detection performance, while also showing high robustness.**

***Index Terms*—Foreign objects detection, data generation, stable diffusion, high-speed railway.**

I. Introduction

IN RECENT years, expansion of high-speed railway (HSR) networks spanning thousands of kilometers has heightened concerns about systemic safety risks arising from foreign object intrusion into catenary systems. Intrusive foreign objects, small as they are, such as kites and balloons, tend to damage catenary systems or disrupt power supply, causing sensor malfunctions or power failure that compromise HSR operational safety [1], [2]. To address the unique challenges of foreign object intrusion detection within the complex HSR environment, researchers have developed innovative detection methods integrating multi-scale sampling, feature fusion, lightweight extraction, and sparse cross-attention technologies [3], [4], [5], [6], [7], [8], [9]. However, due to their heavy reliance on data-driven deep learning algorithms calling for extensive training datasets and the inherent difficulty in collecting authentic intrusion samples to build comprehensive datasets, the capacity of existing detection models to handle complex HSR environment and learn from diverse modalities of small foreign object intrusion is greatly limited [10], [11], [12], [13], [14]. Our early experimental results demonstrated that YOLOv9 models trained on limited data exhibited obvious underfitting when tasked with identifying various types of foreign objects. Data scarcity substantially hinders the development of detection models, highlighting the urgent need for a robust dataset expansion framework that can overcome these limitations.

The advent of generative artificial intelligence technology has offered promising solutions to data scarcity. Early models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have demonstrated potential in image generation [15]. The recent advancement of Stable Diffusion (SD) [16] has enhanced the controllability of the generation process, allowing for high-quality, customized image generation [17]. Cutting-edge research suggests that using generative methods to augment datasets can significantly improve detection accuracy in environments when there is a lack of positive samples [18]. However, we face great challenges in our attempts to overcome data scarcity by generating images of intrusive foreign objects in HSR scenarios. Currently, there is no effective generation method that can produce image samples suitable for training object detection models. The position, size and quantity of foreign objects often fail to reflect real world conditions, and improvement is needed in both the background and overall style of the the images themselves. Thus, generating true-to-reality images of small foreign objects in HSR environment to address data scarcity in data-driven detection models remains a huge task for us.

In the field of object detection, the YOLO series models facilitate real-time detection while maintaining accuracy using regression-based methods [19]. However, in HSR scenario, intrusive foreign objects are often small in size, incomplete in shape, and sometimes even semi-transparent. The features of foreign objects tend to blend with the background, such as tracks, catenary systems, and natural environment, making it difficult to distinguish the objects from the background for effective and accurate detection. Through our early experiments, we found the performance of existing models to be

Received 7 October 2024; revised 7 May 2025 and 14 August 2025; accepted 19 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U2233211 and Grant 62403017 and in part by Beijing Natural Science Foundation under Grant L243026 and Grant 4244088. The Associate Editor for this article was A. Nunez. *(Corresponding author: Liguo Zhang.)*

Quan Hao, Rui Shi, and Liguo Zhang are with the School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China (e-mail: ml.haoquan@gmail.com; ruishi@bjut.edu.cn; zhangliguo@bjut.edu.cn).

Jiaze Li is with Beijing–Dublin International College, Beijing University of Technology, Beijing 100124, China (e-mail: lijiaze@emails.bjut.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2025.3625181, provided by the authors.

Digital Object Identifier 10.1109/TITS.2025.3625181

Fig. 1. Examples of foreign object intrusion images in HSR scenarios, where "Gen" represents images from the generated dataset and "Real" represents images from the collected real-world dataset.

suboptimal and undesirable. Therefore, we started to work out a new approach to accurately detect small foreign objects, such as balloons, plastic bags, kites, etc. in complex HSR environment.

Built upon results of existing research, our innovative study addresses two limitations mentioned in the above: creating true-to-reality images of small intrusive foreign objects in HSR scenarios to overcome data scarcity, and accurately detecting small foreign objects including semi-transparent ones that easily blend into the background in complex HSR environment. To tackle these challenges, we develop a generative approach. As illustrated in Fig. 2, we first develop a generative framework to augment the dataset of intrusive foreign objects and subsequently enhance the YOLOv9 [20] detection model to improve the detection accuracy of small foreign objects. The implementation is divided into two steps:

Image Generation. As depicted in Fig. 2, we propose an innovative generative framework based on diffusion models. To ensure the precision and controllability of the generated images, we utilize the Stable Diffusion model 1.5 (SD 1.5) [16], which enables better control over noise addition and removal processes. To better align the style characteristics of HSR scenario with the images of intruding foreign objects, we apply Low-Rank Adaptation (LoRA) [21] to fine-tune the SD model using real datasets of HSR scenario and foreign object intrusion. Furthermore, inspired by CLIPSeg [22] and aiming to generate true-to-reality and complex images of small objects, we propose a three-stage generation-extraction-integration method. Ultimately, we seamlessly integrate the extracted foreign objects into HSR scenario, producing true-to-reality composite images. Experimental results show that our method outperforms the existing methods in terms of image quality, data sufficiency, and capability of training detection model.

Object Detection. To improve feature extraction and alignment and detection accuracy, we propose the SA-YOLO model. Images depicting small foreign intruding objects in HSR scenario are often complex, containing elements such as the catenary system, tracks, modern railway equipment, natural background, and foreign objects. To capture the deep layer features of these images, we apply the StarNet backbone network [23], which uses element-wise multiplication to map inputs into a high-dimensional, nonlinear feature space, constructing an implicit feature space of extremely high dimensionality. In addition, we introduce an Attention-based Dynamic Sampling (A-DyS) module during the upsampling stage. This module enhances the capability for hard example mining and improves concentration on small foreign objects, which are often difficult to detect due to blurred boundaries. To focus more on the local features of small foreign objects, the SA-YOLO model optimizes the architecture of YOLOv9 [20] for these specific scenario. Ablation studies demonstrate that enhancements to the backbone network and the upsampling module effectively improve the accuracy of small object detection, significantly augmenting the detection capabilities crucial to ensuring safety in HSR operation.

In summary, our contributions are as follows:

- We propose a generation-extraction-integration image generation framework. By generating true-to-reality images, this framework effectively addresses the data scarcity for training object detection models.
- We introduce the SA-YOLO detection model, featured by the star operation in the backbone and A-DyS module. With these innovations, the detection of small, hard-to-identify objects in complex HSR environment is significantly improved.

Experimental results validate both the effectiveness and robustness of our generative approach, ultimately improving mAP_{50} by 6.9%.

II. Related Work

A. Image Generation

In the field of image generation, early generative AI methods exhibit several limitations. Variational Autoencoders (VAEs) [24], [25], [26], [27] can identify latent representations of data but typically generate low-quality images [28]. Generative Adversarial Networks (GANs) [29], [30], [31], [32], [33] often face challenges such as overfitting and complex training processes [34], [35].

With the advent of diffusion models, these limitations are overcome. Ho et al. introduced the Denoising Diffusion Probabilistic Model (DDPM) [36]. This model improves image quality by gradually adding noise to data and then learning a denoising process to generate new samples. Furthermore, CLIP [37], as one of the multimodal learning techniques, shows impressive versatility in handling complex visual and linguistic tasks through learning associations between images and textual descriptions. Built on these advancements, the Stable Diffusion (SD) model [16] was created, fusing U-net model [38] and cross-attention mechanism [39]. It performs the diffusion process in latent space, significantly reducing computational demands while maintaining high-quality image generation. Recent advancements [40], [41], [42], [43], [44], [45], [46], [47] have further demonstrated the strong capabilities of these models in producing photorealistic images.

However, effectively generating complex multi-object images in this scenario remains a huge challenges. Lüddecke and Ecker introduced CLIPSeg [22], a novel image segmentation model utilizing a transformer-based decoder for precise semantic segmentation and localized image extraction. This

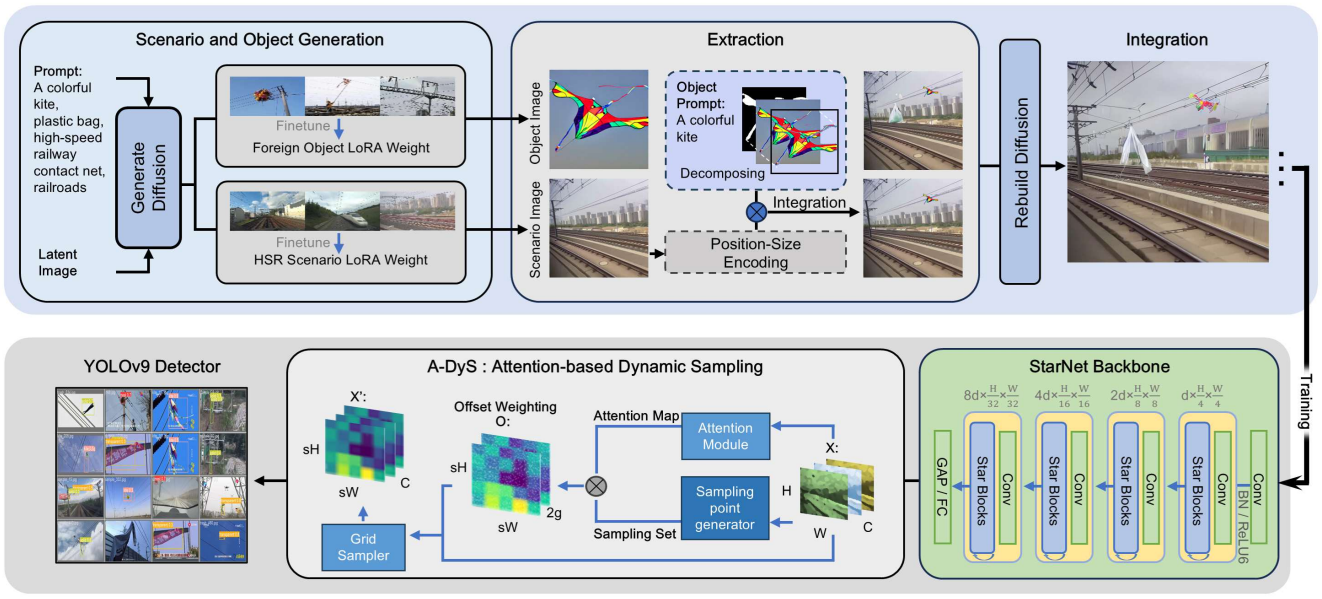


Fig. 2. Overall methodology architecture of the generative enhancement method for detection of intrusive foreign object in HSR scenario: The upper section represents image generation while the lower section represents object detection.

development enables the extraction of small foreign objects and creates new opportunities for generating complex HSR intrusion images.

