

Detection-oriented diffusion super-resolution for autonomous remote sensing perception

Mingming Xie¹, Hongwei Li¹, Mengni Xu¹, Chaojie Zhang¹, Chunhui Wang¹

¹ Zhejiang University, Hangzhou, China

Corresponding author: Chunhui Wang, hytgwch@zju.edu.cn

Low-resolution remote sensing imagery often degrades downstream object detection by blurring object boundaries and suppressing class-discriminative textures, particularly for small and densely distributed targets under sensing and transmission constraints. Although Remote Sensing Image Super-Resolution (RSISR) has achieved substantial progress, most existing methods are optimized primarily for visual reconstruction fidelity. Consequently, higher PSNR or SSIM does not necessarily translate into better downstream detection performance. To address this mismatch, we develop a detection-oriented super-resolution framework built upon a residual diffusion-based super-resolution branch and guided by downstream detection supervision. A frozen YOLOv11 detector is introduced during training to provide task-level guidance, encouraging the super-resolution model to recover structures relevant to object localization and classification. In addition, a Learnable Canny Edge Module (LCEM) is incorporated to enhance contour-sensitive structures that are critical for aerial object detection. Experiments on the DOTA benchmark show that the proposed framework improves mAP50 from 0.5655 for low resolution inputs to 0.5899, while maintaining competitive reconstruction fidelity. Ablation studies further indicate that detector-guided supervision is the main source of the mAP50 improvement, while LCEM provides complementary structural refinement. These results suggest that, for autonomous remote sensing perception, image enhancement should be optimized and evaluated as part of the downstream sensing pipeline rather than as an isolated visual restoration task.

Keywords: Aerial object detection; autonomous remote sensing perception; detection-oriented super-resolution; diffusion models; space AI; task-aware image enhancement

1. INTRODUCTION

Satellite remote sensing is undergoing a transition from human visual interpretation to automated machine perception. Earth observation imagery is increasingly used as the input to object detection, semantic segmentation, and change analysis, which are essential components of intelligent remote sensing systems [1, 2, 3]. In this setting, the value of an image is not determined only by its visual clarity, but also by whether it provides reliable information for downstream machine perception.

Spatial resolution is a critical factor in this transition. In aerial scenes, targets such as ships, aircraft, storage tanks, and small vehicles often occupy only a few pixels and appear in cluttered backgrounds. Their recognition depends heavily on fine boundary structures, local textures, and weak geometric cues. When image resolution decreases, these fragile cues are easily blurred or lost, making it difficult for detectors to distinguish foreground targets from complex backgrounds [2, 4].

Although modern satellite sensors continue to improve, low-resolution imagery remains common in practical remote sensing applications. In practical applications, high-resolution image acquisition is still constrained by imaging conditions, sensor cost, transmission bandwidth, and onboard resource limitations. Algorithmic image enhancement, especially Remote Sensing Image Super-Resolution (RSISR), therefore remains an important and flexible solution for improving the usability of remote sensing data [1, 5].

© International Telecommunication Union, 2026

Some rights reserved. © ⓘ Ⓢ Ⓜ

This work is available under the CC BY-NC-ND 3.0 IGO license: <https://creativecommons.org/licenses/by-nc-nd/3.0/igo/>.

More information regarding the license and suggested citation, additional permissions and disclaimers is available at:

<https://www.itu.int/en/journal/i-fet/Pages/default.aspx>

Recent image and remote sensing super-resolution methods have achieved substantial progress in visual reconstruction. Transformer-based and diffusion-based models can generate sharper and more realistic images than classical interpolation or earlier deep networks [6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. However, most existing methods are still designed as image restoration models. They are usually optimized for reconstruction-oriented metrics such as PSNR and SSIM, implicitly assuming that higher visual fidelity will lead to better downstream perception.

This assumption does not always hold for autonomous remote sensing perception. A super-resolution model may improve visual appearance while failing to restore the contours, target-background transitions, and category-discriminative structures required by an object detector. In some cases, visually plausible textures or over-smoothed boundaries may even reduce detection stability. Previous studies have observed this mismatch between reconstruction quality and detection performance [2, 4, 16], but it remains insufficiently explored in diffusion-based RSISR.

