Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085491

ARTICLE

Motor Temperature Detection System Based on Multi-Scale Feature Screening and Thermal Image Reconstruction

Juhui Zhang, Zongyi Xing* and Chenxiao Cai

School of Automation, Nanjing University of Science and Technology, Nanjing, China

*Corresponding Author: Zongyi Xing. Email: xingzongyi@njust.edu.cn

Received: 12 May 2026; Accepted: 24 August 2026; Published: 22 October 2026

ABSTRACT: As urban rail transit systems continue to expand, greater technical demands are placed on the reliability assessment and temperature monitoring of key bogie components under dynamic operating conditions. To address the complexity of state detection caused by motion blur in high-speed targets, this study proposes a motor temperature detection system based on dynamic infrared video analysis. The system evaluates motor operating conditions by rapidly screening motor images and applying super-resolution reconstruction techniques. First, the system's configuration and operating principles are described. Next, an infrared image screening algorithm is developed based on temperature feature extraction and matching. The algorithm incorporates sliding window grouping and template matching to quickly identify motor targets in infrared video. Finally, the Perception ResNet Attention Generative Adversarial Network (PRAGAN) super-resolution network is employed to improve the spatial resolution of thermal images. It utilizes a multi-scale residual encoder, Visual Geometry Group (VGG)-based perceptual loss, and an uncertainty-based adaptive loss-weighting strategy whose update rule is given explicitly. The system was deployed and validated on Nanjing Metro Line 1. Experimental results demonstrate that the system achieves a temperature measurement accuracy of 2°C and operates reliably at a train speed of 55 km/h. It maintains high measurement accuracy while minimizing equipment complexity, indicating strong engineering applicability.

KEYWORDS: Train motors; condition monitoring; infrared thermal imaging; super-resolution reconstruction

1 Introduction

As a core subsystem for load-bearing, movement, and traction, railway bogies face major reliability challenges under coupled mechanical and electromagnetic stresses. Traction motors operate under high fatigue loads and harsh conditions, including dust and humidity, which can cause bearing wear, lubricant degradation, and electromagnetic-system damage [1,2]. Without timely monitoring, these faults may lead to thermal runaway, traction loss, derailment, or collisions. Therefore, we developed an infrared thermal-imaging and intelligent-detection system for real-time temperature monitoring and fault localization in critical bogie motor components.

Current motor temperature monitoring methods can be classified as manual, contact, and non-contact approaches. Manual inspection and handheld thermometers are inefficient and cannot provide continuous measurements [3]. Contact systems use sensors mounted on bogie components to acquire real-time temperature data [4] and have been deployed on the French railway, Japanese Shinkansen, Shanghai Metro, Guangzhou Metro, and CRH trains [5]. Although these systems provide accurate alarms, extensive

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sensor and controller installation increases maintenance complexity, leaves blind spots, and raises operating costs, limiting comprehensive monitoring under vibration and other complex conditions.

Non-contact detection methods are categorized into point-based and surface-based temperature measurements. Point-based temperature measurement typically uses sensors mounted trackside. In the "5T" monitoring system proposed by China's railway authority, the Temperature Hotbox Detection System (THDS) collects temperature data via these sensors [6]. Lei et al. [7] constructed a trackside temperature monitoring platform to continuously collect real-time temperature data of train axles. Wu et al. [8] adopted infrared imaging technology to realize abnormal state detection of axle boxes and traction motors. Du et al. [9] proposed a bogie temperature detection scheme combining Transformer positioning and affine matching, which lowered the data loss rate to 1.8% and satisfied practical application demands. Our previous study [10] conducted state monitoring of the bogie components using infrared images. However, it assumed sufficient spatial resolution, did not consider the blurring caused by motion, and failed to address the trade-off between resolution and frame rate under the operating speed of the railway line. Kim [11] applied infrared thermography to railway materials and components. More broadly, infrared thermography has been established as a non-contact technique for condition monitoring, temperature measurement, and nondestructive testing [12,13]. The core problem addressed is the recovery of radiometrically faithful, high-resolution motor temperature fields from low-resolution, motion-blurred infrared frames captured at line speed, where high-resolution sensors cannot sustain the frame rate required by a fast-moving train. This entails three methodological challenges: (i) fast retrieval of valid motor frames from highly redundant infrared video; (ii) super-resolution that enhances detail without hallucinating textures that would distort the temperature reading; and (iii) stable generative adversarial network (GAN) training under competing pixel, perceptual, and adversarial objectives.

To address these challenges, this study proposes a temperature monitoring system that uses real-time underbody infrared video to evaluate motor temperature through rapid image screening and super-resolution reconstruction techniques. The main contributions of this study are summarized as follows:

(1) An integrated motor temperature monitoring system that combines multi-scale feature screening and thermal-image reconstruction. The system uses two non-contact thermal imaging cameras to acquire data and evaluate motor conditions.

(2) A fast screening method for motor images in video sequences is proposed, which combines image features and feature matching to efficiently identify key frames.

(3) An infrared thermal image super-resolution technique is introduced. It enhances feature representation using a multi-scale residual generator, improves high-frequency detail recovery through Visual Geometry Group Network (VGGNet) based perceptual loss, and employs adaptive loss weighting to maintain training stability.

The remainder of the study is organized as follows. Section 2 describes the system configuration and its operating principles. Section 3 introduces an infrared image retrieval method based on temperature feature screening and feature matching. Section 4 presents a super-resolution reconstruction algorithm for infrared images. Section 5 presents experimental results and performance analyses conducted in both laboratory and field environments. Finally, Section 6 summarizes the key findings and outlines future research directions.

2 System Overview

2.1 System Configuration

The system consists of four core units: data acquisition, field control, data transmission, and data processing and storage. The data acquisition module integrates a temperature acquisition unit, wheel sensors,

and a train number recognition unit. The temperature acquisition unit comprises two thermal imaging cameras positioned at fixed angles to capture images of the train motor, as shown in Fig. 1. The thermal imaging cameras are triggered by wheel sensor 1. The train number recognition device is installed trackside to identify train identification numbers. The system includes two wheel sensors, spaced 30 m apart according to field deployment requirements.

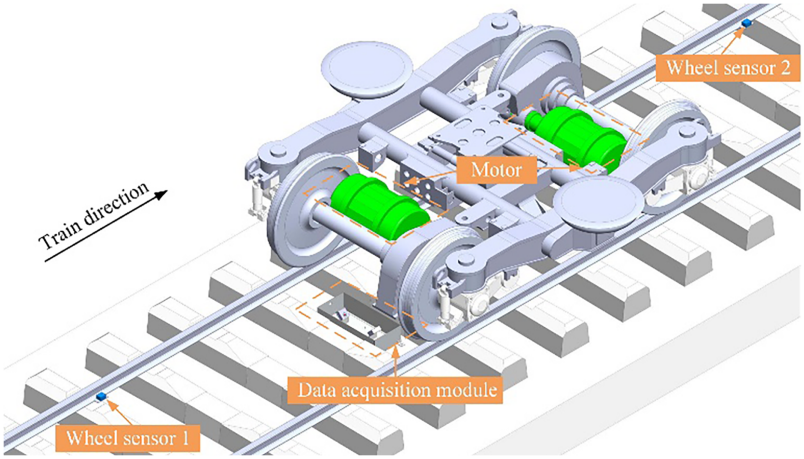


Figure 1: Layout of the field hardware.

2.2 Workflow

When the train triggers wheel sensor 1, the programmable logic controller (PLC) activates data acquisition. After the train passes wheel sensor 2 and the wheel count reaches 24, the PLC stops data acquisition and vehicle number recognition. The collected data are transmitted to the server, where motor localization and super-resolution reconstruction are performed to extract, store, and visualize motor temperature data. The system workflow is shown in Fig. 2a.

3 Infrared Image Retrieval Method Using Temperature-Based Feature Screening and Matching

The input to the localization module is a continuous infrared sequence containing repeated views of the same motor. Because train motion, transient occlusion, and processing latency may cause some frames to be blurred or incomplete, the sequence must first be converted into a set of reliable motor observations. The proposed retrieval procedure therefore performs thermal-feature screening followed by temporal grouping and template-based verification. The complete procedure is illustrated in Fig. 2b.