Despite this advancement, we still see problems in generating high-quality images in HSR scenario. LoRA fine-tuning [21] significantly improves the quality of generated images by applying low-rank decomposition to reduce training parameters while maintaining model efficiency. In a recent study, Choi et al. [48] demonstrated that the LoRA technique for fine-tuning diffusion models worked well to achieve better style alignment and improve generation quality. This technique is expected to show strong capability in generating images in HSR scenario.

B. Object Detection

In recent years, HSR foreign object detection has received widespread attention. Special foreign objects may pose serious safety hazards to the operation of HSR, potentially causing major failure in the overhead catenary system or direct collisions with train components [3], [4], [6]. However, detecting these foreign objects still faces enormous challenges. Due to their unique characteristics and variable appearances, complex lighting changes in railway environment, adverse weather conditions, and image blur as a result of high-speed movement. To address these challenges, researchers have developed various innovative methods. These include intelligent systems using stable sampling modules and feature fusion techniques [3], lightweight feature extraction architectures with adaptive fusion networks [4], sparse cross-attention transformers designed for overhead power systems [5]. Other approaches involve unsupervised methods based on track image symmetry [8] and improved Faster RCNN architectures [6]. The deep SVDD semi-supervised algorithm has achieved significant results in detecting foreign objects on ballastless track beds [7]. Additionally, deep generative methods through adversarial training can detect unknown objects of undefined categories, opening new directions in this field [49].

YOLO series models [50], [51], [52], [53] have been widely applied in object detection and have shown excellent performance, demonstrating outstanding application potential in HSR foreign object detection. The YOLOv5s model combined with railway boundary modeling has demonstrated excellent adaptability and detection efficiency [54]. CF-YOLO [55] has optimized detection capabilities in snowy conditions through cross-fusion blocks, while LR Tiny YOLOv3 [56] has reduced network complexity to meet the requirements of embedded systems. The state-of-the-art (SOTA) model YOLOv9 [20] has further optimized these aspects by introducing programmable gradient information. This innovation addresses the problem of information loss during data transmission in deep networks, thereby improving object detection accuracy and showing significant potential for HSR foreign object detection.

Despite this substantial progress, the detection of small foreign objects such as kites and balloons in HSR environment remains a great challenge. These small foreign objects are difficult to detect primarily because they have limited visible features and tend to blend into the background. The lack of distinctiveness often confuses detection models. Recent advancements have driven notable development in this field, particularly in complex topological relationship processing [57], [58] and hidden layer expansion for abstract scene information representation [58], [59], [60]. Liu et al. [61] developed the DySample approach, which demonstrated enhanced feature extraction capabilities through advanced point sampling methodologies. Moreover, the StarNet framework proposed by Ma et al. [23] implemented element-level multiplication techniques to achieve improved multi-scale feature fusion

outcomes. In addition, Qi et al. enhanced the detection performance of the YOLO architecture in railway scenes by integrating FasterNet and attention mechanisms [62]. Collectively, these technological innovations provide important References for optimizing existing model architectures and improving the detection performance of small objects in challenging environment.

In summary, although prevailing models have significantly improved image quality and foreign object detection, they still face limitations. These include difficulties in merging complex multi-object images, which can lead to poor realism and style alignment, and inadequate performance in accurately detecting small foreign objects in the HSR environment. Our proposed method successfully overcomes these limitations. It aims to refine image styles and extract foreign objects through image segmentation, as well as improve feature alignment and extraction techniques for detection. Ultimately, this work introduces a Generative Approach for Detecting Small Intrusive Foreign Objects in High-Speed Railway Scenario.

III. Generative Method for Creating Dataset of HSR Foreign Objects

A. Generation

First, we construct a Stable Diffusion (SD) model, a generative model based on the latent diffusion process capable of generating high-quality images through noise adding and reducing. This model includes two processes: the forward diffusion process and the reverse generation process. The equations for the forward diffusion process, reverse generation process, and loss function are as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t;\ \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{E}) \tag{1}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_\theta^2(\mathbf{x}_t, t)\mathbf{E}) \tag{2}$$

$$\mathcal{L}_{\text{Diffusion}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right] \tag{3}$$

where $\mathbf{x}_t \in \mathbb{R}^D$ (D is the dimension of the data space) represents the data vector at time step t, α_t controls the degree of noise addition, and t is the scalar time step. $\mathbf{E}$ is the identity matrix. The reverse generation process recovers data from noise, with $\boldsymbol{\mu}_\theta$ and σ_θ^2 being the mean vector and scalar variance parameterized by the neural network θ. The loss function is calculated as the mean squared error (MSE) between the predicted noise vector $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ and the actual noise vector $\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is sampled from a standard Gaussian distribution $\mathcal{N}(0, \mathbf{E})$, and $\mathbf{x}_0$ is the original true image vector.

Next, we introduce low-rank adaptation (LoRA), performing fine-tuning on a dataset constructed from HSR style pictures and images of intrusive foreign object. LoRA decomposes the pretrained weight matrix into two low-rank matrices and introduces a scaling factor α to control the magnitude of updates. The weight updating mechanism is as follows:

$$\mathbf{W}' = \mathbf{W} + \frac{\alpha}{r}(\mathbf{AB}) \tag{4}$$

where $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{m\times n}$ are the pretrained and updated weight matrices, $\mathbf{A} \in \mathbb{R}^{m\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times n}$ are the low-rank matrices, r is the rank (scalar), and α is a scalar controlling the magnitude of adjustment. Adjustments are made by the product of $\mathbf{A}$ and $\mathbf{B}$, controlled by α, ensuring effective adaptation without significantly changing the model structure. Additionally, loss function of LoRA is:

$$\mathcal{L}_{\text{LoRA}} = \lambda_1\mathcal{L}_{\text{task}}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_2\mathcal{L}_{\text{reg}}(\mathbf{A}, \mathbf{B}) \tag{5}$$

where $\mathcal{L}_{\text{task}}$ measures the difference between the generated image vector $\hat{\mathbf{y}} \in \mathbb{R}^D$ and the target image vector $\mathbf{y} \in \mathbb{R}^D$, and $\mathcal{L}_{\text{reg}}$ is a regularization term to prevent overfitting. $\lambda_1, \lambda_2 \in \mathbb{R}$ balance their contributions. During the fine-tuning process, we freeze the pretrained weights $\mathbf{W}$ and train only the low-rank matrices $\mathbf{A}$ and $\mathbf{B}$. This reduces the discrepancy between the generated images and the target domain, ensuring the model retains its pretrained capabilities while learning new features and enhancing performance.

Following this method, we fine-tuned the SD model on images of HSR backgrounds and relevant intrusive foreign objects. The fine-tuning enables the generative model to produce images with complex HSR background features and enhances its capability to generate true-to-reality images of intrusive foreign objects.

B. Extraction

To further embed the foreign object into complex HSR background images, we aim to first extract a clean version of the foreign object. As shown in Fig. 3 (b), we encode the prompt and the original image separately. This process introduces a CLIP Transformer block, where the text encoder and image encoder map the prompt description and the original image to the same embedding space. Then, we use contrastive learning to obtain a prior latent representation and calculate the cosine similarity between the image and text embeddings, as shown in the following equation:

$$\text{sim}(\mathbf{v}_i, \mathbf{t}_i) = \frac{\mathbf{v}_i^\top \mathbf{t}_i}{\|\mathbf{v}_i\|_2\|\mathbf{t}_i\|_2} \tag{6}$$

where $\mathbf{v}_i \in \mathbb{R}^k$ represents image embedding vector, and $\mathbf{t}_i \in \mathbb{R}^k$ represents text embedding vector. In this process, the loss function of the model depends not only on the latent space of the original image but also on the conditioned latent embedding.

The loss function $\mathcal{L}_{\text{CLIP}}$ is calculated based on the similarity of image-text pairs and Softmax normalization as follows:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)} + \log\frac{\exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_j)/\tau)}\right] \tag{7}$$

where N is the batch size, and $\tau \in \mathbb{R}$ is the temperature parameter. By minimizing the contrastive loss function $\mathcal{L}_{\text{CLIP}}$, we increase the similarity of matching pairs and reduce the similarity of non-matching pairs, effectively aligning image and text features.