To address this problem, we develop a detection-oriented super-resolution framework based on a residual diffusion-based super-resolution backbone. The framework introduces task-aware optimization through a frozen YOLOv11 detector, whose classification and localization losses are used to guide the super-resolution branch toward detection-relevant structures. We further incorporate a learnable canny edge module (LCEM) to enhance contour-sensitive features, which are important for small and densely distributed aerial targets. By combining detection-level supervision with structural feature enhancement, the proposed framework aims to improve the utility of super-resolved images for downstream object detection.

The main contributions of this paper are summarized as follows:

1. We propose a detection-oriented training strategy for RSISR that extends the optimization objective from visual fidelity alone to downstream object detection utility.
2. We design a detection-guided optimization strategy based on a frozen YOLOv11 detector, which provides stable task-level supervision for the super-resolution network.
3. We introduce a learnable canny edge module to enhance contour-sensitive representations and complement detector-guided supervision.
4. We conduct experiments on the DOTA benchmark to analyze the relationship between reconstruction fidelity and detection performance, and to evaluate the effect of task-aware optimization on target detectability.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed methodology. Section 4 reports the experimental results and ablation studies. Section 5 discusses the implications and limitations of the proposed framework. Section 6 concludes the paper.

2. RELATED WORK

2.1 Remote sensing image super-resolution

Remote Sensing Image Super-Resolution (RSISR) has evolved from classical interpolation methods to data-driven deep learning approaches [1, 5]. Traditional reconstruction methods struggle to recover high-frequency structural details when confronted with complex, real-world degradations such as atmospheric blur and sensor noise. Deep learning has advanced this domain by learning robust nonlinear mappings from Low-Resolution (LR) to High-Resolution (HR) space. Representative modern architectures include SwinIR, Real-ESRGAN, HAT, and HAUNet [6, 7, 8, 9], which progressively improved perceptual realism and reconstruction accuracy through transformer-based or hybrid feature modeling.

Despite these advances, most RSISR methods remain image-oriented. These models are predominantly trained to minimize pixel-wise or perceptual discrepancies, treating super-resolution as an isolated end goal. Consequently, they rarely optimize the super-resolved features for downstream analytical tasks, leaving a gap between visual fidelity and downstream task utility.

2.2 Diffusion models for image restoration and super-resolution

Diffusion models have recently become a powerful class of generative models and have shown strong performance in image restoration and super-resolution [10, 11, 12]. In image super-resolution, ResShift and SinSR improve diffusion efficiency by reducing the sampling burden, while remote-sensing-oriented methods such as FastDiffSR and EDiffSR adapt diffusion models to RSISR through efficient sampling and conditional modeling [12, 13, 14, 15]. Recent studies have expanded diffusion-based RSISR along several directions, including semantic guidance, domain-specific attention, and continuous-scale reconstruction. For instance, ASDDPM introduces semantic-enhanced diffusion for long-range contextual modeling [17],

EVADM designs variance-aware attention for crop-field aerial image super-resolution [18], and NeurOp-Diff explores neural-operator diffusion for continuous-scale RSISR [19]. These studies demonstrate the rapid progress of diffusion-based RSISR and show that recent work has begun to incorporate semantic guidance, domain-specific modeling, and flexible-scale reconstruction into the diffusion framework.

However, diffusion-based RSISR still faces two major challenges. First, the iterative de-noising process introduces substantial computational overhead, which limits its use in large-scale or resource-constrained remote sensing pipelines. Existing acceleration strategies have partially improved inference efficiency, but computational cost remains an important consideration for practical remote sensing applications [12, 13, 14, 20]. Second, the optimization and evaluation of diffusion-based RSISR remain largely reconstruction-oriented. Although recent studies have begun to incorporate semantic information, domain-specific modeling, or downstream utility into diffusion-based RSISR [17, 18, 19], explicit detector-level supervision is still relatively under-explored. Metrics such as SSIM and LPIPS [21, 22] mainly capture structural or perceptual fidelity, but they do not directly reflect whether the super-resolved images preserve object boundaries and target-discriminative structures required by downstream detectors. This limitation motivates the detection-oriented optimization adopted in this work.

2.3 Task-oriented super-resolution for downstream perception

Task-oriented super-resolution shifts the optimization objective from human-centric visual fidelity to machine-centric perception utility [2, 4, 16]. In remote sensing, early empirical studies by Shermeyer and Van Etten demonstrated that the relationship between SR reconstruction quality and downstream detection performance is highly non-linear and class-dependent [4]. Subsequent joint pipelines further corroborated that explicitly coupling SR with detection feedback can preserve critical local cues, such as contours and semantic boundaries, that generic SR algorithms typically discard [2, 16].