3.1 Coarse Positioning

3.1.1 Feature Calculation

Each video frame is analyzed independently after the infrared stream has been separated into individual images. For a candidate region, the standard deviation (STD) of pixel intensities is used to describe local thermal variation. Regions with larger STD values generally contain stronger temperature changes and are more likely to include motor components or other thermally active structures. A pixel-density filter is subsequently applied to reduce the number of regions passed to the following stages.

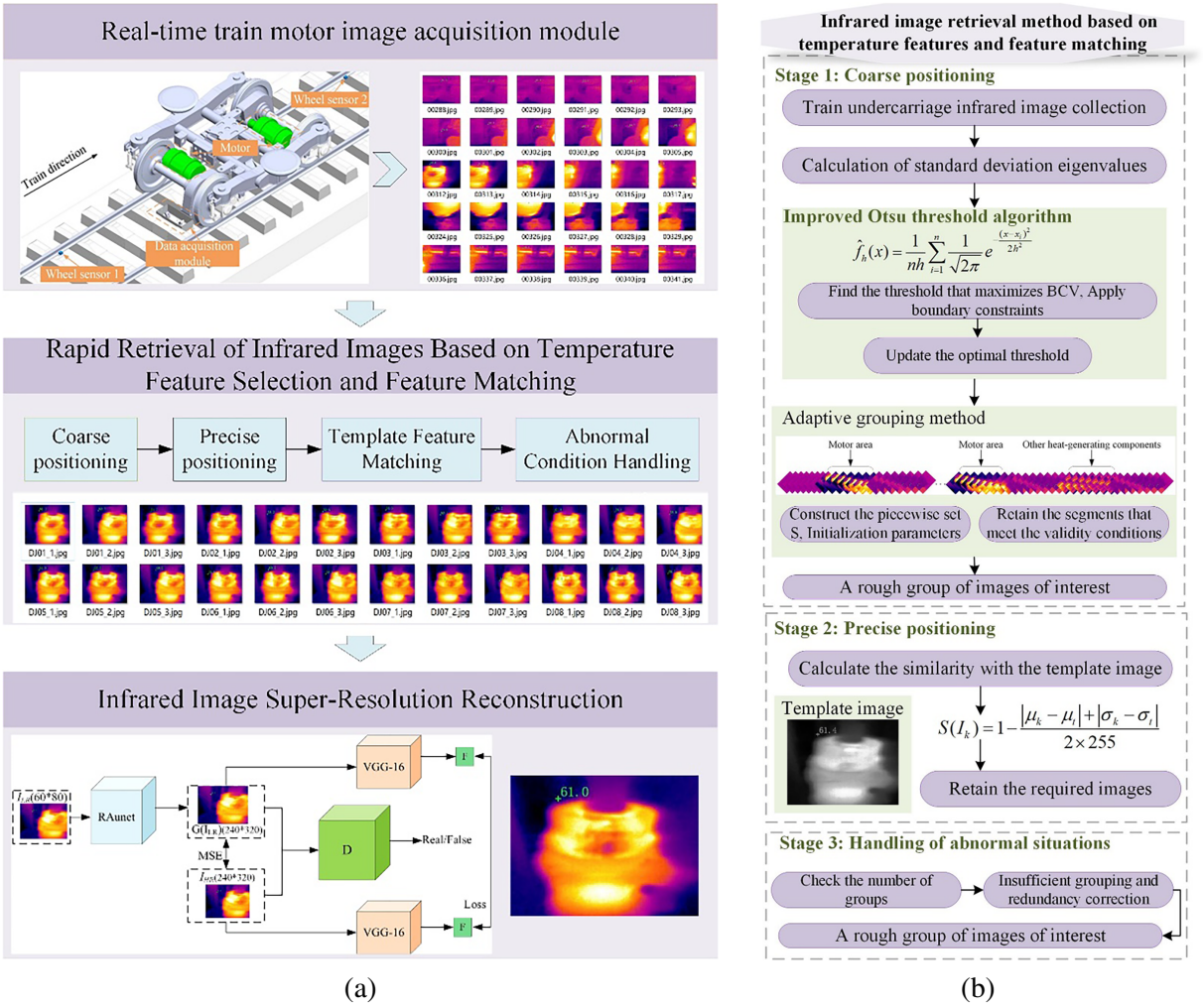


Figure 2: (**a**) System workflow; (**b**) Flowchart of the rapid screening algorithm.

The observed STD distribution changes with both the ambient temperature and the thermal state of the underbody components. Preliminary measurements showed that low-temperature structures and motor-related regions occupy different parts of this distribution. Accordingly, a lower screening limit of $T_{min} = 35$ was introduced for low-ambient-temperature conditions. This limit provides a margin against small fluctuations near the segmentation boundary and prevents unstable candidate selection.

3.1.2 Improvement of Otsu Threshold Algorithm

A single manually selected STD threshold is not reliable when the thermal distribution changes across operating conditions. We therefore estimate the threshold from the STD samples observed in the current image group. Otsu's criterion is retained to determine the point that provides the greatest separation between the two classes. Before applying this criterion, however, the empirical distribution is smoothed with kernel density estimation (KDE) [14]. This KDE-assisted formulation avoids excessive dependence on histogram-bin boundaries and is better suited to the multimodal distributions produced by underbody thermal images.

Let the measured STD values form a sample set. The first step is to obtain a continuous estimate of their probability density.

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-x_i)^2}{2h^2}} \tag{1}$$

where n denotes the number of STD samples and h represents the kernel bandwidth, which is estimated using the following robust rule:

$$h = 0.9 \cdot \min\left(\hat{\sigma}, \frac{I}{1.34}\right) \cdot n^{-1/5} \tag{2}$$

where I is the interquartile range.

The estimated density is evaluated over m points between x_{min} and $x_{max.}$ For a candidate threshold T, the samples are divided into a lower-intensity class and a higher-intensity class. The probability mass and mean value of each class are then calculated according to Eq. (3).

$$\omega_0(T) = \sum_{k=1}^{T} p_k, \omega_1(T) = 1 - \omega_0(T) \tag{3}$$

where $p_k = \frac{\hat{f}_h(x_k^*)\triangle x}{\sum_{i=1}^{m}\hat{f}_h(x_i^*)\triangle x}$, $\triangle x = \frac{x_{\max}-x_{\min}}{m-1}$. The category means for both categories are:

$$\mu_0(T) = \frac{1}{\omega_0(T)}\sum_{k=1}^{T} x_k^* \cdot p_k, \mu_1(T) = \frac{1}{\omega_1(T)}\sum_{k=T+1}^{m} x_k^* \cdot p_k \tag{4}$$

The separation produced by a candidate threshold is measured by the between-class variance. A larger value indicates that the two groups are more clearly separated:

$$\sigma_b^2(T) = \omega_0(T) \cdot \omega_1(T) \cdot [\mu_0(T) - \mu_1(T)]^2 \tag{5}$$

Determine the optimal threshold:

$$T_{\text{otsu}} = \arg\max_T \sigma_b^2(T) \tag{6}$$

The threshold is constrained to remain within the valid STD interval. The final value is obtained after evaluating all admissible candidates and is illustrated in Fig. 3.

3.1.3 Adaptive Grouping Methods

Frame-wise screening alone cannot determine whether neighboring positive detections belong to the same motor passage. Directly joining adjacent positive frames may produce over-segmented sequences, whereas requiring uninterrupted detections may remove valid sequences containing a few abnormal frames. To address both situations, the retained frame indices are connected using a tolerance-based sliding-window procedure.

