ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

# Efficient Image Super-Resolution Guided by Object Detection

Zihan Han*

School of Automation and Electrical Engineering University of Jinan, Jinan 250200, China

*Corresponding author: 15689023658@163.com

## Abstract

**To address the challenges of reduced target recognition accuracy and insufficient real-time performance caused by low image resolution in traditional tower assembly construction, this paper proposes an efficient image super-resolution method guided by object detection. The proposed approach employs YOLOv8 to rapidly locate target regions and introduces a discriminative image segmentation module to divide the image into key and background regions. These regions are then processed separately using a specially designed Super-Resolution Guided Efficient Detail Recovery Network (SGERN) and a Background Lightweight Super-Resolution Network (BLSRN), thereby enhancing image quality while significantly reducing computational cost. The SGERN integrates Generative Adversarial Networks (GAN), a simple gating mechanism, and re-parameterization modules to achieve high-quality reconstruction of fine textures in key regions. Experiments conducted on a self-built dataset of hoisting components and public datasets (UDM10 and Urban100) demonstrate that the proposed method outperforms mainstream approaches in terms of PSNR and SSIM. Moreover, it achieves good real-time performance while ensuring high visual quality in the target areas. This study provides a novel and practical solution for target recognition and image enhancement in complex construction scenarios, with promising engineering application value.**

## Keywords

**Neural Network; Machine Vision; Tower Construction.**

## 1. Introduction

Transmission towers are a vital component of the power transmission system, serving essential roles such as supporting power lines, regulating line tension, and ensuring power system safety [1]. These structures are typically installed in remote or geographically complex areas such as mountains, deserts, and plateaus, and require the use of various lifting equipment for assembly. Tower construction is a high-risk and labor-intensive task, requiring workers to operate at great heights under complex and variable environmental conditions. During traditional tower assembly, crane operators rely solely on visual cues or walkie-talkie communication with workers at height to transport components to the installation location, where they are manually adjusted and fixed. This process is prone to significant subjective errors and is vulnerable to communication interference, thus posing safety risks.

In recent years, machine vision has seen increasing application in industrial fields. As an interdisciplinary technology, machine vision involves how computers analyze digital images and videos, aiming to generate a clearer understanding of objects within images [2]. Image-based object detection, in particular, can greatly enhance detection efficiency and intelligence. When applied to the target classification tasks in tower construction, it replaces manual observation for recognizing targets at high altitudes, offering advantages such as strong continuity, high reliability, and the ability

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

to detect a wide range of object types. Image super-resolution, as an image enhancement technique, improves overall resolution, making details and textures more discernible and facilitating more effective feature analysis compared to the original images.

When the camera is positioned far from the hoisting components, the target occupies only a small portion of the image pixels, making it difficult for low-resolution images to convey sufficient object features. Although using high-resolution images can preserve more details, it also introduces excessive information redundancy, leading to significantly higher transmission and computation costs—making it difficult to meet real-time performance requirements. To balance detection accuracy and real-time efficiency, this paper proposes a super-resolution reconstruction method guided by object detection. This method is detailed from two aspects: a discriminative image segmentation module guided by object detection, and an efficient detail recovery network based on simple gating. These components enable localized and efficient super-resolution reconstruction of low-resolution target regions to ensure both real-time performance and accuracy.

## 2. Related Work

Among classical object detection algorithms, the "You Only Look Once" (YOLO) series is a prominent representative. In 2024, the latest version, YOLOv8 [3], was released, which further enhanced the model's generalization capability by incorporating improved modules for input data augmentation and feature output refinement, leading to higher accuracy and broader applicability. The evolution of the YOLO models from fast detection to increasingly precise recognition has significantly advanced the development of object detection technology.

Single-image super-resolution (SISR) methods based on interpolation primarily rely on upsampling algorithms to process low-resolution images and generate higher-resolution outputs. Common interpolation techniques include bilinear and bicubic interpolation. These methods are simple to implement and computationally efficient, and can improve image resolution to some extent. However, due to their lack of deep understanding of image content, they often fail to recover high-frequency texture details, resulting in limited image quality with noticeable blurring and artifacts.

In contrast, learning-based SISR approaches have demonstrated superior performance in image reconstruction quality, particularly in restoring fine textures and resisting noise, making them the mainstream direction of current research. Ledig et al. [4] introduced Generative Adversarial Networks (GAN) into the super-resolution task, proposing the Super-Resolution Generative Adversarial Network (SRGAN), which effectively addressed the issue of overly smooth and detail-lacking images produced by traditional methods. Subsequently, other GAN-based approaches [5–7] have also made significant progress in improving image resolution while preserving realistic detail and perceptual quality.