During the encoding process, we use the intermediate activation layers of the CLIP model to extract different levels of feature vectors (projections). These extracted feature vectors are then input into a decoder, which uses Feature-wise Linear

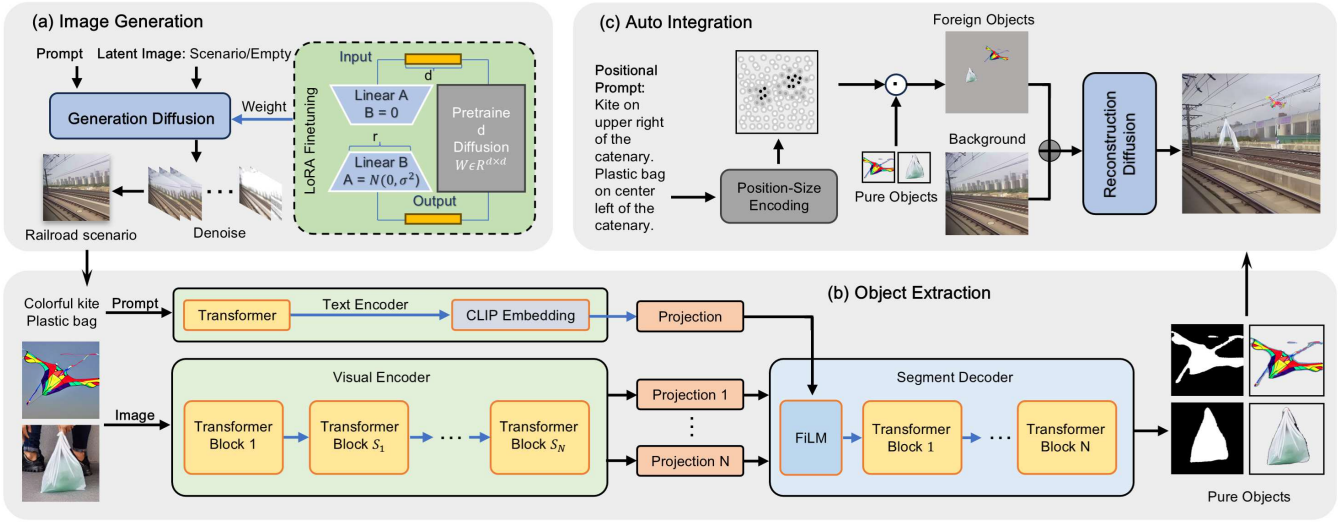


Fig. 3. Detailed diagram of the HSR foreign object intrusion image generation method: (a) LoRA fine-tuning and image generation process, (b) foreign object extraction process, and (c) automatic integration of foreign object intrusion images.

Modulation (FiLM) to modulate the features. Specifically, the scaling parameters $\boldsymbol{\gamma}(\mathbf{t})$ and shifting parameters $\boldsymbol{\beta}(\mathbf{t})$ are generated from the conditional vector $\mathbf{t} \in \mathbb{R}^k$, and the feature modulation process is as follows:

$$\mathrm{FiLM}(\mathbf{M}, \mathbf{t}) = \boldsymbol{\gamma}(\mathbf{t}) \odot \mathbf{M} + \boldsymbol{\beta}(\mathbf{t}) \tag{8}$$

$$\boldsymbol{\gamma}(\mathbf{t}) = \mathbf{W}_\gamma \mathbf{t} + \mathbf{b}_\gamma, \quad \boldsymbol{\beta}(\mathbf{t}) = \mathbf{W}_\beta \mathbf{t} + \mathbf{b}_\beta \tag{9}$$

where $\mathbf{M}$ represents the intermediate layer activations (potentially a tensor or feature map), $\mathbf{t} \in \mathbb{R}^k$ is the conditional vector (text embedding), and $\boldsymbol{\gamma}(\mathbf{t}), \boldsymbol{\beta}(\mathbf{t})$ are the scaling and shifting parameter vectors generated from $\mathbf{t}$, compatible with dimensions of $\mathbf{M}$. $\mathbf{W}_\gamma, \mathbf{W}_\beta$ are weight matrices, and $\mathbf{b}_\gamma, \mathbf{b}_\beta$ are bias vectors. $\odot$ denotes element-wise multiplication. The output is then mapped through a sigmoid activation function to obtain the foreign object mask:

$$\mathbf{M}_{\text{mask}} = \sigma\Big(\text{Decoder}\Big(\{\text{CLIP}(\mathbf{I}_{\text{rail}})_{l_k}\}_{k=1}^{L}, \mathbf{t}\Big)\Big) \tag{10}$$

where $\mathbf{M}_{\text{mask}}$ represents the foreign object mask tensor (with spatial dimensions matching $\mathbf{I}_{\text{rail}}$ and values in [0, 1]), indicating the area in the image tensor $\mathbf{I}_{\text{rail}}$ corresponding to the text prompt T_{object}. The mask is generated using the following components: $\mathbf{t}$ is the text embedding vector derived from T_{object}. $\{\text{CLIP}(\mathbf{I}_{\text{rail}})_{l_k}\}_{k=1}^{L}$ is the set of feature vectors for $\mathbf{I}_{\text{rail}}$ extracted from different layers l_k of the CLIP model. Decoder is the decoder network that uses FiLM internally, conditioned on $\mathbf{t}$, to process the CLIP features. σ is the sigmoid activation function applied to the output of the decoder. We then use the mask $\mathbf{M}_{\text{mask}}$ to process the foreign object image tensor $\mathbf{I}_{\text{object}}$, obtaining a clean foreign object image with a transparent background. Through these steps, the automatic segmentation and extraction of images of foreign objects is accomplished.

C. Reconstruction

Foreign object intrusions often occur on railway catenaries or tracks. When merging foreign objects with the background, it is important to consider the real condition of the position and size of objects. To achieve automated fusion and reconstruction, we design a position-size encoding fusion method that seamlessly integrates the foreign objects into the HSR background, ultimately generating true-to-reality images of foreign objects entangled with HSR catenary system.

Firstly, inspired by cutting-edge research and widespread applications of large language models [63], we use Qwen 1.5 [64] to process prompts, generating reasonable position coordinate vectors and size vectors. Then, based on the size vector, we adjust the size of the clean foreign object image tensor $\mathbf{I}_{\text{object}}$ to get $\mathbf{I}'_{\text{object}}$. Based on the position vector, we generate a coordinate map matrix $\mathbf{M}_{\text{map}}$, where its elements $\mathbf{M}_{\text{map}}(i, j) \in \{0, 1\}$ indicate whether the coordinates (i, j) are within the covered area of the foreign objects. Finally, we overlay the processed foreign object image $\mathbf{I}'_{\text{object}}$ onto the background image tensor $\mathbf{I}_{\text{background}}$, resulting in the final image tensor $\mathbf{I}_{\text{final}}$. The calculation process is as follows:

$$\mathbf{I}_{\text{final}} = \mathbf{I}_{\text{background}} \odot (\mathbf{1} - \phi(\mathbf{M}_{\text{map}})) + (\mathbf{I}'_{\text{object}} \odot \phi(\mathbf{M}_{\text{map}})) \tag{11}$$

where $\odot$ represents pixel-wise multiplication, $\mathbf{1}$ is a tensor of ones with the same dimensions as $\mathbf{I}_{\text{background}}$, and ϕ is an activation function mapping the values in $\mathbf{M}_{\text{map}}$ to 0 or 1. Thus, where $\mathbf{M}_{\text{map}}$ equals 1, $\mathbf{I}'_{\text{object}}$ is overlaid on $\mathbf{I}_{\text{background}}$, and where $\mathbf{M}_{\text{map}}$ equals 0, $\mathbf{I}_{\text{background}}$ is retained, ensuring the foreign object image is correctly overlaid onto the background image, thereby generating the final composite image.

As shown in Fig. 3 (c), we effectively integrate clean foreign object images with the HSR scene images. Although the foreign objects are seamlessly incorporated into the background, the real status of environmental interactions requires further refinement. Consequently, we further enrich the prompts, introduce noise to the composite images, and reprocess them using the SD model for reconstruction, ultimately generating true-to-reality images of HSR foreign intrusive objects.

To address real-world challenges, we also consider adverse weather conditions. We developed extreme weather masks for heavy rain and dense fog using computer vision techniques. By modifying lighting effects, shadow properties, and color parameters, we integrated these masks into the original images before generation and reconstruction. This process yielded images under adverse weather conditions.

IV. SA-YOLO for HSR Intrusive Foreign Object Detection

A. StarNet Backbone

In the HSR foreign object intrusion scenario, some objects (such as plastic bags) are semi-transparent and small in size. The complex background of HSR comprises various elements such as catenaries, tracks, railway barriers, and natural landscapes, making small foreign objects easily blend with the background and hence difficult to detect. This poses great challenges for existing object detection models, particularly in feature extraction and alignment. Drawing theoretical inspiration from StarNet, we developed an innovative star operation of elementwise multiplication to construct a sophisticated, high-dimensional nonlinear feature representation. Based on this principle, we strategically refined the YOLOv9 backbone architecture, significantly enhancing its capability in extracting complex image features and mining hard samples.

By stacking multiple layers with width d, StarNet can implicitly generate a feature space belonging to $\mathbb{R}^p$, where the dimension p follows the exponential relationship $p = (d/\sqrt{2})^{2^l}$ after l star operation blocks. For example, given a 10-layer ($l = 10$) isotropic network with a width $d = 128$, the resulting implicit feature dimension p approximates 90^{1024}. This dimension is extremely large and can be reasonably approximated as infinite. Consequently, the layered stacking process enables substantial exponential amplification of the implicit feature dimension through the star operation.