The indexed set of screened high pixel images is $S = \{s_1, s_2, \cdots, s_n\}$, construct the set of segments $G = \{[s_k, e_k]\}$, where each segment $[s_k, e_k]$ represents a continuous frame interval. First, initialize the parameters, the current segment start point $s_c = t_1$, the current segment termination point $e_c = t_1$. After

setting the initial values, iterative processing $\forall t_i \in S\ (i \geq 2)$ is performed if the neighborhood condition is satisfied:

$$\triangle_i \leq \delta + 1 \tag{7}$$

where, $\triangle_i = s_i - s_{i-1}$, δ is set to 3, allowing up to three non-anomalous frames within a segment. The current segment $e_c \leftarrow t_i$ is extended if the condition is met; otherwise, the segment is submitted and the process is reset:

$$\begin{cases} G \leftarrow G \cup \{[s_c, e_c]\} \\ s_c \leftarrow s_i \\ e_c \leftarrow s_i \end{cases} \tag{8}$$

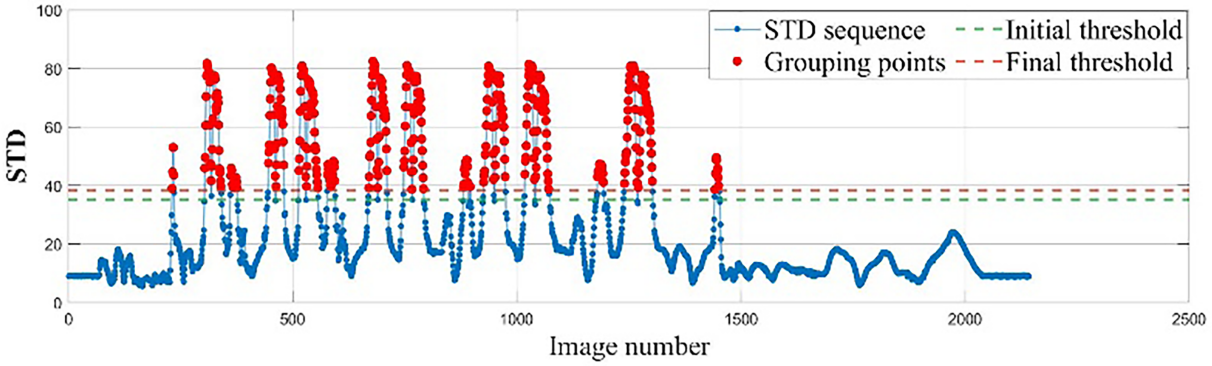


Figure 3: Determined final threshold.

The minimum effective length threshold is defined as $L_{\min} = 10$, which is used to filter out short-lived pseudo-activated segments. For each candidate segment S_k, compute its duration:

$$Gv = \{S_k \in G \,|\, L_k \geq L_{\min}\} \tag{9}$$

Retain the subsegments that meet the predefined validity criteria:

$$L_k = e_k - s_k + 1 \tag{10}$$

By integrating STD feature analysis with an enhanced sliding window grouping algorithm, effective motor image segments can be efficiently identified and constructed.

3.2 Feature Matching Algorithm

Thermal sources other than the traction motor may also pass the coarse screening stage, resulting in multiple candidate intervals for a single passage. A second verification stage is therefore required to distinguish the motor view from unrelated hot structures. We use a reference image with a stable motor position, clear thermal contrast, and minimal occlusion. Each candidate is converted into a comparable representation and evaluated against this reference using brightness-distribution and spatial-pattern information. Candidates whose similarity score falls below the acceptance threshold are rejected; the remaining candidates are forwarded to reconstruction and temperature analysis.

A bank of templates could improve adaptability when the target appearance changes substantially, but it would increase both matching time and template-maintenance requirements. Because the deployment configuration keeps the motor position relatively stable, one representative template is used in this study. Before similarity calculation, both the template and candidate images are binarized to reduce the influence of absolute temperature offsets and sensor noise. The selected template is shown in Fig. 4.

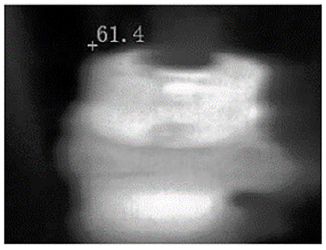


Figure 4: Template image.

4 Image Super-Resolution Reconstruction

High-resolution sensors lower the imaging frame rate due to extended pixel scanning time, making it challenging to meet real-time requirements during high-speed train operation. In such scenarios, overemphasis on image quality at the expense of frame rate can lead to motion blur and detail loss, negatively impacting subsequent analysis. Therefore, a balance between image clarity and acquisition speed is essential under real-world conditions. Single-frame super-resolution (SR) from low-resolution (LR) images offers a practical solution. SR significantly enhances infrared image detail and clarity while reducing hardware costs, and has broad applications in medical imaging, satellite remote sensing, and security surveillance.

Traditional infrared super-resolution methods based on the frequency or spatial domain have limited capability in recovering fine details [15,16]. Deep-learning methods have progressed along three main lines. Convolutional neural network (CNN)-based methods, including multi-scale residual networks [17], SRCNN [18], RCAN [19], and Efficient Sub-Pixel Convolutional Neural Network (ESPCN) [20], efficiently learn nonlinear mappings from low-resolution (LR) to high-resolution (HR) images but may over-smooth high-frequency details. Transformer-based methods, including SwinIR and its variants [21,22], improve long-range dependency modeling but impose higher computational costs. Generative adversarial network (GAN)-based methods can reconstruct sharper textures, although they may introduce training instability and texture hallucination; region-aware adversarial learning has therefore been investigated for infrared super-resolution [23]. Few existing methods specifically address low-texture, noise-dominated trackside thermal images or verify whether super-resolution improves temperature-related spatial information rather than only peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). This motivates the proposed Perception ResNet Attention Generative Adversarial Network (PRAGAN) and its evaluation.

4.1 Generative Adversarial Networks

Goodfellow et al. introduced generative adversarial networks (GANs) as a minimax game between a generator and a discriminator [24]. For super-resolution, the generator maps an LR input to an SR image, whereas the discriminator distinguishes generated outputs from real HR images and supplies adversarial feedback. This competition can sharpen reconstruction but may also cause texture distortion and unstable training, motivating the loss design and stabilization strategy used in PRAGAN.

However, GAN-based reconstruction results under real-world train operating conditions remain suboptimal. Several challenges currently hinder the effectiveness of super-resolution reconstruction:

(1) Texture distortion: Adversarial training often results in unnatural textures in localized image regions.

(2) Loss of high-frequency details: Pixel-level loss functions often produce over-smoothed outputs.

(3) Training instability: Maintaining dynamic equilibrium between the generator and discriminator remains challenging.

4.2 Improved GAN

Therefore, we propose a novel model PRAGAN, which incorporates the following key improvements:

(1) A residual generator with channel attention is designed to enhance feature representation;

(2) VGG-based hierarchical perceptual loss is introduced to improve the reconstruction of high-frequency details;

(3) An adaptive loss weighting strategy is proposed to stabilize the training process.

Unlike generic SRGAN/ESRGAN pipelines that apply a fixed weighted sum of losses, PRAGAN couples an attention-gated multi-scale residual U-shaped Convolutional Neural Network (U-Net) generator with homoscedastic-uncertainty-based adaptive loss-weighting strategy. This combination is designed specifically for the low-texture, noise-dominated regime of trackside infrared motor imagery, where conventional perceptual SR tends to hallucinate textures that corrupt the radiometric (temperature) reading. The individual primitives (residual blocks, attention, perceptual loss) are standard.

4.2.1 Multi-Scale Feature Fusion Module

To recover high-frequency details and enhance feature representation, we propose an improved multi-scale encoder–decoder structure. This design integrates residual learning, skip connections, and attention mechanisms into a U-Net architecture, forming a perception-driven subnetwork. Fig. 5 shows the module architecture, which integrates layered feature extraction, multi-scale fusion, and interlayer attention to enable deep feature representation and high-fidelity image reconstruction.

The encoder includes a two-stage residual unit and max pooling, reducing spatial resolution while increasing channel depth to extract deep features. Residual blocks mitigate the vanishing gradient problem, enhance training stability, and preserve edge information. Residual blocks capture signals better than standard convolution, enhancing the network's representational capacity. The decoder upsamples with PixelShuffle and merges the result with high-resolution features from the corresponding encoder layer. PixelShuffle avoids checkerboard artifacts and preserves image texture continuity better than deconvolution. U-Net's symmetric skip connections introduce low-level features into the decoder, recovering edges and details while improving texture reconstruction.