Recent studies have further strengthened this task-oriented view. In general vision, Kim et al. proposed SR4IR, where a task-driven perceptual loss transfers task-specific knowledge from downstream recognition networks to the SR model, enabling the generated images to better support semantic segmentation, object detection, and image classification [23]. In remote sensing, Liu et al. embedded an SR perceptual branch into a small-object detection framework and associated

SR and detection through joint training, showing that SR-related supervision can improve the detectability of small aerial targets [24]. Different from such detector-integrated designs, the present work keeps the downstream detector frozen and evaluates whether the super-resolved image itself improves detection under a controlled fixed-detector protocol. This line of work is particularly relevant to aerial imagery, where targets are often small, densely distributed, and weakly contrasted. Despite recent progress in both diffusion-based RSISR and task-oriented SR, these two directions have largely developed along separate lines. Existing task-aware SR or SR-assisted detection methods often rely on detector-integrated architectures, while many diffusion-based RSISR methods are still mainly designed for image reconstruction. Directly incorporating detector-level supervision into a diffusion-based remote sensing super-resolution framework therefore remains relatively under-explored. This gap motivates the detection-oriented formulation adopted in this work, where detector feedback is explicitly incorporated into diffusion-based RSISR training under a frozen detector setting.

2.4 Remote sensing object detection under low-resolution conditions

Aerial object detection differs from natural image detection due to extreme scale variations, arbitrary target orientations, and highly cluttered backgrounds, as evidenced by benchmarks like DOTA [3]. Under low-resolution conditions, the limited pixel support of each target weakens local feature representations. Blurred boundaries conflate targets with background noise, while the suppression of category-specific textures exacerbates semantic confusion among visually similar classes.

Modern single-stage detectors, including the YOLO family, are sensitive to such resolution degradation because their predictions rely on robust multiscale features [25]. Consequently, generic image upsampling or purely fidelity-driven enhancement may be insufficient for preserving detector-relevant structures. These challenges motivate super-resolution methods that preserve structural and geometric cues relevant to downstream detectors.

3. METHODOLOGY

3.1 Problem formulation and overall framework

The objective of this work is to improve the usefulness of remote sensing image super-resolution for downstream

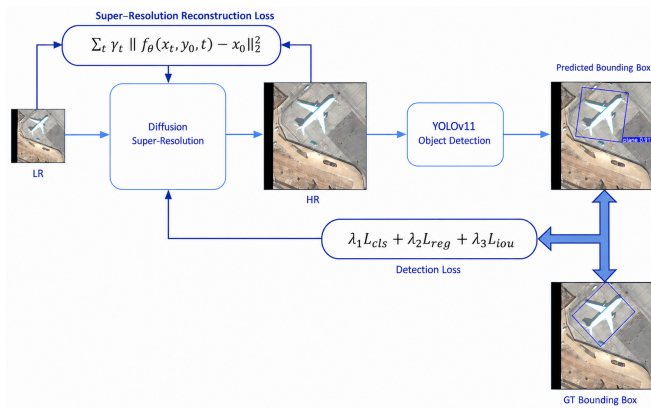


Figure 1 – Problem formulation of the proposed detection-oriented super-resolution framework.

object detection. Given a low-resolution input image x_{lr} and its corresponding high-resolution reference x_{hr} , a super-resolution branch G_θ is used to generate an enhanced image:

$$\hat{x}_{sr} = G_\theta(x_{lr}), \quad (1)$$

where θ denotes the learnable parameters of the super-resolution branch. Conventional super-resolution methods mainly optimize the generated image according to reconstruction-oriented objectives, such as pixel-level similarity or perceptual fidelity [1, 6, 7, 8]. However, in autonomous remote sensing perception, the final value of super-resolution is not limited to visual quality. The generated image should also preserve structures that are useful for downstream object localization and classification [2, 4, 16].

As illustrated in Fig. 1 and further described in the following subsections, the proposed framework consists of four main components: a residual diffusion-based super-resolution branch, a learnable canny edge module (LCEM), a frozen YOLOv11 detector for detection-guided supervision, and a skip-based inference strategy for improving computational efficiency. The residual diffusion branch serves as the basic image enhancement backbone. A Swin-UNet structure is adopted as its prediction network to model multi-scale spatial structures and local details in remote sensing imagery [6]. LCEM is inserted into the bottleneck feature layer of the super-resolution branch to strengthen contour-sensitive representations before image reconstruction.