In terms of network lightweighting, beyond adopting shallower architectures, Wang et al. [8] applied a knowledge distillation strategy to achieve better performance with fewer parameters. While significant progress has been made in addressing key challenges in SISR, increasing demands for real-time performance and fine-detail restoration in engineering applications still present challenges. Although lightweight models have achieved some breakthroughs, they are still insufficient for domains with strict real-time requirements. In particular, further improvements are needed in processing images with unclear textures or small, distant targets. To address these limitations, this paper proposes an efficient image super-resolution method guided by object detection to better meet the needs of practical applications.

## 3. Method Introduction

The overall framework of the proposed efficient image super-resolution method guided by object detection is illustrated in Figure 1. YOLOv8 is employed for object detection to rapidly identify target regions. To ensure real-time performance in super-resolution computation, a discriminative image segmentation module guided by object detection is designed. Based on the detection results, the input

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

image is divided into two parts: the Guided Image (containing the target region) and the Background Image (non-target regions).In the target regions, the proposed Simple Gated Efficient Detail Recovery Network (SGERN) is used to preserve and enhance fine texture details. For the background regions, a lightweight model named Background Lightweight Super-Resolution Network (BLSRN) is utilized to reduce computational cost and ensure rational allocation of processing resources.

To further enhance the detail recovery capabilities and reconstruction quality of SGERN, a Generative Adversarial Network (GAN) framework is integrated into its architecture. This method achieves a balance between high-quality super-resolution reconstruction and computational efficiency, making it well-suited for real-time tower assembly applications where both speed and visual fidelity are critical.

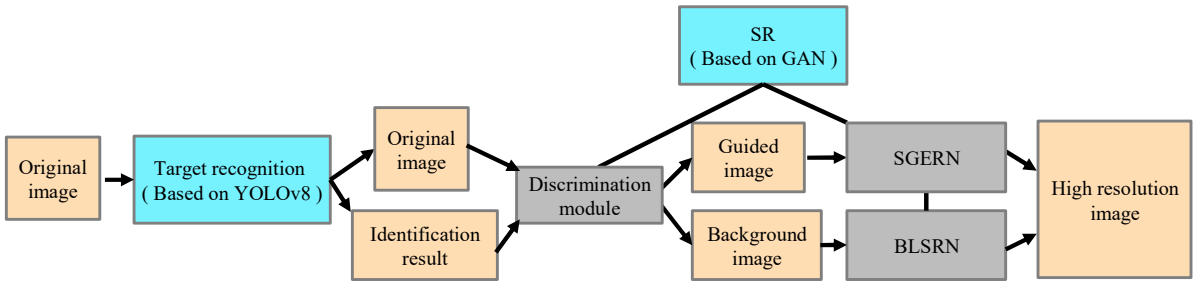


**Figure 1.** Method flow

### 3.1 Discriminative Image Segmentation Module and BLSRN

To ensure real-time performance in image super-resolution, a region-wise processing strategy is adopted through a discriminative image segmentation module. This module relies on object detection results to divide the image into distinct regions, effectively preventing excessive computational resource consumption and enabling a balance between precision and efficiency in practical engineering applications.

As shown in Figure 2, after object detection is performed on the hoisting component image, the bounding box of the target can be obtained based on the detection results. The area within this bounding box is designated as the high-precision super-resolution target region, with its boundaries defined by the pixel coordinates output by the object detection algorithm. To enable more refined processing within this region, the proposed method applies a binary mask $M(x,y)$, which is calculated according to Equation (1) as follows:

$$M(x,y) = \begin{cases} 1, & x_1 \le x \le x_2, \quad y_1 \le y \le y_2 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

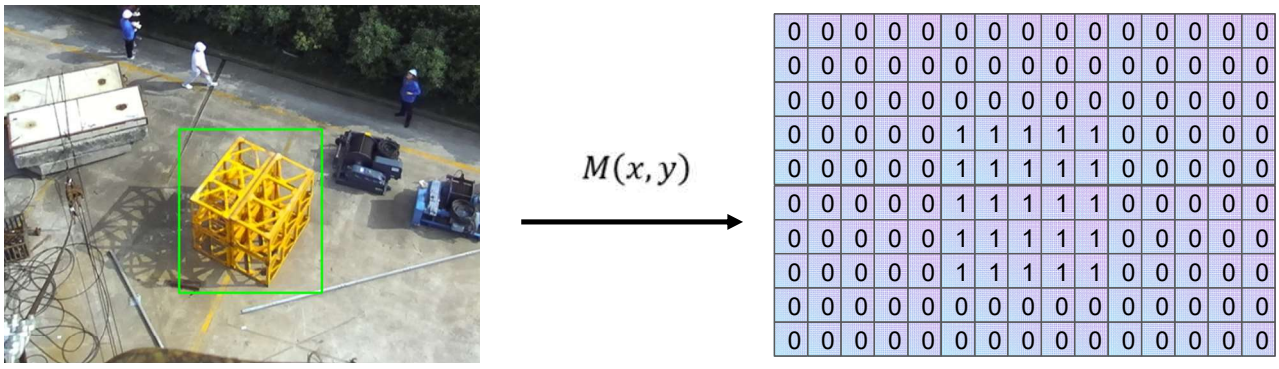


**Figure 2.** Identification result mask conversion diagram

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

In the equation, $(x_1, y_1)$ to $(x_2, y_2)$ denote the coordinate range of the target region obtained from the object detection results.