An Attention Module after each upsampling layer weighs feature importance, highlights key texture responses, and suppresses noise, enabling more accurate texture reconstruction. The Attention Module facilitates feature reorganization and scale selection, acting as a "gating unit" in multi-scale fusion. It dynamically focuses on regions of interest, enhancing perceptual consistency and detail restoration.

4.2.2 VGG-Based Hierarchical Perception

VGG16 uses stacked 3 × 3 convolutions to learn progressively higher-level image representations [25]. In PRAGAN, the pretrained network is used as a fixed feature extractor rather than a classifier. Selected convolutional features provide perceptual supervision by measuring the discrepancy between reconstructed and HR images in feature space. The VGG16 structure is illustrated in Fig. 6.

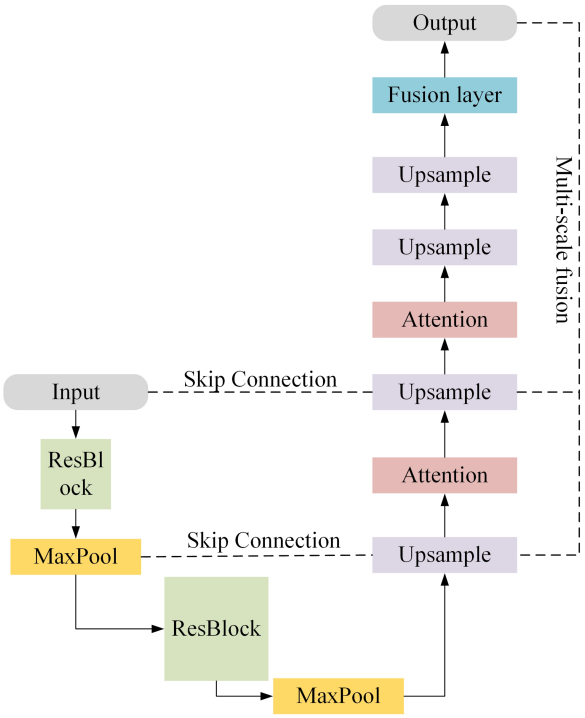


Figure 5: Multi-scale feature fusion module.

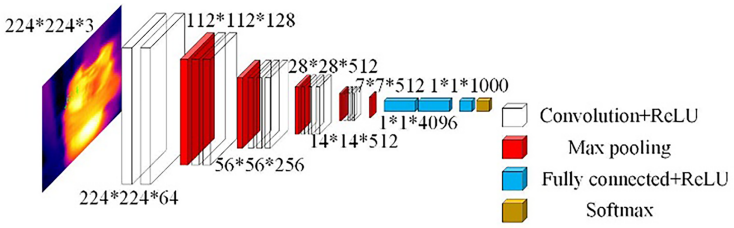


Figure 6: VGG-16 network structure.

The multiscale generator captures rich features across multiple scales, while the hierarchical perceptual loss imposes distinct constraints on high-frequency details and low-frequency structures at various levels of the feature space, thereby jointly improving reconstruction quality. This multilevel supervision enables the network to simultaneously optimize high-frequency textures and low-frequency structures across scales.

4.2.3 Loss Function

Generator training combines adversarial, content, and perceptual objectives. The fixed-weight formulation provides a non-adaptive baseline, but manually selected coefficients cannot respond to differences in loss scale. PRAGAN therefore uses homoscedastic-uncertainty weighting to learn the effective contribution of each term during optimization. The individual objectives are defined below.

Adversarial Loss: emphasizes the global realism of the image to make the generated results closer to the real distribution.

$$\hat{I}_{HR} = G\left(I_{LR}\right) \tag{11}$$

$$L_{adv} = E_{I_{LR}}\left[D\left(G\left(I_{LR}\right)-1\right)^2\right] \tag{12}$$

$$L_D = \frac{1}{2}E_{I_{LR}}\left[\left(D\left(I_{HR}\right)-1\right)^2\right] + \frac{1}{2}E_{I_{LR}}\left[\left(D(G\left(I_{LR}\right)^2\right] \tag{13}$$

where, I_{LR} denotes the input low-resolution image. $G(\cdot)$ represents the generator that maps, I_{LR} to the super-resolved image $G\left(I_{LR}\right)$; $D(\cdot)$ represents the PatchGAN discriminator that receives either a real high-resolution image IHR or a generated image $G\left(I_{LR}\right)$ and outputs a spatial realness map, where values close to 1 indicate real HR patches and values close to 0 indicate generated SR patches.

Pixel-level Content Loss: enforces global structural consistency between the generated and target images at the pixel level, thereby preventing pattern collapse.

$$L_C = \left\|G\left(I_{LR}\right) - I_{HR}\right)\right\|_1 \tag{14}$$

where I_{HR} denotes the real high-resolution image, $G\left(I_{LR}\right)$ represents the generated high-resolution image, and $\|\cdot\|_1$ refers to the pixel-level L_1 loss.

Hierarchical Perceptual Loss: constrains high-frequency detail and semantic consistency across multiple feature levels to enhance visual richness.

$$L_P = \sum_{l=1}^{L}\beta_l \left\|\phi_l\left(G(I_{LR}\right) - \phi_l\left(I_{HR}\right)\right\|_2^2 \tag{15}$$

where L denotes the number of selected layers in the VGG16 network; $\phi_l(\cdot)$ represents the feature extraction function at the l-th layer of the pre-trained VGG16 model; and β_l is the weighting coefficient for the features of the l-th layer, which controls the constraint strength for each layer.

The total generator loss function is:

$$L = \lambda_g L_G + \lambda_c L_C + \lambda_p L_P \tag{16}$$

where λ_g, λ_c, and λ_p denote the fixed weighting coefficients for the adversarial, content, and perceptual losses, respectively, and are set to 1, 100, and 10 in the non-adaptive baseline. In the full PRAGAN model, these fixed coefficients are replaced by the learnable homoscedastic-uncertainty weights defined below.

We adopt a homoscedastic-uncertainty-based adaptive weighting strategy for the three generator objectives, i.e., adversarial, content, and perceptual losses. For each loss term L_k, where $k \in \{$adv, content, perceptual$\}$, we introduce a learnable uncertainty parameter σ_k. The total generator loss is formulated as:

$$L_G = \sum_k \left(\frac{1}{2\sigma_k^2}L_k + \log\sigma_k\right) \tag{17}$$

In implementation, we optimize $s_k = \log {\sigma_k}^2$ for numerical stability, leading to the equivalent form

$$L_G = \sum_k \left(\frac{1}{2}\exp\left(-s_k\right)L_k + \frac{1}{2}s_k\right) \tag{18}$$

Therefore, the effective weight of each loss term is automatically determined as:

$$\lambda_{\mathrm{k}} = \frac{1}{2}\exp\left(-s_k\right) = \frac{1}{2\sigma_k^2} \tag{19}$$

where s_k is the learnable log-variance parameter associated with the k-th generator loss. During training, s_k is updated together with the generator parameters by minimizing L_G, whereas the discriminator parameters are updated separately by minimizing L_D. No explicit epoch-dependent loss schedule is used. When a loss term becomes relatively large or difficult to optimize, its corresponding uncertainty increases, thereby reducing its effective weight and preventing it from dominating the generator optimization. Meanwhile, the regularization term 1/2 s_k prevents the trivial solution $\sigma_k \to \infty$. The initial log-variance parameters were set to $s_{adv} = 0$, $s_{content} = -2$, and $s_{perceptual} = -1$.

The three losses provide complementary supervision: adversarial loss promotes perceptual realism, content loss preserves pixel-level fidelity, and perceptual loss constrains VGG feature-space consistency. In full PRAGAN, their effective weights are learned through the uncertainty-based formulation rather than an epoch-dependent schedule. Fig. 7 summarizes the complete training architecture.