Fig. 2 illustrates the architecture of the StarNet backbone. It comprises four main stages, each composed of Star Blocks and standard convolutional layers. Given input features with spatial dimensions $H \times W$, the backbone processes these features sequentially through the four stages. Across these stages, the spatial resolution is systematically reduced to $H/32 \times W/32$, while the number of feature channels increases progressively, scaling with a base dimension parameter d up to a maximum of $8d$. This design progressively builds richer, more abstract feature representations by increasing channel depth while reducing spatial redundancy. Leveraging the Star Block modules and their use of elementwise multiplication, StarNet effectively maps inputs to a high-dimensional nonlinear feature space. This enhances the capacity of the model to comprehend complex scene patterns and multi-scale feature interactions, demonstrating exceptional performance in detecting small foreign object intrusions in HSR scenarios.

B. A-DyS Upsampling Module

For precise feature map reconstruction during upsampling, we propose A-DyS, a high-precision upsampling module designed to address scenarios where small objects are camouflaged within complex backgrounds, and therefore hard to identify. As illustrated in Fig. 4(b), its core principle involves generating dynamic offsets using an attention mechanism and then dynamically resampling the feature map.

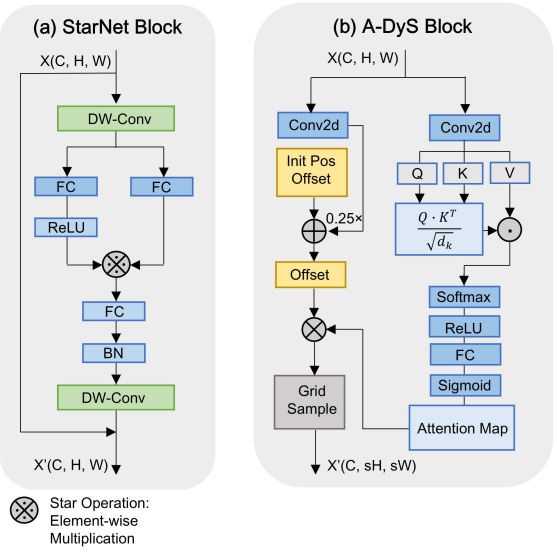


Fig. 4. SA-YOLO detailed design. (a) StarNet block, where ⋇; (star operation) means element-wise multiplication; (b) Attention-based Dynamic Sample (A-DyS) block.

The A-DyS module computes dynamic offsets for the feature map, performing fine-grained spatial sampling on the input features. This adaptive sampling process captures detailed information, enabling the extracted feature vectors to reflect deeper image characteristics, thus enhancing the ability of the model to identify small foreign objects.

Once the dynamic offsets are calculated, we incorporate the attention mechanism. This mechanism first computes attention information based on the input feature map, then uses this information to adjust the dynamic sampling grid, allowing the model to focus on critical feature regions. This approach significantly improves the extraction of features of small objects.

The A-DyS module, detailed in Algorithm 1, refines the upsampling process through attention-guided dynamic sampling. Initially, a dynamic offset tensor, $\mathbf{W}_{\text{offset}}$, is computed from the input features $\mathbf{X}$ via a 2D convolution (Conv2d). This offset adjusts an initial grid $\mathbf{G}$ to yield a preliminary sampling grid $\mathbf{S}$. Concurrently, Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) tensors are generated from $\mathbf{X}$ through convolutional layers. These tensors facilitate the computation of a spatial attention map $\mathbf{A}$ via scaled dot-product attention, followed by an activation function σ. Crucially, this attention map $\mathbf{A}$ spatially modulates the preliminary grid $\mathbf{S}$ to produce the final attention-weighted sampling grid $\mathbf{S}'$. This step effectively directs the sampling process towards more salient feature regions identified by the attention mechanism. Prior to sampling, this weighted grid $\mathbf{S}'$ is normalized (Normalizing) based on the input feature map's spatial dimensions (H, W) to create $\mathbf{S}_{\text{norm}}$. Subsequently, a

Algorithm 1 A-DyS Algorithm (Revised Grid Variables)

Input: Feature map $\mathbf{X}$, Initial position grid $\mathbf{G}$, Upsampling scale s

Output: Upsampled feature map $\mathbf{X}'$

Initialization:

$\mathbf{W}_{\text{offset}} \leftarrow \text{Conv2d}(\mathbf{X})$

$\mathbf{S} \leftarrow \mathbf{G} + 0.25 \times \mathbf{W}_{\text{offset}}$

Attention-based dynamic sampling:

$\mathbf{Q}, \mathbf{K}, \mathbf{V} \leftarrow \text{Conv2d}(\mathbf{X})_{Q,K,V}$

$\mathbf{A} \leftarrow \sigma\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V}$

$\mathbf{S}' \leftarrow \mathbf{S} \times \mathbf{A}$

$B, C, H, W \leftarrow$ dimensions of $\mathbf{X}$

$\mathbf{S}_{\text{norm}} \leftarrow \text{Normalizing}(H, W, \mathbf{S}', \mathbf{X})$

$\mathbf{F} \leftarrow \text{GridSampling}(\mathbf{S}_{\text{norm}}, \mathbf{X})$

$\mathbf{X}' \leftarrow \text{Reshape}(\mathbf{F}, (B, -1, sH, sW))$

return $\mathbf{X}'$

GridSampling operation utilizes $\mathbf{S}_{\text{norm}}$ to interpolate features from the original map $\mathbf{X}$, yielding an intermediate feature representation $\mathbf{F}$. Finally, $\mathbf{F}$ is reshaped (Reshape) to match the target upsampled spatial resolution ($sH \times sW$), producing the output feature map $\mathbf{X}'$. By dynamically adjusting sampling locations based on feature attention, A-DyS module enhances the extraction of fine-grained details, proving particularly beneficial for detecting small foreign objects.

V. Experiment

A. Experimental Setup

1) Dataset and Preprocessing: We developed a comprehensive dataset of 543 high-quality images depicting foreign object intrusions in the high-speed railway (HSR) scenario through rigorous collection and filtering processes. The dataset comprises 84 images of kites, 79 of balloons, 122 of semi-transparent objects (such as plastic bags), 124 of opaque objects (including banners and fabrics), and 134 of natural foreign objects (like branches and nests). From this collection, we selected 100 representative images to form a validation set for evaluating detection accuracy of our model in real-world scenarios. For all experiments, we consistently used this real-world validation dataset to evaluate performance.

To evaluate performance of our method, we constructed datasets for generation and validation experiments. Given the limited size of datasets, we focused on evaluating methods through the validation set.

Additionally, we used computer vision techniques to create an adverse weather dataset by adding rainfall and fog effects to images, further evaluating the generalization performance of our method.

a) Real Dataset: We collected 543 images of real-life small target foreign object intrusions, including 443 images for training and 100 for validation.

b) Gen Dataset: We combined the Real dataset with an increasing number of generated images, ranging from 100 to 900.

c) Adverse Weather Dataset: We transformed our datasets with rainfall and heavy fog effects, comprising 443 real images and 400 generated images (totaling 843) for training, while maintaining 100 consistent images for validation.

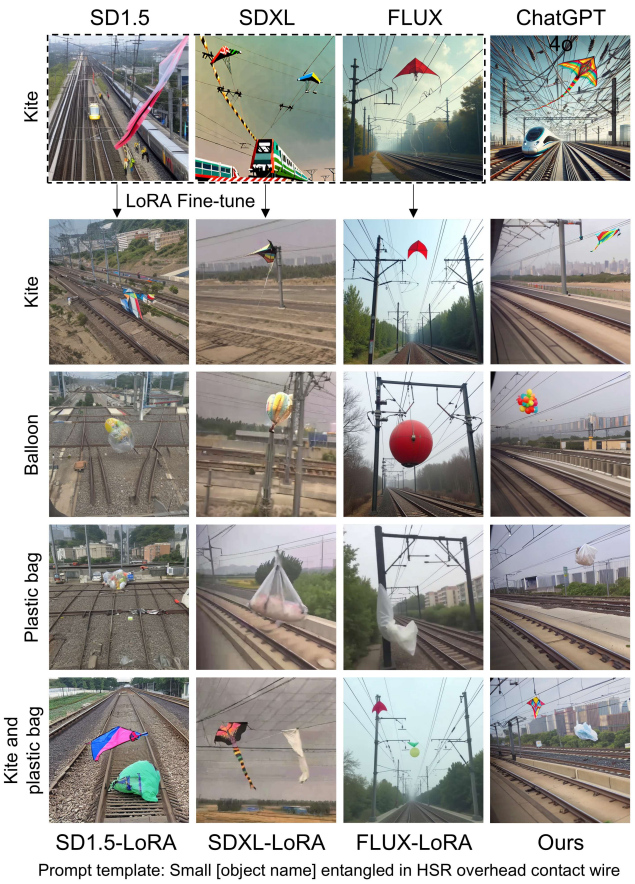


Fig. 5. Comparison of generated effects with SOTA models. The upper part shows the original outputs generated by baseline SOTA models: SD 1.5, SDXL, FLUX, and ChatGPT 4o. The lower part contrasts the results from our method with the outputs of these SOTA models (SD 1.5, SDXL, FLUX) after LoRA fine-tuning.