In the matrix, the region with a value of 1 represents the target area, which is defined in this paper as the guided region $P_1$, while the region with a value of 0 corresponds to the background region $P_0$.

Next, a discriminative gating mechanism is introduced, taking both the discriminative matrix and the feature map as inputs and aligning them accordingly. Through this gating operation, features located in the guided region can be precisely extracted, effectively segmenting the target area from the original image. This allows for subsequent fine-grained processing. The effect of the discriminative gate module is illustrated in Figure 3.

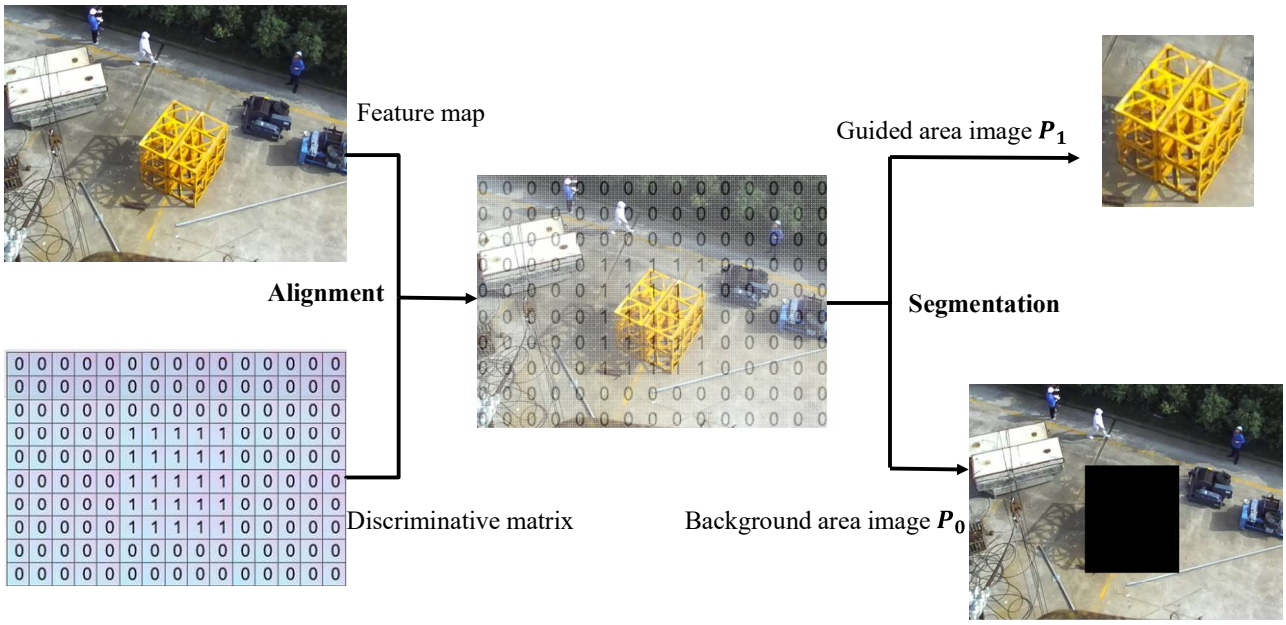


**Figure 3.** Schematic diagram of the discriminant gate module processing

To improve computational efficiency and reduce parameter overhead, BLSRN as illustrated in Figure 4-is designed for super-resolution reconstruction of the background region $P_0$. BLSRN utilizes a compact model for feature extraction and incorporates a Channel-Split Feature Extractor (CSFE) in a stacked configuration. The CSFE adopts a channel-splitting strategy to divide the input feature map into multiple segments, each of which is processed separately for feature extraction. This approach effectively reduces computational cost and parameter complexity while preserving essential feature information. After initial processing through a 3×3 convolution, the image features are passed into the CSFE. The extracted features are then fed through a 2×2 convolution layer, followed by sub-pixel convolution for upsampling, resulting in the super-resolved background image $P_{0SR}$.