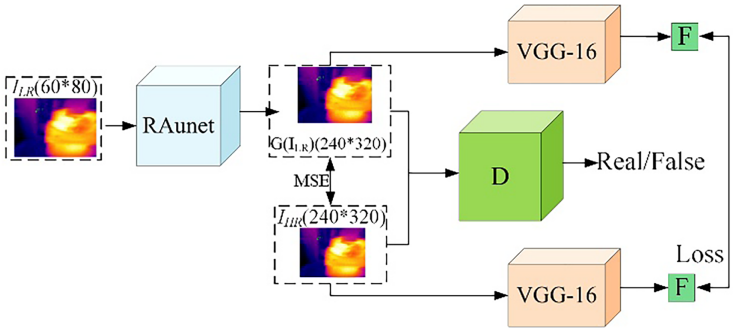


Figure 7: Model diagram of the improved GAN network.

PRAGAN extracts shallow features using a 9 × 9 convolution, processes them with sixteen residual blocks, and applies a global residual connection and channel attention before two-stage PixelShuffle upsampling. A final 9 × 9 convolution with Tanh activation produces the SR output. The PatchGAN discriminator and pretrained VGG16 extractor then provide adversarial, pixel-level content, and feature-space perceptual losses, whose adaptive combination is used to update the generator.

5 Experiments

5.1 Experimental Environment

The data used in this study were obtained from photographs of equipment implemented at Nanjing Metro Line 1, China. The site installation is shown in Fig. 8. Additionally, two laboratory experiments were conducted, and the detailed hardware configuration and experimental setup are provided in Table 1. The dataset is self-collected (no public benchmark was used). The dataset used in the reported experiments contains 401 infrared thermal images with an original resolution of 320 × 240 pixels (W × H). These images were used as HR targets at a resolution of 240 × 320 pixels in H × W notation. To generate the corresponding LR inputs, Gaussian blur with a radius of 1.2 was first applied, followed by bicubic downsampling to 60 × 80 pixels and the addition of Gaussian noise with a standard deviation of 0.015. This procedure produced paired LR–HR images for 4× super-resolution. The dataset was divided at the image level into training, validation, and test subsets at a ratio of approximately 70:15:15, comprising 280, 60, and 61 images, respectively.

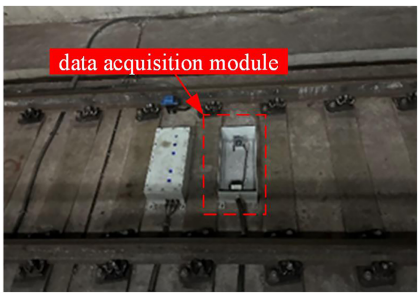


Figure 8: Installation of equipment on site.

Table 1: Hardware configuration and software environment.

Type of Experiment	Hardware	Configure	Software	es
Rapid Screening Experiment		Windows 10	Matlab	R2020a
Super-resolution reconstruction experiments	GPU	NVIDIA GeForce RTX 2080 Ti	VS Code	11 G Video memory

5.2 Screening Experiments

5.2.1 Evaluation Indicators

Recall and F1 score are used to quantify target-screening accuracy, while processing speed evaluates the algorithm's real-time performance.

Recall: a measure of the proportion of all actual motor areas successfully detected by the system, defined as:

$$R = \frac{TP}{TP + FN} \tag{20}$$

where TP denotes the number of correctly detected target regions and FN denotes the number of true target regions that are not detected.

$F1$ score: It is used to synthesize the system's balance between positioning accuracy (Precision) and Recall, and is calculated by the formula:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{21}$$

where P denotes the proportion of detected targets that are real targets.

Processing Speed: This metric evaluates real-time performance by measuring the number of frames processed per second, expressed in frames per second (FPS).

5.2.2 Experimental Analysis

Four screening strategies were evaluated to examine the behavior of the proposed method under different ambient conditions. Method A represents the complete pipeline developed in this study, whereas Methods B–D provide alternative combinations of thresholding, frame grouping, template matching, and structural comparison. Their main processing characteristics are summarized in Table 2.

Table 2: Configurations of the evaluated screening methods.

Method	Method Description
Method A	Improved Standard Deviation + Kernel Density Estimation + Dynamic Threshold + Sliding Window Clustering+ Template Matching + Anomaly Correction
Method B	Fixed Thresholding + Continuous Frame Screening + Template Matching
Method C	Image Grouping by K-means Clustering Method + Template Verification
Method D	Structural Similarity Screening by Region of Interest + Template Structural Comparison

The screening performance was evaluated using two groups of train images acquired at ambient temperatures of 30.7°C and 22.3°C. For each condition, recall, F1 score, and processing speed were calculated using the same evaluation procedure. The quantitative results are reported in Table 3.

Table 3: Experimental comparison of methods.

Environmental Temperature	Method	Recall	F1-Score	Processing Speed (FPS)
30.7	A	0.267	0.421	763.52
	B	0.083	0.153	836.44
	C	0.012	0.023	829.97
	D	0.002	0.004	243.92
22.3	A	0.116	0.208	761.34
	B	0.068	0.128	795.20
	C	0.014	0.027	818.73
	D	0.001	0.002	314.56

As shown in Table 3, Method A achieves the highest recall and F1 score at both ambient temperatures. At 30.7°C, its recall and F1 score are 0.267 and 0.421, respectively; at 22.3°C, the corresponding values are 0.116 and 0.208. These results indicate that the combination of adaptive thresholding, temporal grouping, and template verification improves the retention of valid motor regions while suppressing pseudo-hotspot candidates. Method A processes more than 760 frames per second, although Methods B and C are slightly faster. Thus, the proposed method provides the best detection accuracy and a favorable accuracy–speed trade-off rather than the highest absolute processing speed. The final screened image groups are shown in Fig. 9.

Table 3 provides a more detailed comparison of the four screening strategies. At 30.7°C, Method A obtains a recall of 0.267, compared with 0.083 for Method B, and its F1 score reaches 0.421, compared with 0.153 for Method B. At 22.3°C, Method A also maintains the highest recall and F1 score, with values of 0.116 and 0.208, respectively. Its processing speeds are 763.52 FPS and 761.34 FPS at the two temperatures, which are lower than those of Methods B and C but substantially higher than those of Method D. These results show that the proposed pipeline prioritizes reliable motor-region retrieval while maintaining real-time processing capability. The comparison is visualized in Fig. 10.

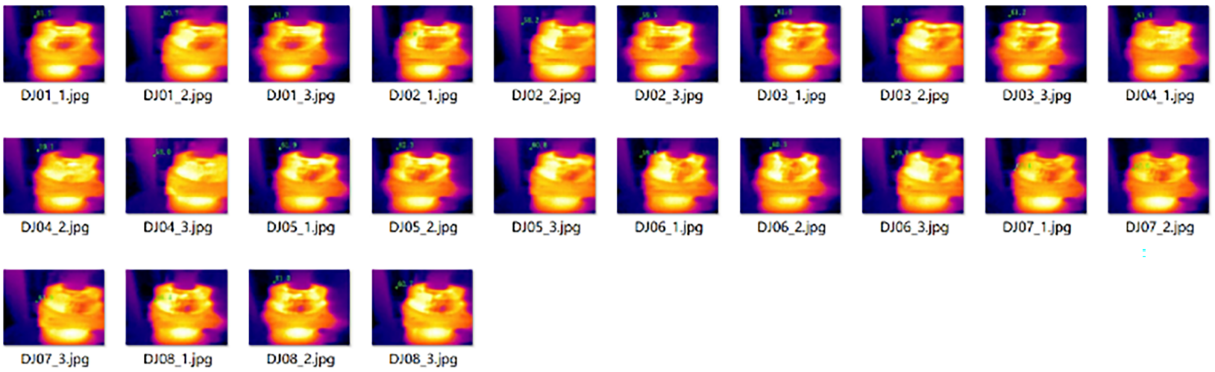


Figure 9: Final groups of screened images.