During training, the super-resolved image $\hat{x}_{sr}$ is fed into a frozen YOLOv11 detector [25]. The detector provides task-level supervision through detection-related losses, while its own parameters remain fixed. The gradients from the detection loss are backpropagated through the detector to update only the super-resolution branch and LCEM. Therefore, the generated image is optimized not only for reconstruction fidelity but also for downstream detection utility.

During inference, the detection-guided loss is no longer used. The trained diffusion super-resolution branch directly generates enhanced images from low resolution inputs, and the same detector is used for downstream evaluation. To reduce the computational cost of diffusion sampling, a skip-based inference strategy is applied inside the reverse diffusion process. This strategy reuses intermediate features from selected de-noising steps to reduce redundant computation in the Swin-UNet prediction network. In summary, the residual diffusion branch provides the reconstruction foundation, LCEM enhances structure-aware representations, the frozen detector provides task-aware supervision during training, and the skip-based strategy improves inference efficiency during deployment.

3.2 Residual diffusion super-resolution branch

The proposed framework adopts a residual diffusion-based super-resolution branch as the image enhancement backbone. The purpose of this branch is to provide a reliable reconstruction foundation for detection-oriented optimization. Different from conventional diffusion models that generate images from pure Gaussian noise [10, 11], residual diffusion constructs a transition process between the low-resolution observation and the high-resolution target [12]. This design is suitable for super-resolution because the low-resolution image already contains the global scene layout, while the missing information mainly lies in high frequency residual details, including object contours, local textures, and target-background transitions.

Before entering the diffusion process, the low-resolution image is first upsampled to the same spatial size as the high-resolution reference. Let y_0 denote the upsampled low-resolution image and x_0 denote the high-resolution target. For clarity, the residual transition between them can be conceptually expressed as:

$$x_t = (1 - \eta_t)x_0 + \eta_t y_0 + \sigma_t \epsilon, \quad (2)$$

where x_t denotes the intermediate state at timestep t , η_t controls the transition degree from the high-resolution target to the low-resolution observation, σ_t denotes the noise scale, and ϵ is Gaussian noise. This expression is used to illustrate the intuition of residual diffusion: instead of corrupting the image toward pure noise, the diffusion trajectory is constructed between the HR target and the LR observation. In the reverse process, the model progressively recovers the missing residual information from the low-resolution condition and generates the super-resolved image.

In this work, the residual diffusion branch is used as an image reconstruction backbone rather than as an isolated contribution. A Swin-UNet-based prediction network is

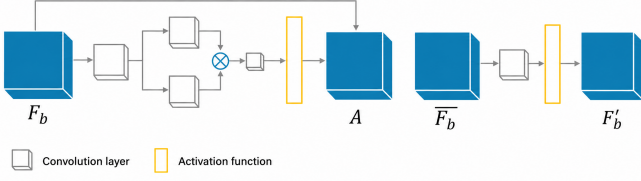


Figure 2 – Learnable canny edge module (LCEM).

adopted within the diffusion branch. Its encoder-decoder structure and skip connections allow the model to combine high-level contextual features with low-level spatial details, which is important for remote sensing scenes containing small objects, complex backgrounds, and large scale variations. The output of this branch is further constrained by the LCEM-enhanced feature representation and the detection-guided supervision described in the following subsections.

3.3 Learnable canny edge module

Remote sensing object detection is highly sensitive to boundary and contour information, especially for small and densely distributed targets. When the input resolution decreases, object boundaries and target-background transitions may become blurred or even disappear, making it difficult for the detector to distinguish small objects from complex backgrounds. To address this issue, we introduce a learnable canny edge module (LCEM) into the super-resolution branch.

LCEM is inserted into the bottleneck feature layer of the Swin-UNet-based super-resolution branch before the final reconstruction stage, as illustrated in Fig. 2. This position allows the module to operate on compact high-level features while still preserving spatial information related to object contours. Different from a conventional canny operator used as an external post-processing step [26], LCEM is embedded inside the network and trained jointly with the super-resolution branch. Therefore, it provides a learnable structural prior rather than a fixed edge map.