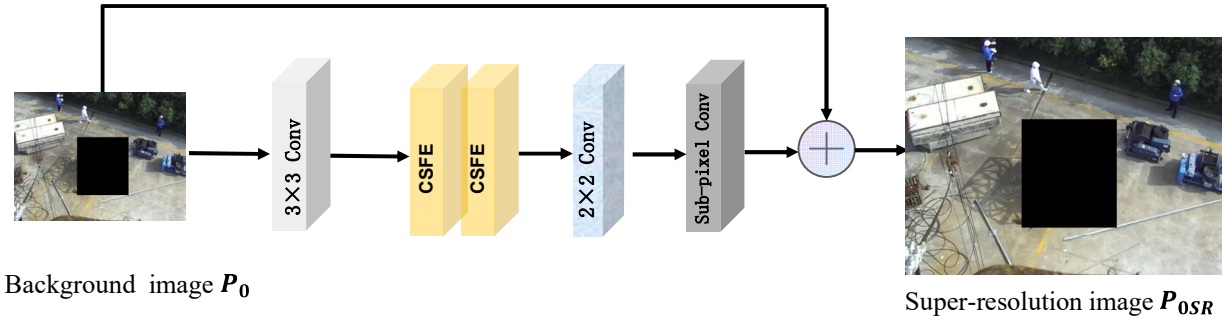


**Figure 4.** BLSRN structure

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

### 3.2 Super-Resolution Guided Efficient Detail Recovery Network introduction

To enhance the detail recovery capability and reconstruction quality of the image, GAN is introduced to perform super-resolution reconstruction on the guided region $P_1$. The training approach of GAN does not significantly increase network complexity, thereby avoiding additional computational overhead.Building on this, SGERN is proposed to more effectively reconstruct high-quality super-resolved images. The overall architecture of SGERN consists of a generator and a discriminator. The generator integrates two key modules: the Simple Gated Reconstruction Module (SGPM) and the Re-parameterization Reconstruction Module (RRM), which together enable efficient extraction of fine-grained features and enhancement of image quality. Bicubic interpolation is used as a preprocessing step for upsampling, aiming to improve both super-resolution performance and inference efficiency. The discriminator adopts a standard fully convolutional structure to enhance discriminative ability and improve the realism of the reconstructed images. The architecture of the SGERN generator is shown in Figure 5.

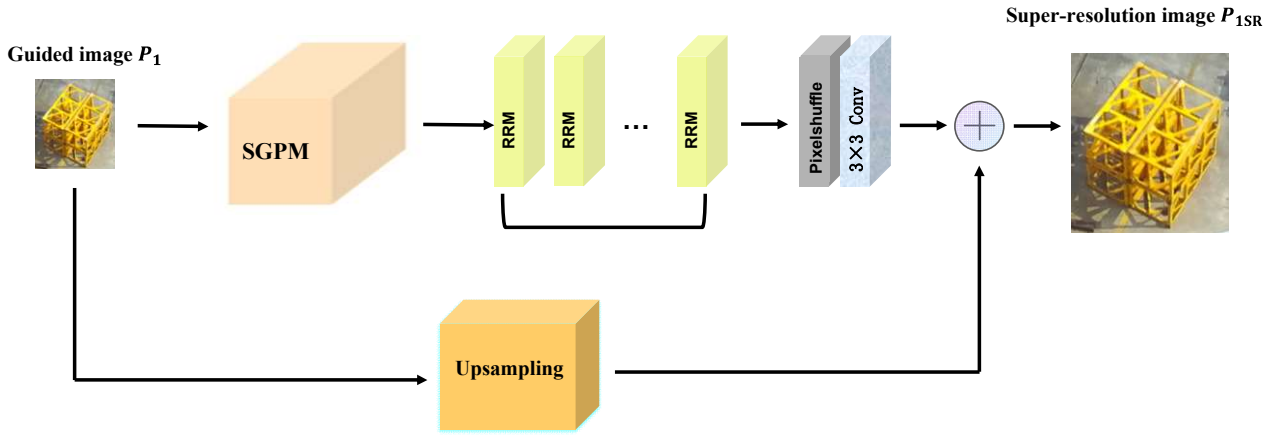


**Figure 5.** SGERN structure

In complex video scenarios, texture details are often difficult to accurately restore, and the accumulation of errors may lead to the loss of fine details or the generation of artifacts. To address this issue, a gating mechanism is introduced to effectively mitigate error propagation and enhance information filtering capabilities. By dynamically modulating the features, the gating mechanism enables the network to better select and preserve critical information, thereby improving the quality of detail reconstruction.Due to its efficient information processing ability, the gating mechanism has been widely applied in tasks such as image super-resolution and denoising. The simple gating mechanism adopted in this work not only preserves important details but also reduces redundant computations, thereby improving overall efficiency.As illustrated in Figure 6, the input feature map channels are evenly split into two parts, and , and the output is computed using Equation (2) as follows:

$$F_{\text{out}}(i,j) = F_a(i,j) \times F_b(i,j) \tag{2}$$

In the equation, $F_a$ serves as the gating component to suppress redundant feature interference by determining which features should be enhanced or suppressed, while $F_b$ acts as the primary source of feature information.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