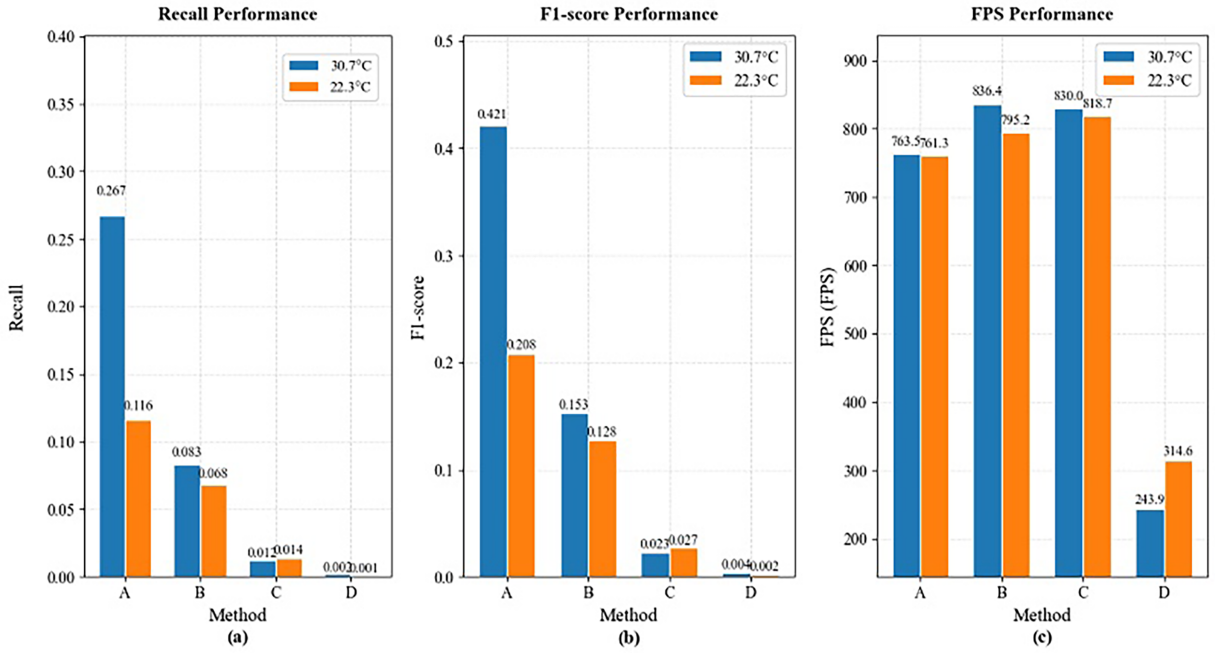


Figure 10: Comparison results of rapid screening methods. (**a**) Recall performance; (**b**) F1-score performance; (**c**) processing speed (FPS).

5.3 Super-Resolution Reconstruction Experiments

5.3.1 Evaluation Indicators

To verify the performance advantages of the proposed PRAGAN (Perception ResNet Attention GAN) in image super-resolution tasks, the network is comprehensively evaluated based on image reconstruction quality, detail restoration, and perceptual visualization. A side-by-side comparison with mainstream methods is conducted. In addition, ablation experiments are performed to assess the effectiveness of the perceptual loss and attention modules.

PSNR [26] which reflects the magnitude of the image error, the higher the better.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{22}$$

where *MAX* denotes the maximum pixel value in the image and *MSE* is the mean square error between the super resolution image and the high resolution image.

SSIM [27] which is particularly important in super-resolution tasks. Unlike the PSNR, SSIM takes into account not only the inter-pixel error, but also the structure, luminance, and contrast of the image for a more comprehensive assessment of the visual quality of the image, with the closer to 1 the better.

$$SSIM\,(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_x\sigma_y + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{23}$$

where μ is the arithmetic mean of the image, σ is the standard deviation of the image contrast, C_1 and C_2 are constants.

Higher PSNR and SSIM indicate that our method can recover finer thermal structural details from low-resolution infrared frames, which is beneficial for subsequent temperature estimation and thermal anomaly detection.

5.3.2 Experimental Analysis

In this study, the following hyper parameter settings were adopted to ensure the validity and performance of the model. To reduce computational load during training, the batch size was set to 32, the initial learning rate was 0.0001, and the Adam optimizer was used. All methods in the primary comparison were trained for 100 epochs, with validation performed after each epoch. An in-depth comparative analysis was conducted to evaluate the performance of SRCNN [25], RCAN [26], ESPCN [27], and the proposed method in the motor reconstruction task. Fig. 11 presents the generated outputs of these methods on the same dataset, while Table 4 summarizes their evaluation metrics.

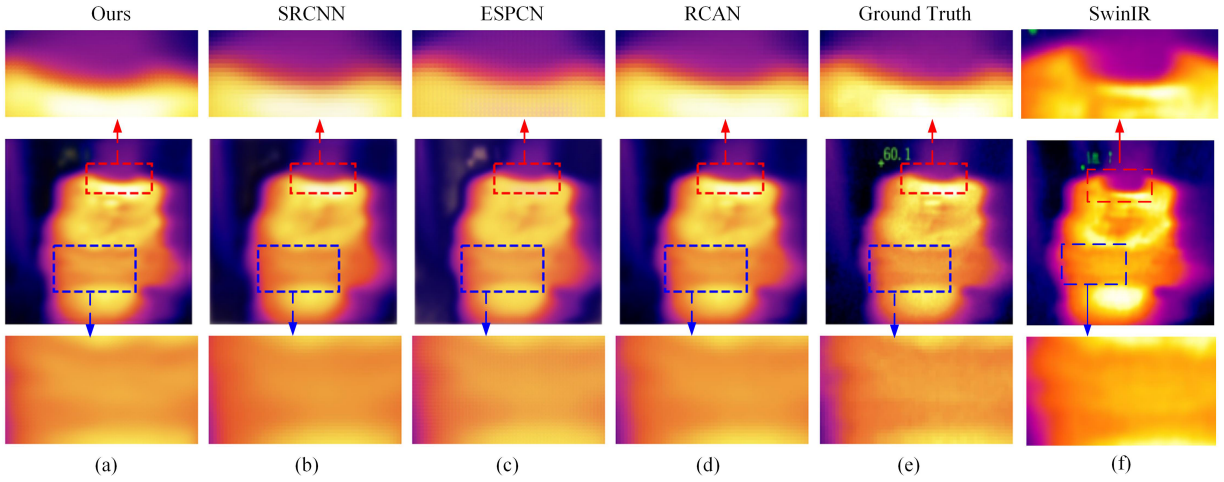


Figure 11: Results generated by different methods. (**a**) ours; (**b**) SRCNN; (**c**) ESPCN; (**d**) RCAN; (**e**) ground truth; (**f**) SwinIR.

The quantitative and visual results in Table 4 demonstrate the effectiveness of the proposed method for image super-resolution. It achieves the highest PSNR of 31.5420 and SSIM of 0.9829, indicating lower reconstruction error and better structural preservation than the compared models. Visual comparisons further show sharper edges and textures with fewer blurring and distortion artifacts. SRCNN and ESPCN

suffer from detail loss or over-smoothing, while RCAN remains inferior in reconstructing local high-intensity regions and structural contrast. These results confirm that the proposed architecture and loss function improve both pixel-level accuracy and structural fidelity.

Table 4: Experimental comparison of methods evaluation metrics of different algorithms on test set.

Method	PSNR	SSIM
Ours	31.5420	0.9829
SRCNN	26.2557	0.9632
ESPCN	28.2255	0.9722
RCAN	28.9315	0.9791
SwinIR	31.4190	0.9470

To evaluate the contribution of each module to overall model performance, ablation experiments were designed to analyze the individual effects of the ResNet structure, attention mechanism, and perceptual loss. The detailed results are presented in Table 5.

Table 5: Comparison of ablation experiments.