2) Evaluation Metrics: Mean Average Precision (mAP) aggregates AP scores across categories and thresholds to measure object detection performance comprehensively. Specifically, mAP_{50} represents mean AP at an IoU threshold of 0.50, while $mAP_{50\text{-}95}$ averages over thresholds from 0.50 to 0.95 in 0.05 increments, assessing detection at varying strictness levels.

3) Experimental Environment and Hyperparameters: Our experiments were conducted with eight NVIDIA RTX 2080Ti GPUs. We performed LoRA fine-tuning with a rank of 32. During training, the batch size was set to 32, and the initial learning rate was established at 0.001. The variability observed in our experiment is approximately $\pm 0.3\%$.

B. Evaluation of Image Generation Effects

We evaluated our proposed method by comparing it with several State-of-the-Art (SOTA) models: SD 1.5 [71], SDXL [72], FLUX [73], and ChatGPT-4o [74], using identical prompts. As shown in Fig. 5, we presented both original SOTA outputs and results after fine-tuning SD 1.5, SDXL, and FLUX on our custom HSR foreign object intrusion dataset.

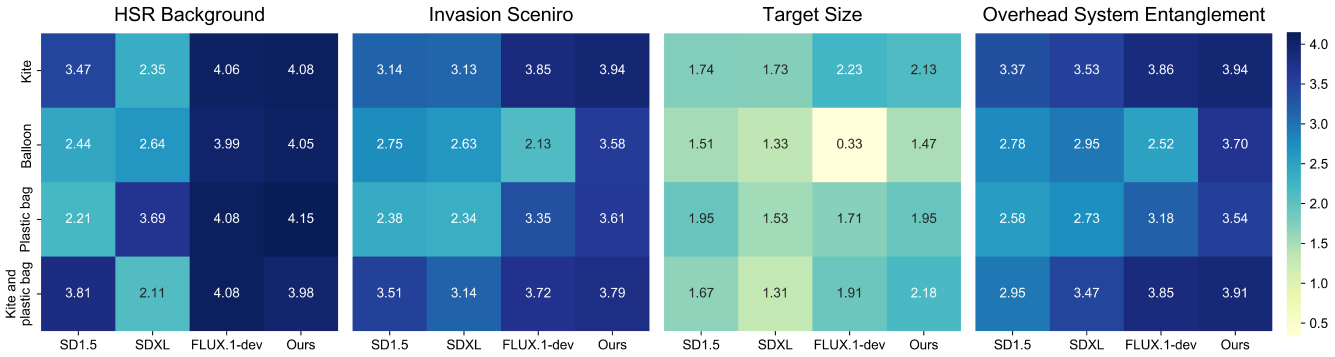


Fig. 6. Detailed realism evaluation results for 16 generated images are presented. From left to right, the evaluation results show the high-speed railway (HSR) background effect, intrusion scene effect, and foreign object characteristics (size and catenary entanglement). These results are visualized using heatmap blocks corresponding to Fig. 5, in which darker colors signify better performance.

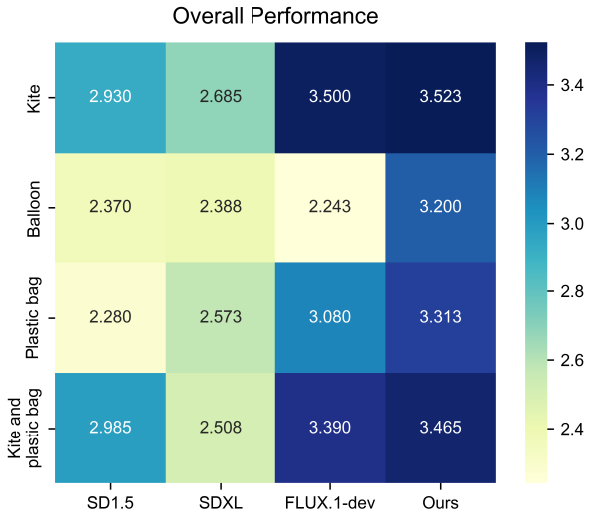


Fig. 7. Comprehensive evaluation results of the realism performance for 16 generated images, with the heatmap blocks corresponding to Fig. 5. In this heatmap, darker colors represent better performance.

TABLE I
AVERAGE SCORES OF FOUR FINE-TUNED IMAGE GENERATION METHODS

LoRA Fine-tuned Model	SD 1.5	SDXL	FLUX	Ours
Average Score	2.641	2.539	3.053	**3.375**

ChatGPT-4o was used only for baseline comparison as it is a closed-source model and lacked a fine-tuning interface. The results qualitatively show improved realism in fine-tuned models, especially for HSR background, with performance generally scaling with model size among the SOTAs.

To quantitatively assess realism, we conducted human evaluations via questionnaires, scoring images from 1 (poor) to 5 (excellent) on four key aspects: (1) HSR background quality, (3) overall intrusion scene realism, (2) small target object characteristic adherence, and (4) catenary entanglement effects. As shown in Fig. 6, these scores are presented using heatmaps, where darker colors signify better performance.

The evaluation results, illustrated in Fig. 6 and Fig. 7, demonstrate that our method significantly outperforms the fine-tuned SD 1.5 and SDXL models across all four evaluated aspects. A notable comparison involves the fine-tuned FLUX model. Despite operating at a scale more than four times larger than our base model, the FLUX model only approached our method's performance in HSR background generation. Furthermore, our method maintained superior results in the other crucial aspects. Overall, our approach achieved a composite score of 3.375, as detailed in Table I. This score is markedly higher than those of the compared SOTA models, representing a substantial improvement of 0.73 (27.8%) over the base SD 1.5 and 0.322 points (10.5%) over the significantly larger fine-tuned FLUX.

The importance of fine-tuning on our specialized dataset using LoRA is evident from Fig. 5. This process significantly improves the generation capabilities of existing generative models. Among these fine-tuned comparative models, FLUX performed best, while the performance of SD 1.5 was notably hampered by its limited model scale. In contrast stood our multi-step generation method, built upon the foundational SD 1.5 architecture. It exhibited markedly superior generation quality. This level of performance substantially surpassed that of all other models compared, including their fine-tuned versions. This overall result strongly validates the effectiveness and necessity of our proposed multi-step generation strategy.

C. Evaluation of Object Detection Performance

To validate the effectiveness of our generated image data and model architecture, we conducted a comprehensive comparison of the SA-YOLO method against existing SOTA models, including CNN-based YOLO series models, the Transformer-based RT-DETR architecture, and YOLOv9 enhancement models.

Experimental results demonstrated that when trained solely on the Real dataset, our SA-YOLO model achieved an mAP_{50} of 74.0 and an $mAP_{50\text{-}95}$ of 40.7. Although YOLOv12 showed marginally higher mAP_{50} (74.3), our model tied with MobileNetV4 for the best $mAP_{50\text{-}95}$ performance, indicating SA-YOLO's excellent detection precision on the Real dataset. The CNN-based YOLO family models generally performed

TABLE II
COMPARISON OF DIFFERENT SOTA DETECTION MODELS TRAINED WITH THE REAL DATASET AND THE GEN DATASET

Model	Real		Real + Gen400		Gen Dataset Improvement	
	mAP_{50}	$mAP_{50\text{-}95}$	mAP_{50}	$mAP_{50\text{-}95}$	ΔmAP_{50}	ΔmAP_{50-95}
YOLOv8 [65]	69.4	38.9	74.8	43.1	+5.4%	+4.2%
YOLOv9 [20]	69.8	37.4	<u>75.9</u>	43.1	+6.1%	+5.7%
YOLOv10 [66]	69.9	40.3	72.8	<u>43.8</u>	+2.9%	+3.5%
YOLOv11 [67]	72.9	40.3	75.7	43.4	+2.8%	+3.1%
YOLOv12 [68]	**74.3**	40.2	75.4	41.4	+1.1%	+1.2%
RT-DETR [69]	67.4	38.7	73.2	41.9	+5.8%	+3.2%
MobileNetV4 [70]	73.9	**40.7**	75.8	43.0	+1.9%	+2.3%
StarNet [23]	73.2	40.3	73.6	42.6	+0.4%	+2.3%
SA-YOLO (Ours)	<u>74.0</u>	**40.7**	**76.7**	**45.5**	+2.7%	+4.8%

*The best results are marked in **bold**, and the second ones are marked with <u>underlined</u>.*

well, while RT-DETR achieved a relatively lower performance with mAP_{50} of 67.4, suggesting that for this particular task, CNN-based architectures might be more suitable than Transformer-based ones.

When integrating generated data (Real + Gen400), all models showed improvements, confirming the value of our generated data approach. SA-YOLO's advantages significantly expanded after using generated images, achieving the highest performance among all models, with mAP_{50} and $mAP_{50\text{-}95}$ reaching 76.7 and 45.5 respectively. Compared to training with real data alone, these metrics improved by 2.7% and 4.8% respectively, outperforming all comparison models in absolute terms. Notably, RT-DETR, despite its lower baseline performance, demonstrated substantial improvement (+5.8% in mAP_{50}), indicating that our generated data could enhance diverse model architectures. The YOLOv9 enhancement models showed varying responses to generated data, with MobileNetV4 demonstrating more substantial improvement than StarNet. However, YOLOv12 showed the most limited improvement with generated data (mAP_{50} increased by only 1.1%), potentially due to its residual efficient layer aggregation networks (R-ELAN) structure which prioritized generalized stability over specialized data adaptation. These results conclusively demonstrated that SA-YOLO possessed exceptional adaptability and learning capability when integrating generated data, efficiently extracting features from diverse data sources to enhance detection precision. This experiment validated not only the quality of our generated data but also the architectural innovations of SA-YOLO, particularly effective for the HSR foreign object intrusion detection.