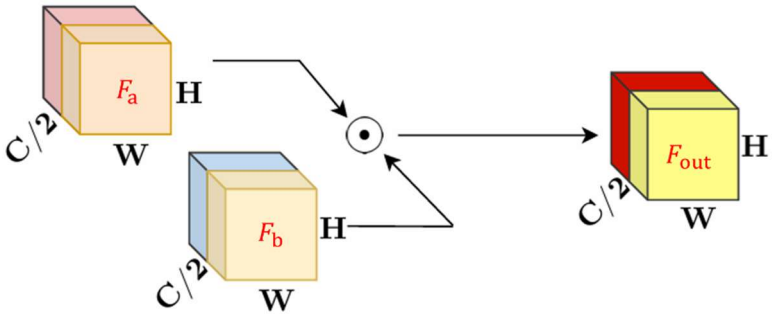


**Figure 6.** Simple gating structure

To better reconstruct detailed features in complex video scenarios, and inspired by the work of Chen et al. [9], we design the SGPM, as illustrated in Figure 7.First, the input guided image $P_1$ undergoes a 3×3 convolution followed by normalization for initial feature extraction. The features are then passed through two parallel branches to enhance representational capability:

In the first branch, a 1×1 convolution is applied to capture local features with a small receptive field. A simple gating mechanism is introduced to adaptively filter and enhance key information, resulting in feature map $F_1$.

The second branch consists of a 1×1 convolution followed by a 3×3 convolution in series, aiming to extract rich detail features and strengthen the contextual representation of complex regions. A simple gating mechanism is also applied here to produce feature map $F_2$.

Next, the outputs of both branches are fused through a gated interaction, where an element-wise multiplication is performed to generate $F_3$. Since $F_3$ aggregates feature channels from different branches, a 1×1 convolution is used to merge information across channels and reduce dimensionality, yielding feature map $F_4$, which helps minimize computational load in the following stages.

To further enhance the representational strength of the features, a Lightweight Channel Attention (LCA) module is introduced for feature recalibration. The process is described by Equation (3):

$$F_{LCA}=F_4 \otimes (\mathrm{Conv}_{1\times1}(Pooling(F_4))) \tag{3}$$

In this equation, Pooling(·) denotes global average pooling to extract global channel-level information, $\mathrm{Conv}_{1\times1}$ represents a 1×1 convolution to adjust inter-channel weights, and ⊗ indicates channel-wise multiplication, where the generated attention weights are applied to enhance the original feature map $F_4$.

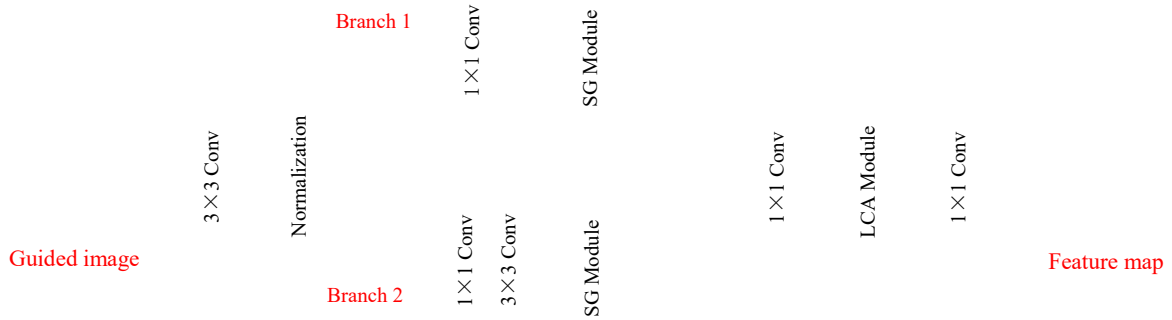


**Figure 7.** SGPM structure

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

The resulting attention-enhanced feature map $F_{LCA}$ is then passed through a 1×1 convolution and added to the residual path of the input image $P_1$, which has undergone a 3×3 convolution. This completes the SGPM computation and yields the final output feature map $F_{SGPM}$.

In the subsequent feature processing stage, residual network layers are employed to enhance feature representation and improve the network's nonlinear modeling capabilities. However, traditional residual blocks may introduce redundant computations and parameter overhead, which can impact inference efficiency.Inspired by Ding et al. [10], this paper incorporates the RRM into the generator network, as shown in Figure 8. During training, the module is designed with sufficient depth and redundancy to improve the model's fitting capacity. During inference, RRM leverages structural re-parameterization to fold multiple convolution layers into a single equivalent 3×3 convolution layer, effectively reducing computational cost and accelerating inference, thereby enhancing deployment efficiency.The feature map $F_{SGPM}$, obtained from the SGPM, is passed through multiple re-parameterized blocks to produce $F_{RRM}$.