Given a bottleneck feature map F_b , LCEM first estimates horizontal and vertical gradient responses using learnable convolutional operations:

$$G_x = \text{Conv}_x(F_b), \quad G_y = \text{Conv}_y(F_b), \quad (3)$$

where G_x and G_y denote the horizontal and vertical gradient-like responses, respectively. The gradient magnitude map is then computed as:

$$M = \sqrt{G_x^2 + G_y^2 + \varepsilon}, \quad (4)$$

where ε is used for numerical stability. The gradient magnitude is transformed into an edge attention map:

$$A = \sigma(\text{Conv}_{1 \times 1}(M)), \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The attention map is further used to recalibrate the bottleneck feature through residual modulation:

$$\bar{F}_b = F_b + \gamma(F_b \odot A), \quad (6)$$

where $\odot$ denotes element-wise multiplication and γ is a learnable scaling factor.

Through this design, LCEM encourages the super-resolution branch to preserve contour-sensitive representations during image reconstruction. During training, it helps the model respond more effectively to detection-guided supervision. During inference, it remains part of the super-resolution branch and continues to enhance structure-aware features in the generated image.

3.4 Detection-guided joint optimization

Although super-resolution methods can improve visual fidelity, higher reconstruction quality does not necessarily lead to better object detection performance. Some visually plausible details may be irrelevant or even harmful to downstream detectors, while some detector-sensitive structures may not be sufficiently emphasized by reconstruction losses alone [2, 4, 16]. Therefore, we introduce detection-guided joint optimization to align the super-resolution process with downstream detection utility.

During training, the super-resolved image $\hat{x}_{sr}$ is sent to a YOLOv11 detector D [25]. The detector is frozen throughout the optimization process and serves only as a task-level supervisor. Its parameters are not updated. The detection loss is computed using the detection predictions and the ground-truth annotations, and the resulting gradients are backpropagated through the detector to the super-resolution branch and LCEM. In this way, the super-resolution model is encouraged to generate images that are not only visually consistent with the HR reference but also more favorable for object localization and classification.

The detection-guided loss is formulated as:

$$\mathcal{L}_{det} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{reg} + \lambda_3 \mathcal{L}_{iou}, \quad (7)$$

where $\mathcal{L}_{cls}$ denotes the classification loss, $\mathcal{L}_{reg}$ denotes the bounding box regression loss, and $\mathcal{L}_{iou}$ denotes the IoU-related localization loss. The coefficients λ_1 , λ_2 , and λ_3 are used to balance the contributions of different detection-related terms.

The overall training objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{SR}(\hat{x}_{sr}, x_{hr}) + \mathcal{L}_{det}(D(\hat{x}_{sr}), y), \quad (8)$$

where $\mathcal{L}_{SR}$ denotes the reconstruction loss between the generated super-resolved image and the HR reference, and $\mathcal{L}_{det}$ denotes the detection-guided loss computed from the frozen detector.

3.5 Skip-based efficient inference

Diffusion-based super-resolution usually requires multiple reverse de-noising steps, which increases inference cost [10, 11, 12]. This issue is particularly important in remote sensing applications, where large-scale imagery must often be processed under limited computational resources. To reduce the computational burden, we adopt a skip-based inference strategy inspired by feature caching and reuse mechanisms [20].

The strategy is based on the observation that the noise prediction network may produce similar intermediate representations across consecutive diffusion steps. Recomputing all feature transformations at every de-noising step can therefore introduce redundant computation. In the Swin-UNet prediction network, the forward path contains both a main branch and skip connections. During reverse diffusion, intermediate features from selected de-noising steps are cached and reused in subsequent steps, so that part of the repeated computation in the prediction network can be replaced by feature retrieval.

This skip-based inference strategy is used only during inference. It does not change the training objective, the LCEM design, or the frozen detector setting. Its role is to improve computational efficiency while preserving the detection-oriented enhancement learned during training. Therefore, the proposed method separates the functions of different components: the residual diffusion branch reconstructs the enhanced image, LCEM strengthens contour-aware representations, detection-guided supervision aligns super-resolution with downstream object detection, and skip-based inference reduces redundant computation during deployment.