D. Ablation Studies

To evaluate the effectiveness of our proposed components and the impact of using the Gen dataset to enhance training, we conducted comprehensive ablation studies.

The results were summarized in Table III. The standard YOLOv9 model served as our baseline, achieving a mAP_{50} of 69.8 and a $mAP_{50\text{-}95}$ of 37.4, utilizing 51.18M parameters and requiring 239.9 GFLOPs. Replacing the baseline backbone with StarNet improved mAP_{50} to 73.2 and $mAP_{50\text{-}95}$ to 40.3, while reducing parameters to 41.86M and GFLOPs to 193.7. This confirmed the ability of StarNet to enhance both accuracy and efficiency. Integrating only the A-DyS module increased mAP_{50} to 73.5 and $mAP_{50\text{-}95}$ to 40.6, with slightly increased computational requirements (51.27M parameters, 240.0 GFLOPs). Combining StarNet and A-DyS to create SA-YOLO improved accuracy (mAP_{50} to 74.0, $mAP_{50\text{-}95}$ to 40.7) while maintaining efficiency (41.94M parameters, 193.8 GFLOPs), indicating effective synergy between components.

Using the Gen dataset to enhance training of the baseline model yielded significant improvements: mAP_{50} increased to 75.9 (6.1% increase) and $mAP_{50\text{-}95}$ to 43.1, with no additional inference overhead. When training enhancement with the Gen dataset was applied to the StarNet-enhanced model, mAP_{50} improved to 73.6 and $mAP_{50\text{-}95}$ to 42.6. Similarly, for the A-DyS-enhanced model, training with the Gen dataset boosted mAP_{50} to 76.3 and $mAP_{50\text{-}95}$ to 44.7.

Our complete method, SA-YOLO with training enhanced by the Gen dataset, achieved optimal performance with 76.7 mAP_{50} and 45.5 $mAP_{50\text{-}95}$. This represents significant improvements over the baseline (mAP_{50} by 6.9%, $mAP_{50\text{-}95}$ by 8.1%) while requiring fewer parameters (41.94M vs. 51.18M) and GFLOPs (193.8 vs. 239.9). These studies demonstrate both the individual and synergistic effectiveness of our architectural components and highlight the critical role of Gen dataset-enhanced training in significantly improving detection performance.

E. Image Generation Prompt Ablation Studies

To evaluate the impact of prompting strategies on image generation quality and subsequent detection performance, we designed an ablation experiment featuring three distinct prompt configurations: Full Prompt, Positional Prompt, and Minimal Prompt. The Full Prompt configuration included detailed descriptions, positioning information, and negative prompts. The Positional Prompt configuration used only simple object descriptions and position information. Lastly, the Minimal Prompt configuration contained only the names of the foreign object and the background. For this experiment, we generated images across all three groups while maintaining consistency in background, object type, and position. Due to the inherent randomness in image generation, we implemented a rigorous quality control process to ensure data integrity and experimental fairness. We generated 1000 images for each

TABLE III
ABLATION STUDY RESULTS

YOLOv9	StarNet	A-DyS	Gen	mAP_{50}	$mAP_{50\text{-}95}$	Layers	Parameters	GFLOPs
✓				69.8	37.4	962	51.18M	239.9
✓	✓			73.2	40.3	796	41.86M	193.7
✓		✓		73.5	40.6	978	51.27M	240.0
✓	✓	✓		**74.0**	**40.7**	812	41.94M	193.8
✓			✓	75.9	43.1	962	51.18M	239.9
✓	✓		✓	73.6	42.6	796	41.86M	193.7
✓		✓	✓	76.3	44.7	978	51.27M	240.0
✓	✓	✓	✓	**76.7**	**45.5**	812	41.94M	193.8

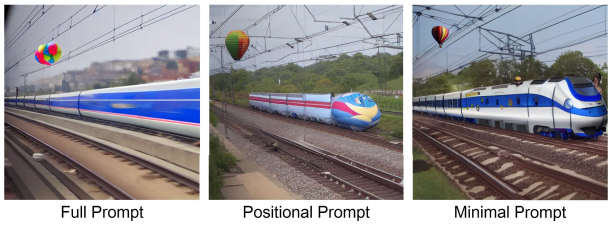


Fig. 8. Comparison of generated effects with different prompts.

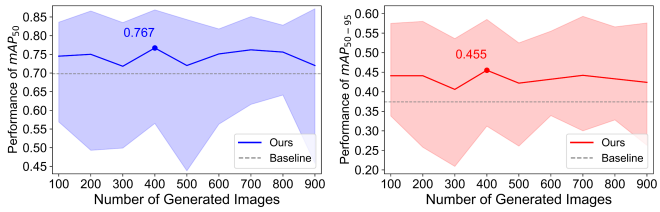


Fig. 9. Effect of enhanced training with varying numbers of generated data on detection performance. The solid line shows average performance, the colored area indicates detection thresholds for five foreign objects, and the dashed line represents YOLOv9 baseline.

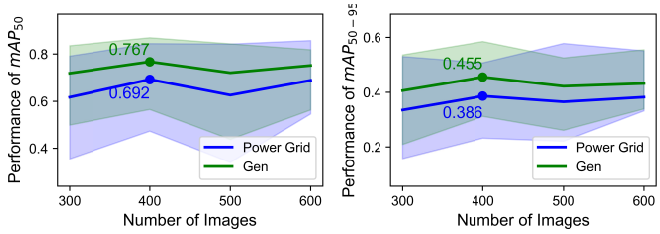


Fig. 10. Comparison of detection results using power grid foreign object data. Two solid lines represent performance across two datasets, with the colored area indicating detection thresholds for five foreign object types.

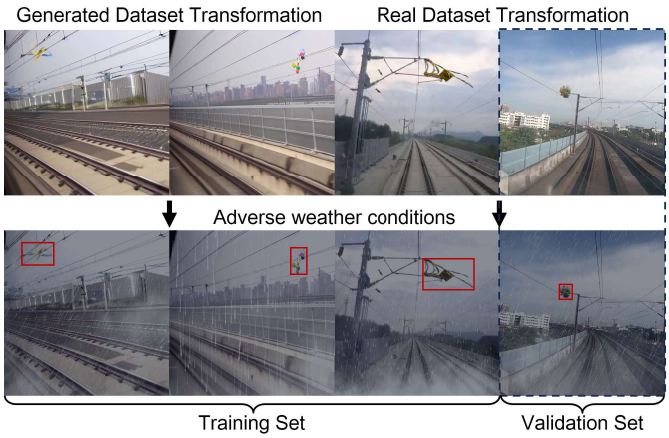


Fig. 11. Adverse weather transformation process. The top row displays original images, while the bottom row shows their counterparts transformed under rain and fog conditions. Images on the left are synthetically generated, whereas images on the right are captured from real environment. The validation dataset was constructed exclusively using the authentic images shown in the rightmost column.

TABLE IV
SA-YOLO PERFORMANCE ACROSS DIFFERENT PROMPT CONDITIONS

Condition	mAP_{50}	$mAP_{50\text{-}95}$	Failure Rate
Real Images	74.0	40.7	-
Minimal Prompt	72.3 (-1.7%)	41.4 (+0.7%)	12.25%
Positional Prompt	70.9 (-3.1%)	42.4 (+1.7%)	17.75%
Our Method	**76.7 (+2.7%)**	**45.5 (+4.8%)**	0.50%

prompt group, which were randomly shuffled and evaluated by three independent assessors using the four criteria from Section B. We selected the top 300 highest-scoring images from each group for object detection experiments.

Our analysis revealed distinct visual limitations in images generated using secondary prompts compared to the Full Prompt strategy. As shown in Fig. 11, train heads exhibited balloon-like appearances under suboptimal prompting conditions. We attribute this phenomenon to several factors. First, insufficient text-image feature alignment within SD 1.5 causes the model to struggle with sparse prompt features, resulting in cross-associated features during fusion. Second, balloons and streamlined train heads possess inherent similarities as both follow fluid dynamics principles, creating natural visual parallels that confuse the generation process. Additionally, cross-multiplicative modeling of positional and object features may interfere with prompt interpretation when features lack clarity. Despite these challenges, our comprehensive Full Prompt strategy effectively mitigates these issues through in-depth analysis of HSR scenarios and train characteristics. Additional generated examples can be found in Fig. A.1 in the appendix (See the Supplementary Material).

As shown in Fig. 8 and Table IV, prompting strategies significantly influenced both image generation quality and object detection performance. Comprehensive prompt engineering in our method produced photorealistic intrusion scenarios,

resulting in superior detection metrics with mAP_{50} of 76.7 (2.7% improvement over real images) and $mAP_{50\text{-}95}$ of 45.5 (4.8% improvement). In contrast, the Minimal Prompt group yielded lower mAP_{50} at 72.3 (1.7% decrease) despite a slight increase in $mAP_{50\text{-}95}$ to 41.4 (0.7% improvement). The Positional Prompt group showed further performance degradation with mAP_{50} dropping to 70.9 (3.1% decrease), though $mAP_{50\text{-}95}$ improved to 42.4 (1.7% increase).