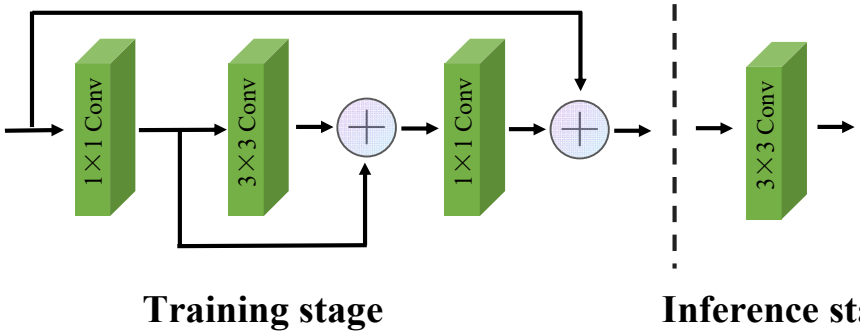


**Figure 8.** RRM structure

This output $F_{RRM}$ is then processed by a PixelShuffle module for feature rearrangement, and combined with a series of 3×3 convolution layers to enhance features and restore spatial resolution. Finally, it is added to a bicubic-interpolated version of the input guided image $P_1$, producing the final super-resolved output of the target region, denoted as $P_{1S}$ .The super-resolved background image $P_{0SR}$, previously generated by the lightweight background network, is then concatenated with $P_{1SR}$ to reconstruct the full image in high resolution.

## 4. Experiments and Results

### 4.1 Experimental Setup

Cross structure | Double-cross structure integral sheet | Standard section

**Figure 9.** Hoisting component sample

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

(1) Datasets: As shown in Figure 9, the Hoisting Component Dataset consists of 150 images captured on-site by a construction company. For each scene, images were taken at multiple resolution settings using the same camera setup, resulting in a multi-resolution dataset that includes low, medium, and high-resolution versions of each image.

To improve the model's generalization capability and strengthen the experimental reliability, two public datasets were also used for supplementary evaluation. UDM10 Test Set: Contains 320 high-quality images covering a wide range of scenes. These images exhibit rich detail and high-definition textures, with a resolution of 1280×720. Urban100 Dataset: Includes 100 images focused on urban architectural environments, which closely align with the visual characteristics of tower assembly scenes. The dataset contains various types of texture details and is well-suited for evaluating the performance of the proposed simple gating-based detail recovery network in texture reconstruction tasks.

(2) Evaluation Metrics: The experiments in this chapter adopt two widely-used objective metrics for evaluating super-resolution quality: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). All evaluations are performed on the luminance (Y) channel in the YUV color space. The scaling factors for both training and testing are set to 2× and 4×.

(3) Experimental Settings: The experiments were conducted on an Ubuntu 18.04 system equipped with an NVIDIA GTX 3090 GPU, 24 GB of RAM, CUDA 11.2, PyTorch 1.8, and Python 3.9. The Adam optimizer was used for training with a batch size of 16. For each batch, the input images were cropped into patches of size 64×64. The momentum parameters were set to $\beta_1$=0.9 and $\beta_2$=0.99, with an initial learning rate of 0.0001.

### 4.2 Analysis of Experimental Results

To verify the performance of the method proposed in this paper, SGERN is compared with five popular SR algorithms: RCNN, FSRCNN, SRDCNN [11], MSRN [12], and SelfExSR [13]. Additionally, due to the network architecture's use of Bicubic interpolation, Bicubic is included as a control group. The comparison results of SGERN with other mainstream networks for 2x and 4x super-resolution, evaluated using PSNR and SSIM as metrics, are shown in Tables 1 and 2. Results from other public datasets, except for the lifting component dataset, are cited from the respective network papers. The experimental results demonstrate that, for scaling factors of 2x and 4x, the SGERN proposed in this chapter achieves the highest evaluation metrics on both a self-built dataset and two public datasets.

**Table 1.** The comparison results of multiple networks for 2x super-resolution

| Network | Scale | Diaojian(150) PSNR/SSIM | UDM10(320) PSNR/SSIM | Urban100 PSNR/SSIM |
|---|---|---|---|---|
| Bicubic | 2 | 28.39/0.844 | 27.69/0.848 | 26.89/0.841 |
| SRCNN | 2 | 31.03/0.889 | 30.27/0.886 | 29.53/0.896 |
| FSRCNN | 2 | 31.38/0.892 | 30.66/0.911 | 29.88/0.902 |
| SRDCNN | 2 | 33.43/0.891 | 32.75/0.930 | 31.93/0.925 |
| MSRN | 2 | 33.72/0.901 | 32.97/0.927 | 32.22/0.933 |
| SelfExSR | 2 | 31.05/0.887 | 30.23/0.896 | 29.55/0.898 |
| SGERN | 2 | **33.94/0.902** | **33.14/0.938** | **32.44/0.933** |

By training the model to convergence, a visual comparison of images processed by various super-resolution networks was also conducted. The region with rich textures (Region 1) in the image was magnified and qualitatively analyzed. Figure 10 presents the results of 2x super-resolution reconstruction on standard images from the diaojian dataset. Most methods are able to recover the main structural features, but the details become blurred, indicating that these models have not effectively learned the image gradient information. The MSRN network also exhibited pixel shift

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

issues, demonstrating its limited generalization ability. In contrast, the proposed SGERN network outperformed the others in every aspect, showing the best detail edge recovery comparable to SelfExSR, and presenting the most realistic texture details without any noticeable edge distortion.