Model Configuration	PSNR	SSIM
Basic Model	26.723	0.8914
Basic Model + residual network (ResNet)	27.45	0.8580
Basic Model + RESNet + Attention	29.2395	0.9345
Basic Model + RESNet + Attention + Perceptual Loss	31.5420	0.9829

As shown in Table 5, the overall results demonstrate that the progressive integration of the proposed modules improves reconstruction quality. Combining ResNet with the attention mechanism increases the PSNR to 29.2395 and the SSIM to 0.9345, corresponding to improvements of 9.4% and 4.83% over the baseline, respectively. This indicates that residual feature extraction and attention-guided feature selection jointly strengthen image representation and the recovery of important structural regions. After perceptual loss is introduced, the PSNR further increases to 31.5420 and the SSIM reaches 0.9829, representing overall gains of 18.0% and 10.26%. These results support the complementary roles of residual learning, attention, and perceptual supervision in improving structural fidelity and visual detail.

The five-run stability analysis used different random seeds and a training budget of 30 epochs per run, whereas the primary model comparison in Table 4 used 100 epochs. Across the five runs, PRAGAN achieved an average PSNR of 29.1613 ± 0.9788 dB and an average SSIM of 0.8910 ± 0.0075. As shown in Table 6, PRAGAN contains 1.550 M parameters, requires 10.647 billion floating-point operations (GFLOPs), and achieves 27.76 FPS. Compared with RCAN, it reduces the parameter count and computational cost by 63.5% and 76.3%, respectively, while providing 4.62 times the inference speed.

To determine whether super-resolution can recover temperature-related spatial information lost during acquisition with low-resolution, high-frame-rate sensors, we evaluated hotspot-region overlap and peak-location accuracy after reconstruction. This experiment aimed to examine whether PRAGAN can preserve and localize critical thermal anomalies without sacrificing image acquisition speed, rather than relying solely on PSNR and SSIM to assess reconstruction quality. Bicubic interpolation, RCAN, Swin Transformer for Image Restoration (SwinIR), and PRAGAN were compared using the same random seed, training budget,

and test images. Fig. 12a evaluates the overlap between the reconstructed and HR-reference hotspot regions using hotspot-mask intersection over union (IoU), whereas Fig. 12b measures the spatial displacement of the hottest point using peak-location error. Green temperature annotations and crosshair markers were excluded from the analysis. Compared with bicubic interpolation, PRAGAN increased the hotspot IoU from 0.8762 to 0.8920 and reduced the peak-location error from 10.43 to 5.78 pixels, demonstrating improved hotspot localization.

Table 6: Comparison of computational complexity and inference efficiency.

Method	Params (M)	GFLOPs	FPS
SRCNN	0.069	12.622	168.66
ESPCN	0.037	26.356	684.67
RCAN	4.242	44.873	6.01
SwinIR	0.413	8.258	10.08
PRAGAN	1.550	10.647	27.76

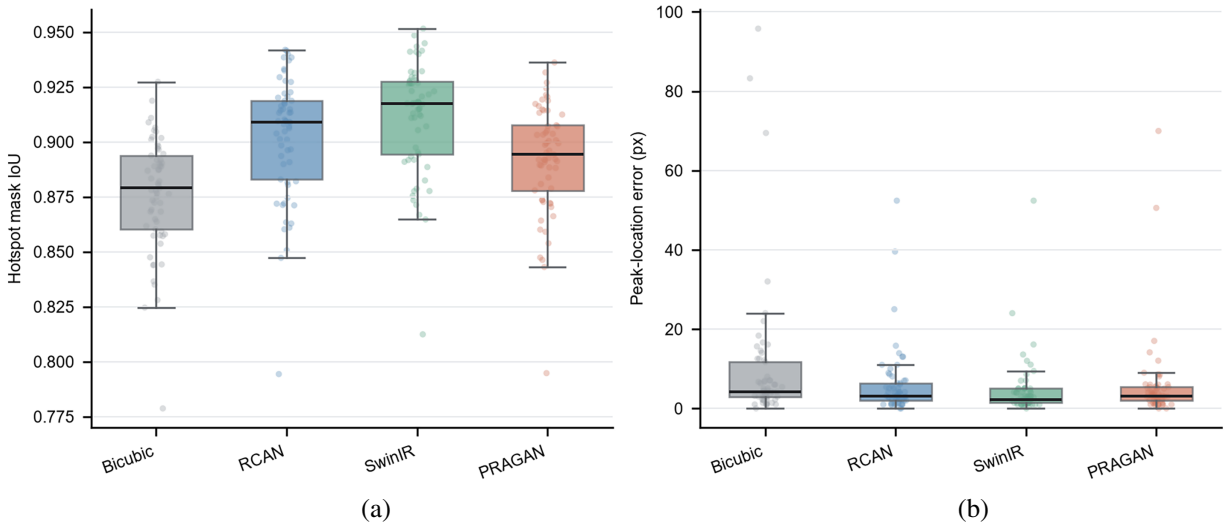


Figure 12: Hotspot localization performance (**a**) Hotspot-mask intersection over union. (**b**) Peak-location error.

5.4 Precision Analysis

To further verify the stability and accuracy of the experimental system, temperature data from various components were collected over 20 passes within 2 days during a selected week and compared with data obtained from other systems. For each component, the experimental system recorded the highest observed temperature. These measurements were compared with actual values to analyze the system's temperature measurement deviation. Temperature data were collected from 160 motors located in different parts of 10 carriages. To visualize the distribution of motor temperature deviations, the differences between the true values and the system-measured temperatures are plotted, as shown in Fig. 13.

As shown in Fig. 13, the absolute deviations between the system measurements and the reference values remained below 4°C. Across 20 train passages and four measurement positions, the mean absolute deviation was 1.9817°C, with a standard deviation of 0.702°C, as summarized in Table 7. The remaining

deviation mainly arises because the proposed system calculates an average temperature over the motor-bottom region, whereas the reference system measures local regions that are more susceptible to heating. The results demonstrate stable temperature measurement during train operation at 55 km/h, as shown in Fig. 14.

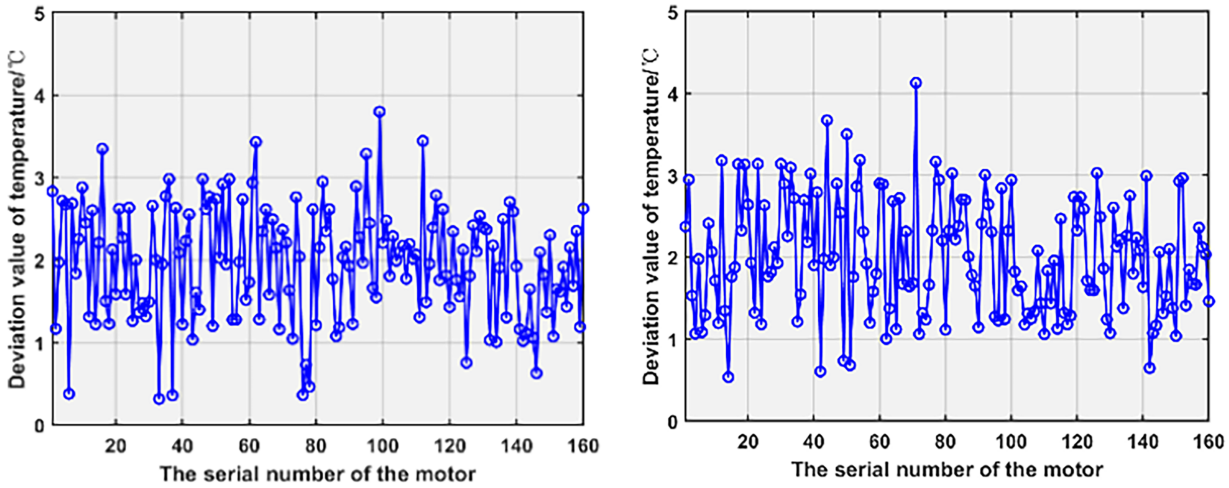


Figure 13: Error between true value and system detection.

Table 7: Error analysis.

	Group 1	Group 2	Mean Error
Mean error/°C	1.9601	2.0032	1.9817
Standard deviation/°C	0.6876	0.7164	0.702



Figure 14: Actual speed at the system installation location.