Notably, our method achieved a remarkably low failure rate of just 0.50%. Here, failure referred to instances where the model failed to generate proper foreign object intrusion images. In comparison, the Minimal Prompt group had a failure rate of 12.25%, and the Positional Prompt group showed an even higher failure rate of 17.75%. Since our method was based on foreign object mask map fusion and image reconstruction, position prompts had minimal impact on the final object placement. This validated the robustness of our generation approach for small target foreign object intrusion scenarios. These results demonstrated that well-engineered prompts not only enhanced the visual quality of generated images but also substantially improved object detection performance.

F. 100-Step Generated Image Data Experiment

This experiment evaluated the impact of varying quantities of generated images on the detection of small target foreign objects in the HSR scenario. By incrementally adding 100 generated images at a time to the training dataset, up to a total of 900 images, we assessed the performance of the model using the mAP_{50} and $mAP_{50\text{-}95}$ metrics.

The results showed that incorporating up to 400 generated images significantly improved detection performance, with mAP_{50} and $mAP_{50\text{-}95}$ reaching 76.7 and 45.5 respectively. However, beyond 400 images, performance began to decline, likely due to increased noise and overfitting. Interestingly, at 300 and 500 images, the training results deviated from the overall trend, possibly due to high similarity among the sampled images. This issue could be addressed through improved sampling methods in future work.

Notably, with 400 generated images, the ratio of generated to real data reached 1:1, which indicated an ideal data augmentation strategy. These findings suggest that effectively managing both the quantity and quality of generated data is crucial for optimizing model performance.

G. Comparative Analysis of Gen Dataset Versus Power Grid Dataset in Training the SA-YOLO Model

This study compared the impact of generated datasets (Gen) and power grid foreign object datasets (Power Grid) on training the SA-YOLO model. Notably, the power grid dataset performed well when 400 images were used, achieving mAP_{50} and $mAP_{50\text{-}95}$ scores of 69.2 and 38.6 respectively. By evaluating datasets with 300, 400, 500, and 600 additional images (either generated or from power grid sources) added to the real dataset, the results showed that generated images significantly improved model performance compared with the power grid data. Specifically, in the case of 400 images, the performance difference reached its peak, with mAP_{50} and $mAP_{50\text{-}95}$ gaps of 7.5% and 6.9% respectively. These results indicated that, compared with foreign object intrusion data from specific scenarios, the generated data contained more effective features. This further proves the effectiveness of our proposed generative method in addressing sample scarcity and underscores the superior capability of generated data in enhancing detection performance.

TABLE V
COMPARISON OF SA-YOLO AND BASELINE MODELS TRAINED WITH THE ADVERSE WEATHER DATASET

Model	Real (Transformed)		Real + Gen400 (Transformed)	
	mAP_{50}	$mAP_{50\text{-}95}$	mAP_{50}	$mAP_{50\text{-}95}$
YOLOv9	68.6	36.5	74.4(+5.8%)	42.7(+6.6%)
SA-YOLO	73.1	39.1	76.4(+3.3%)	45.3(+6.2%)

H. HSR Foreign Object Intrusion Detection and Generation Experiments Under Adverse Weather Conditions

To evaluate the effectiveness of our proposed generation method under challenging conditions, we investigated its ability to improve HSR foreign object detection performance using the YOLOv9 and SA-YOLO models in simulated heavy rain and dense fog scenarios. In this experiment, we introduced factors such as heavy rain, dense fog, and overcast skies, and transformed real-world datasets using computer vision techniques. As shown in the first row of Fig. 11, the original training and validation datasets contain both generated HSR foreign object intrusion images and authentic foreign object intrusion images. We proportionally allocated real foreign object intrusion images collected from operational HSR lines using specialized monitoring equipment throughout the training and validation sets.

The transformed image effects are shown in the second row of Fig. 11. The transformed images exhibit heavy rainfall with extremely poor visibility, achieving realistic adverse weather effects in both training and validation sets. Despite being severely constrained by sample limitations, these authentic images and transformed images provide the most comprehensive reflection of our model's real-world performance under current conditions.

The results presented in Table V clearly demonstrate the benefit of the augmentation strategy. Training with the generated dataset improved detection performance significantly in these adverse weather conditions: the detection performance of YOLOv9 model increased by 5.8% in mAP_{50} and 6.6% in $mAP_{50\text{-}95}$. Similarly, the detection performance of SA-YOLO improved by 3.3% in mAP_{50} and 6.2% in $mAP_{50\text{-}95}$ subsequent to augmentation. We also observed that SA-YOLO consistently achieved higher mAP_{50} and $mAP_{50\text{-}95}$ scores than YOLOv9, both before and after augmentation. This superior performance is attributed to the enhanced feature extraction capabilities of SA-YOLO, proving particularly advantageous when dealing with the blurred edges and reduced contrast characteristics due to poor visibility.

Notably, the SA-YOLO detector trained with our generative method's augmented data yielded substantial improvements of 7.8% in mAP_{50} and 8.8% in $mAP_{50\text{-}95}$ detection accuracy compared to the baseline YOLOv9 operating on the original dataset without augmentation. These findings confirm that our image generation approach possesses strong generalization capabilities, offering considerable potential for practical application in enhancing object detection robustness under extreme environmental conditions.

VI. Conclusion

In this work, we propose an image generation framework for high-speed railway (HSR) foreign object intrusion detection that effectively addresses the challenges of data scarcity and small object detection. In terms of generation, the framework integrates processes of generation, extraction, and integration to produce true-to-reality images of foreign object intrusions in HSR scenarios. Comparative experiments with power grid foreign object intrusion data show that our generative method is more effective in enhancing intrusion detection models, improving mAP_{50} detection accuracy by 7.5% in the optimal configuration. For detection, we use an improved SA-YOLO model by integrating StarNet and Attention-based Dynamic Sampling mechanisms to enhance small object detection capabilities, successfully improving detection performance for lightweight small objects such as plastic bags, kites, and balloons. Ablation experiments indicate that our improved StarNet backbone and A-DyS module can effectively improve detection accuracy. Overall, the SA-YOLO model improves detection performance, and this enhancement is further magnified when using our generated data compared to the baseline model. Comparative experiments with state-of-the-art detection models provide evidence of the generalizability of our generative method in enhancing model performance, with our approach improving detection accuracy by 2.4% compared to the latest YOLOv12 with Real dataset. These results highlight the potential advantages of our method in HSR foreign object intrusion detection across various scenarios. Moreover, experiments under adverse weather conditions demonstrate that our generative method effectively improves detection accuracy, underscoring its potential for application in extreme scenarios.

Despite promising results, this study has key limitations. These limitations include a data collection focus primarily on catenary system intrusions, which inadequately covers track intrusions, and the detection model's underperformance with partially occluded objects. Additionally, the evaluation of generation effectiveness in our experiments remains insufficiently comprehensive, and conducting large-scale manual evaluation experiments requires enormous resources.

refore, further research will continue to expand data collection for diverse intrusion scenarios and further enhance model robustness. Additionally, while our framework allows for some annotation of foreign objects in intermediate image results, automatically annotating the final generated outputs is not yet feasible. Recent advances in multimodal large language models and intelligent agent systems offer promising directions for addressing these limitations. Inspired by collaborative agent frameworks [75] and multimodal foundation models for autonomous driving [76], we will explore intelligent agent-based approaches to expand the applicability of our method in future work. Meanwhile, following the developments in multimodal models for low-level visual perception [77], [78], we will focus on adopting MLLM-based methods for more comprehensive and automated quality assessment of generated images.

References

[1] Z. Li, Z. Rao, L. Ding, B. Ding, J. Fang, and X. Ma, "YOLOv5s-D: A railway catenary dropper state identification and small defect detection model," *Appl. Sci.*, vol. 13, no. 13, p. 7881, Jul. 2023.

[2] C. Liu, S. He, H. Liu, J. Chen, and H. Dong, "WindTrans: Transformer-based wind speed forecasting method for high-speed railway," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 6, pp. 4947–4963, Jun. 2024.

[3] T. Ye, C. Ren, X. Zhang, G. Zhai, and R. Wang, "Application of lightweight railway transit object detector," *IEEE Trans. Ind. Electron.*, vol. 68, no. 10, pp. 10269–10280, Oct. 2021.

[4] T. Ye, Z. Zhao, S. Wang, F. Zhou, and X. Gao, "A stable lightweight and adaptive feature enhanced convolution neural network for efficient railway transit object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17952–17965, Oct. 2022.

[5] S. Zheng, Z. Wu, Y. Xu, and Z. Wei, "Intrusion detection of foreign objects in overhead power system for preventive maintenance in high-speed railway catenary inspection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[6] Z. Liu, Y. Lyu, L. Wang, and Z. Han, "Detection approach based on an improved faster RCNN for brace sleeve screws in high-speed railways," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4395–4403, Jul. 2020.

[7] Z. Chen et al., "Foreign object detection for railway ballastless trackbeds: A semisupervised learning method," *Measurement*, vol. 190, Feb. 2022, Art. no. 110757.

[8] T. Yang, Y. Liu, Y. Huang, J. Liu, and S. Wang, "Symmetry-driven unsupervised abnormal object detection for railway inspection," *IEEE Trans. Ind. Informat.*, vol. 19, no. 12, pp. 11487–11498, Dec. 2023.

[9] D. Li, H. Deng, T. Yu, and L. Zhang, "Multivehicle cooperative localization using a TOA-based simulated annealing extended Kalman filter in urban canyons," *IEEE Internet Things J.*, vol. 12, no. 13, pp. 22832–22846, Jul. 2025.