**Table 2.** The comparison results of multiple networks for 4x super-resolution

| Network | Scale | Diaojian(150) PSNR/SSIM | UDM10(320) PSNR/SSIM | Urban100 PSNR/SSIM |
|---|---|---|---|---|
| Bicubic | 4 | 21.81/0.649 | 23.98/0.656 | 23.15/0.660 |
| SRCNN | 4 | 23.18/0.741 | 25.22/0.743 | 24.53/0.725 |
| FSRCNN | 4 | 23.14/0.734 | 25.40/0.717 | 24.62/0.728 |
| SRDCNN | 4 | 25.29/0.799 | 27.30/0.795 | 26.64/0.803 |
| MSRN | 4 | 24.55/0.784 | 26.84/0.775 | 26.04/0.790 |
| SelfExSR | 4 | 23.47/0.730 | 25.51/0.759 | 24.83/0.740 |
| SGERN | 4 | **25.36/0.810** | **27.45/0.807** | **26.68/0.805** |

Figure 11 shows the results of 4x super-resolution reconstruction on samples from the Urban100 dataset. The objective evaluation metrics of the proposed method exhibit the best performance. When compared to the real images, the proposed method performs the most realistically in terms of detail restoration for high-magnification super-resolution images. The SGERN network demonstrated remarkable texture recovery, with no distortion, pixel shift, or edge distortion issues.