Currently, motor-based temperature detection systems mainly include contact temperature sensors, non-contact temperature sensors and infrared thermal camera methods. To evaluate the accuracy of the proposed system, its performance was compared with previously reported contact and non-contact temperature-monitoring systems, as summarized in Table 8.

Table 8: Comparison of reported temperature-measurement performance.

	Reported Error or Accuracy/°C	Component	Speed/Mounting Position
Ours	±2	Motor	55 km/h
Contact [7]	±1	Axle box	Vehicle depot
Non-contact [16]	±1	Axle box, Motor, Elv box	Vehicle depot
Non-contact [18]	2	Motor, Gearbox	5–10 km/h
Non-contact [15]	/	Axle box	65–75 km/h

As shown in Table 8, the proposed system achieved a mean absolute temperature deviation of 1.9817°C during operation at 55 km/h, with a maximum observed deviation below 4°C. Previous systems reported values ranging from ±1°C to 2°C under different components, operating speeds, and measurement protocols. Because these studies used different definitions of accuracy and different test conditions, the reported values should be interpreted as an engineering comparison rather than a strictly controlled quantitative ranking.

6 Conclusion

This study proposes a motor temperature detection system that utilizes multi-scale feature screening and thermal image reconstruction to evaluate motor operating status based on collected thermal data. An infrared thermal camera captures thermal video of the train during mainline operations. The recorded data is transmitted to a server, where multi-scale feature screening and thermal image reconstruction are performed. Experimental results show that the system achieves a mean absolute temperature deviation of 1.9817°C, with a standard deviation of 0.702°C and a maximum observed deviation below 4°C, during train operation at 55 km/h. Notably, in addition to infrared thermal imaging, visible light images offer valuable references for assessing motor appearance and analyzing the surrounding environment. These images provide field engineers with a more intuitive basis for evaluation. Therefore, future work will focus on integrating visible light image data to enhance overall sensing capabilities and improve the accuracy of motor condition evaluation.

Acknowledgement: The authors would like to thank the field experts for their effective suggestions during the commissioning of the system.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Juhui Zhang contributed to conceptualization, methodology, software development, validation, and manuscript drafting, Zongyi Xing and Chenxiao Cai supervised the study and reviewed and edited the manuscript. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to the operational data-management requirements of Nanjing Metro.

Ethics Approval: Not applicable. This study did not involve human participants, human biological materials, or animals.

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

References

1. Zhao J, Fu Y, Ji G, Liu S, Zhou C. Research on over-temperature fault prediction technology of traction motors in subway vehicles. Electron Compon Inf Technol. 2024;8(7):71–4. (In Chinese). doi:10.19772/j.cnki.2096-4455.2024.7.022.
2. Ma J, Yang M, Qian B, Li L, He D, Lien FS. Experimental study on temperature rise characteristics and influencing factors of underbody motors in subway trains. Therm Sci Eng Prog. 2025;58:103230. doi:10.1016/j.tsep.2025.103230.
3. Wang X. Train axle temperature detection and communication design based on wireless sensor. Int J Syst Assur Eng Manage. 2023;3:1–12. doi:10.1007/s13198-023-01986-1.
4. Tao J. Research and application of auto-generating vehicle-mounted axle temperature detection system for Shanghai rail transit train. Urban Mass Transit. 2023;26(1):132–6. (In Chinese). doi:10.16037/j.1007-869x.2023.S1.029.
5. Lee YS, Lee BG, Hwang IK, Han SH. Implementation of the measurement equipment to measure return current and axle temperature of high speed railway. Trans Korean Inst Electr Eng. 2016;65(4):695–703. doi:10.5370/kiee.2016.65.4.695.
6. Zheng XY, Liu PJ, Bi FY. Adaptability analysis of trace hotbox detection system for railway heavy duty freight cars. Railway Locomot Car. 2022;42(5):52–5. (In Chinese). doi:10.3969/j.issn.1008-7842.2022.05.08.
7. Lei L, Song D, He P, Lin H. Research on rapid position of axle box in high-noise infrared images acquired by trackside. IEEE Sens J. 2024;24(1):554–63. doi:10.1109/JSEN.2023.3334013.
8. Wu CH, Xia ZC, Gao YF, Chen XY, Zhou X, Liu KF. Research and design of ground monitoring system for metro vehicle running gears. Urban Rapid Rail Transit. 2021;34(4):69–74. (In Chinese). doi:10.3969/j.issn.1672-6073.2021.04.011.
9. Du K, Jiao S, Li Y, Liu W, Lei Z, Xin J. Bogie temperature detection method for subway trains based on dual optical images. In: Intelligent robotics. Singapore: Springer Nature; 2023. p. 257–66. doi:10.1007/978-981-99-0301-6_20.
10. Zhang J, Xing Z, Cai C, Chen Y. Infrared image-based system for the condition monitoring of key bogie components. IEEE Sens J. 2025;25(8):13316–29. doi:10.1109/JSEN.2025.3542236.
11. Kim J. Non-destructive characterization of railway materials and components with infrared thermography technique. Materials. 2019;12(24):4077. doi:10.3390/ma12244077.
12. Bagavathiappan S, Lahiri BB, Saravanan T, Philip J, Jayakumar T. Infrared thermography for condition monitoring—a review. Infrared Phys Technol. 2013;60:35–55. doi:10.1016/j.infrared.2013.03.006.
13. Usamentiaga R, Venegas P, Guerediaga J, Vega L, Molleda J, Bulnes FG. Infrared thermography for temperature measurement and non-destructive testing. Sensors. 2014;14(7):12305–48. doi:10.3390/s140712305.
14. Parzen E. On estimation of a probability density function and mode. Ann Math Statist. 1962;33(3):1065–76. doi:10.1214/aoms/1177704472.
15. Liu J, Dai S, Guo Z, Zhang D. An improved POCS super-resolution infrared image reconstruction algorithm based on visual mechanism. Infrared Phys Technol. 2016;78:92–8. doi:10.1016/j.infrared.2016.07.010.
16. Liu H, Yuan K, Shen B, Xing W, Huang M. Blind super-resolution reconstruction of infrared images based on dual-domain feature extraction. Opt Lasers Eng. 2025;193:109109. doi:10.1016/j.optlaseng.2025.109109.
17. Feng X, Li X, Li J. Multi-scale fractal residual network for image super-resolution. Appl Intell. 2021;51:1845–56. doi:10.1007/s10489-020-01909-8.
18. Dong C, Loy CC, Tang X. Accelerating the super-resolution convolutional neural network. In: Computer Vision—ECCV 2016; 2016 Oct 11–14; Amsterdam, The Netherlands. p. 391–407. doi:10.1007/978-3-319-46475-6_25.
19. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y. Image super-resolution using very deep residual channel attention networks. In: Computer Vision—ECCV 2018; 2018 Sep 8–14; Munich, Germany. p. 294–310. doi:10.1007/978-3-030-01234-2_18.
20. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 1874–83. doi:10.1109/CVPR.2016.207.

21. Sun Z, Zhang J, Chen Z, Hong L, Zhang R, Li W, et al. Image super-resolution reconstruction using Swin Transformer with efficient channel attention networks. Eng Appl Artif Intell. 2024;136:108859. doi:10.1016/j.engappai.2024.108859.
22. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R. SwinIR: image restoration using swin transformer. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops; 2021 October 11–17; Montreal, BC, Canada. p. 1833–44. doi:10.1109/ICCVW54120.2021.00210.
23. Deng F, Wang S, Yang J. Region-aware discriminative learning GAN for super-resolution reconstruction of infrared imagery. Neurocomputing. 2025;639:130202. doi:10.1016/j.neucom.2025.130202.
24. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44. doi:10.1145/3422622.
25. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Internet]. 2015 [cited 2026 Jan 1]. Available from: https://mlanthology.org/iclr/2015/simonyan2015iclr-very/.
26. Wang Z, Bovik AC. Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process Mag. 2009;26(1):98–117. doi:10.1109/MSP.2008.930649.
27. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12. doi:10.1109/tip.2003.819861.