[10] M. Arikilla and B. Raviteja, "Foreign object debris detection in aerodromes using deep learning approaches," in *Proc. Int. Conf. Inf. Commun. Technol. Intell. Syst.*, Jan. 2023, pp. 587–598.

[11] S. Yang, H. Lu, and J. Li, "Multifeature fusion-based object detection for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1126–1133, Jan. 2023.

[12] J. Huamin, Y. Chang, Z. Zhao, C. Wang, and K. See, "Defect detection and classification of railway track system for in-service MRT in tropical regions using a contactless TBMS and adaptive-DBSCAN," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11130616

[13] J. He, W. Wang, F. Lv, H. Luo, G. Zhang, and Z. Chen, "Multi-scale CNN-transformer hybrid network for rail fastener defect detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 6, pp. 8894–8906, Jun. 2025.

[14] Z. Chen et al., "Foreign object detection method for railway catenary based on a scarce image generation model and lightweight perception architecture," *IEEE Trans. Circuits Syst. Video Technol.*, early access, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10988810

[15] Y. Cao et al., "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," 2023, *arXiv:2303.04226*.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[17] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, Apr. 2024.

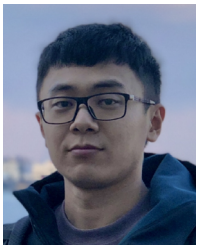
[18] R. Voetman, M. Aghaei, and K. Dijkstra, "The big data myth: Using diffusion models for dataset generation to train deep detection models," 2023, *arXiv:2306.09762*.

[19] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022.

[20] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2024, pp. 1–21.
[21] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.
[22] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7076–7086.
[23] X. Ma, X. Dai, Y. Bai, Y. Wang, and Y. Fu, "Rewrite the stars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5694–5703.
[24] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3 VAE: Self-supervised sequential VAE for representation disentanglement and data generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6537–6546.
[25] I. Higgins et al., "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–22. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl
[26] H. Shao et al., "ControlVAE: Controllable variational autoencoder," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2020, pp. 8655–8664.
[27] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[28] I. Daunhawer, T. M. Sutter, K. Chin-Cheong, E. Palumbo, and J. E. Vogt, "On the limitations of multimodal VAEs," 2021, *arXiv:2110.04121*.
[29] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[30] H. Zhang et al., "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.
[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.
[32] Y. Li et al., "StoryGAN: A sequential conditional GAN for story visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6322–6331.
[33] E. Richardson et al., "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.
[34] M. Durgadevi, "Generative adversarial network (GAN): A general review on different variants of GAN and applications," in *Proc. 6th Int. Conf. Commun. Electron. Syst. (ICCES)*, 2021, pp. 1–8.
[35] D. Saxena and J. Cao, "Generative adversarial networks (GANs): Challenges, solutions, and future directions," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–42, Apr. 2022.
[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.
[37] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, 2015, pp. 234–241.
[39] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[40] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22503–22513.
[41] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "LayoutDiffusion: Controllable diffusion model for layout-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22490–22499.
[42] Y. Li et al., "GLIGEN: Open-set grounded text-to-image generation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jan. 2023, pp. 22511–22521.
[43] Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, and J. Xu, "Shifted diffusion for text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10157–10166.
[44] B. Kawar et al., "Imagic: Text-based real image editing with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6007–6017.
[45] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.
[46] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.
[47] M. Peng et al., "Diffusion models for intelligent transportation systems: A survey," 2024, *arXiv:2409.15816*.
[48] J. Young Choi, J. R. Park, I. Park, J. Cho, A. No, and E. K. Ryu, "Simple drop-in LoRA conditioning on attention layers will improve your diffusion model," 2024, *arXiv:2405.03958*.
[49] T. Wang, Z. Zhang, and K.-L. Tsui, "A deep generative approach for rail foreign object detections via semisupervised learning," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 459–468, Jan. 2023.
[50] G. Jocher, "Ultralytics YOLOv5," Zenodo, Tech. Rep., Nov. 2022, doi: 10.5281/zenodo.7347926. [Online]. Available: https://zenodo.org/records/7347926
[51] C. Liu et al., "YOLO-CSM-based component defect and foreign object detection in overhead transmission lines," *Electronics*, vol. 13, no. 1, p. 123, Dec. 2023.
[52] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.
[53] L. Liu et al., "YOLO-3DMM for simultaneous multiple object detection and tracking in traffic scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9467–9481, Aug. 2024.
[54] S. Ning, F. Ding, and B. Chen, "Research on the method of foreign object detection for railway tracks based on deep learning," *Sensors*, vol. 24, no. 14, p. 4483, Jul. 2024.
[55] Q. Ding et al., "CF-YOLO: Cross fusion YOLO for object detection in adverse weather with a high-quality real snow dataset," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 10749–10759, Oct. 2023.
[56] L. Falaschetti, L. Manoni, L. Palma, P. Pierleoni, and C. Turchetti, "Embedded real-time vehicle and pedestrian detection using a compressed tiny YOLO v3 architecture," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 19399–19414, Dec. 2024.
[57] J. Zhan, L. Zhang, and J. Qiao, "Boundary consensus of networked hyperbolic systems of conservation laws," *IEEE Trans. Autom. Control*, vol. 70, no. 8, pp. 4989–5004, Aug. 2025.
[58] R. Shi, T. Li, Y. Yamaguchi, and L. Zhang, "Traffic scene-informed attribution of autonomous driving decisions," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 7, pp. 9175–9186, Jul. 2025.
[59] T. Li, R. Shi, Q. Zhu, L. Zhang, and T. Kanai, "Spectrum-enhanced graph attention network for garment mesh deformation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 8, pp. 7153–7170, Aug. 2025.
[60] R. Shi, T. Li, Y. Yamaguchi, and L. Zhang, "Attribution explanations for decision-making in deep lane-change models," *Transp. Res. C, Emerg. Technol.*, vol. 180, Nov. 2025, Art. no. 105361. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X25003651
[61] W. Liu, H. Lu, H. Fu, and Z. Cao, "Learning to upsample by learning to sample," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4512–4522.
[62] Z. Qi, D. Ma, J. Xu, A. Xiang, and H. Qu, "Improved YOLOv5 based on the attention mechanism and FasterNet for foreign object detection on railway and airway tracks," in *Proc. Asian Conf. Commun. Netw. (ASIANComNet)*, Oct. 2024, pp. 1–6.
[63] H. Li, J. Leung, and Z. Shen, "Towards goal-oriented prompt engineering for large language models: A survey," 2024, *arXiv:2401.14043*.
[64] J. Bai et al., "Qwen technical report," 2023, *arXiv:2309.16609*.
[65] R. Varghese and M. Sambath, "YOLOv8: A novel object detection algorithm with enhanced performance and robustness," in *Proc. Int. Conf. Adv. Data Eng. Intell. Comput. Syst. (ADICS)*, Apr. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10533619
[66] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, and J. Han, "YOLOv10: Real-time end-to-end object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 107984–108011.
[67] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.
[68] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," 2025, *arXiv:2502.12524*.
[69] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.
[70] D. Qin et al., "MobileNetV4: Universal models for the mobile ecosystem," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2024, pp. 78–96.
[71] S. A. CompVis, "Stable diffusion V1.5," 2022. [Online]. Available: https://github.com/Kameronski/stable-diffusion-1.5
[72] D. Podell et al., "SDXL: Improving latent diffusion models for high-resolution image synthesis," 2023, *arXiv:2307.01952*.

[73] Black Forest Labs. (2024). *FLUX*. [Online]. Available: https://github.com/black-forest-labs/flux
[74] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[75] Y. Wei et al., "Editable scene simulation for autonomous driving via collaborative LLM-agents," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15077–15087.
[76] Z. Xu et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robot. Autom. Lett.*, vol. 9, no. 10, pp. 8186–8193, Oct. 2024.
[77] H. Wu et al., "Q-bench: A benchmark for general-purpose foundation models on low-level vision," 2023, *arXiv:2309.14181*.
[78] H. Wu et al., "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25490–25500.

Quan Hao received the bachelor's degree from the College of Software, Beijing University of Technology (BJUT), Beijing, China, in 2022. He is currently pursuing the Ph.D. degree with the School of Information Science and Technology, BJUT. His current research interests include intelligent transportation and generative artificial intelligence.



Rui Shi received the Ph.D. degree in graphic and computer sciences from The University of Tokyo, Tokyo, Japan, in 2022. He worked as a Visiting Researcher with the Department of General Systems Studies, The University of Tokyo. He is currently a Lecturer with the School of Information Science and Technology, Beijing University of Technology, Beijing, China. His current research interests include autonomous driving, neural networks, and explainable artificial intelligence.

Jiaze Li (Student Member, IEEE) is currently pursuing the dual bachelor's degree in electronic information engineering with University College Dublin and Beijing University of Technology. His research interests include generative artificial intelligence and computer vision.

Liguo Zhang (Senior Member, IEEE) received the Ph.D. degree in control theory and applications from Beijing University of Technology (BJUT), Beijing, China, in 2006. Since 2014, he has been a Full Professor with the School of Electronic Information and Control Engineering, BJUT. He is currently the Deputy Director of the School of Information Science and Technology, BJUT. His research interests include hybrid systems, intelligent systems, and control of distributed parameter systems. He is an Associate Editor of *IMA Journal of Mathematical Control and Information* and the Guest Editor of *International Journal of Distributed Sensor Networks*.