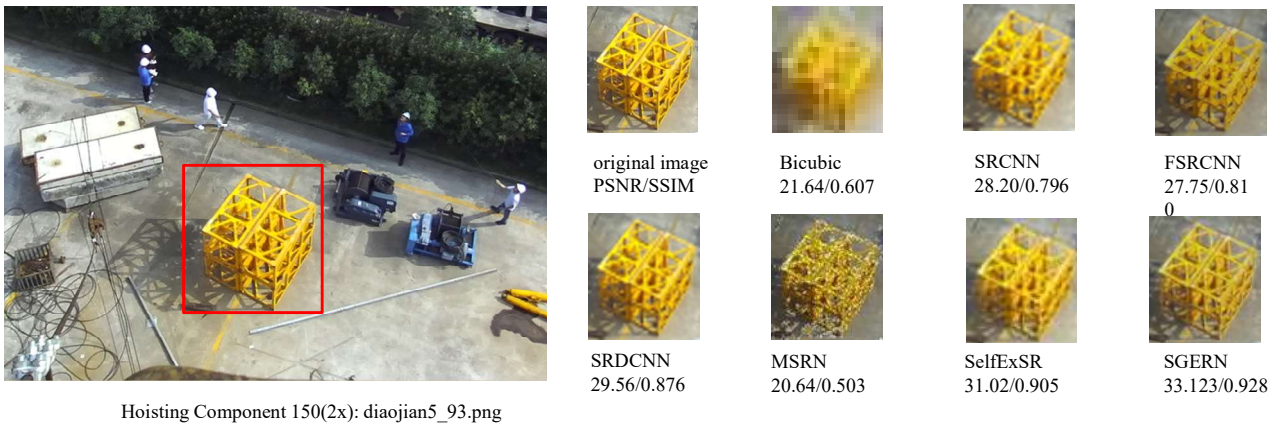


**Figure 10.** Comparison of visual effects with twice the SR of Hoisting Component Dataset

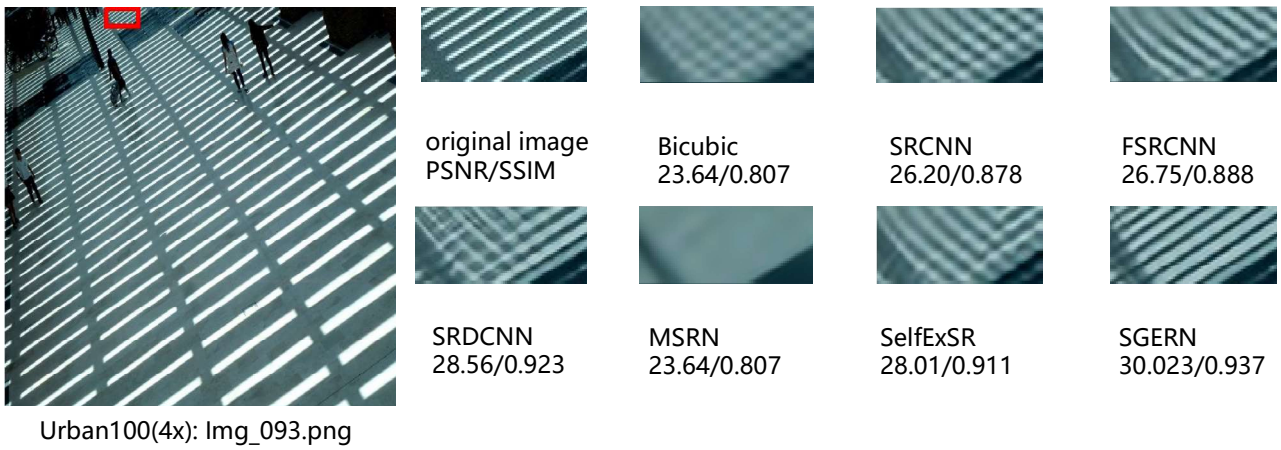


**Figure 11.** Comparison of visual effects with four times the SR of Urban100 Dataset

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

Since the target recognition-guided discriminative image segmentation module relies on the target recognition results for image segmentation, the verification experiment of its computational efficiency must be conducted on the self-built lifting component dataset. The core of the experiment lies in evaluating the speed of applying super-resolution processing to different segmented regions and comparing it with the full-image super-resolution processing using SGERN and BLSRN. To this end, images with a resolution of 1920×1080 from the lifting component dataset are used, with the results of target recognition as fixed parameters, and three sets of comparative experiments are designed to verify the performance, using frame rate (FPS) as the evaluation metric.

The experimental results are shown in Table 3: using BLSRN for full-image super-resolution reconstruction has an extremely fast computational speed, far exceeding the full-image super-resolution using SGERN (only 4.57 FPS). After introducing the discriminative segmentation module, the super-resolution processing speed is significantly improved compared to the full-image SGERN approach, with only 7.08 fewer frames processed per second than the full-image BLSRN. The experimental results indicate that the introduction of the discriminative segmentation module greatly enhances the processing speed of the fine-grained super-resolution tasks for key regions.

**Table 3.** The experimental results of the processing speed comparison

| Group | SGERN | BLSRN | Introduce the discriminative module | Processing speed (FPS) |
|---|---|---|---|---|
| Bicubic | √ | | | 4.57 |
| SRCNN | | √ | | **17.40** |
| FSRCNN | √ | √ | √ | 10.32 |

## 5. Conclusion

This paper designs a super-resolution method based on discriminative module and simple gating module from the perspectives of regional processing and network architecture optimization. The discriminative module enables the joint application of the target recognition network and the super-resolution network, providing new insights for super-resolution tasks. Additionally, the introduction of the simple gating module and reparameterization module strengthens the network's ability to extract details, and combined with the discriminative module, the processing speed is improved. Experimental comparisons with other mainstream super-resolution networks show that the proposed method achieves the best results in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) for image quality. In subjective qualitative results, it also provides the best detail restoration. Efficiency experiments verify that the proposed method meets the real-time processing requirements for tower construction.

## References

[1] Wang G, Ding P, Huang C, et al. A novel lifting point location optimization method of transmission line tower based on improved grey wolf optimizer[J]. Scientific Reports, 2023, 13(1): 21914.

[2] Zhou L, Zhang L, Konz N. Computer vision techniques in manufacturing[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 53(1): 105-117.

[3] Varghese R, Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness[C]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE, 2024: 1-6.

[4] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[5] Chan E R, Lin C Z, Chan M A, et al. Efficient geometry-aware 3d generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16123-16133.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202506_11(6).0005

[6] Navidan H, Moshiri P F, Nabati M, et al. Generative Adversarial Networks (GANs) in networking: A comprehensive survey & evaluation[J]. Computer Networks, 2021, 194: 108149.

[7] De Souza V L T, Marques B A D, Batagelo H C, et al. A review on generative adversarial networks for image generation[J]. Computers & Graphics, 2023, 114: 13-25.

[8] Wang H, Chen X, Ni B, et al. Omni aggregation networks for lightweight image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22378-22387.

[9] Chen L, Chu X, Zhang X, et al. Simple baselines for image restoration[C]. European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 17-33.

[10] Ding X, Zhang X, Ma N, et al. Repvgg: Making vgg-style convnets great again[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13733-13742.

[11] Sun L, Pan J, Tang J. Shufflemixer: An efficient convnet for image super-resolution[J]. Advances in Neural Information Processing Systems, 2022, 35: 17314-17326.

[12] Feng X, Li X, Li J. Multi-scale fractal residual network for image super-resolution[J]. Applied Intelligence, 2021, 51(4): 1845-1856.

[13] Meng G, Wu Y, Li S, et al. Video Super-Resolution with Long-Term Self-Exemplars[J]. arXiv preprint arXiv:2106.12778, 2021.