4. EXPERIMENTS

4.1 Experimental settings and implementation details

The proposed framework is evaluated on the DOTA v1.0 benchmark to assess its effectiveness for detection-oriented remote sensing super-resolution. DOTA contains complex aerial scenes with 15 categories of oriented objects, including planes, ships, small vehicles, large vehicles, bridges, storage tanks, harbors

and sports fields [3]. Since this work focuses on downstream object detection rather than generic image reconstruction, the experiments are conducted under a fixed detection pipeline to isolate the influence of different super-resolution strategies.

For LR-HR pair generation, the original DOTA images are first cropped into 1024×1024 HR patches. The corresponding LR inputs are then generated by bicubic downsampling under the fixed super-resolution scale setting used in this work. Before being fed into the residual diffusion branch, the LR images are upsampled to the HR spatial size, so that the LR condition and HR reference share the same spatial resolution during diffusion-based reconstruction. The training, validation and testing splits follow the official DOTA protocol, and the same preprocessing protocol is used for all compared methods. The compared inputs include LR images, bicubic upsampled images [5], SwinIR super-resolved images [6], and the proposed detection-oriented super-resolution results. All super-resolved images are fed into the same YOLOv11 detector, so that the observed detection differences mainly come from the super-resolution stage rather than changes in the detector.

Downstream detection performance is evaluated using mean Average Precision at an intersection-over-union threshold of 0.50 (mAP50), which reflects whether the enhanced images improve object localization and classification under the fixed detector. In addition, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) are reported as auxiliary reconstruction metrics. Higher PSNR, SSIM and mAP50 indicate better performance, whereas lower LPIPS indicates better perceptual similarity. In the tables, $\uparrow$ and $\downarrow$ denote that higher and lower values are better, respectively.

The model is implemented in PyTorch and trained on four NVIDIA RTX 3090 GPUs. The Adam optimizer is used with an initial learning rate of 0.0002 and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. During detection-oriented training, the YOLOv11 detector is first trained on the DOTA training set and then kept frozen [25]. The detector parameters are not updated during SR training; instead, the classification, localization, and IoU-related losses are backpropagated through the frozen detector to update only the super-resolution branch and LCEM. In implementation, the detection-related loss weights are set to $\lambda_1 = \lambda_2 = 0.1$ and $\lambda_3 = 0.05$, so that the reconstruction loss remains the dominant term while the detection losses serve as auxiliary task-level constraints.

Table 1 – Image-level reconstruction metrics and detection performance on the DOTA dataset.

Input	PSNR (dB) $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	mAP50 $\uparrow$
LR	33.1659	0.6182	0.2820	0.5655
bicubic [5]	34.1563	0.7331	0.2557	<u>0.5896</u>
SwinIR [6]	35.7844	0.7629	0.1956	0.5690
Ours	35.8972	0.7754	0.1977	0.5899
HR reference	–	–	–	0.6526

4.2 Reconstruction-detection misalignment analysis

Before analyzing category-level detection results, we first examine whether reconstruction quality is always consistent with downstream detection performance. Table 1 reports both image-level reconstruction metrics and mAP50 on the DOTA dataset. All inputs are evaluated using the same YOLOv11 detector, so the comparison reflects the impact of different super-resolution strategies under a fixed downstream detection pipeline.

A noticeable mismatch can be observed between reconstruction fidelity and detection performance. SwinIR achieves strong reconstruction metrics, especially in PSNR, SSIM and LPIPS, but its mAP50 is lower than that of bicubic upsampling. This indicates that improving image-level fidelity does not necessarily preserve detector-relevant structures. The proposed method improves mAP50 from 0.5655 on LR inputs to 0.5899, while maintaining competitive reconstruction quality. Although the gain over bicubic upsampling is marginal, the comparison remains informative: SwinIR achieves better reconstruction metrics than bicubic but lower detection accuracy, indicating that reconstruction fidelity and detection performance are not strictly aligned. The HR reference achieves an mAP50 of 0.6526, which remains higher than all low-resolution and super-resolved inputs, suggesting that SR can improve the usability of degraded images but cannot fully recover information lost during resolution degradation.

To further understand where the overall detection gain comes from, we decompose the mAP50 result into category-level AP in the following section.

4.3 Category-level detection performance

To further analyze how detection-oriented super-resolution affects different aerial object categories, Table 2 reports the category-level Average Precision (AP) results on DOTA. Combining the overall mAP50 results in Table 1 with the category-level AP results in Table 2, the detection gain of the proposed method should be interpreted as category-dependent rather

than uniformly distributed across all object types. As shown in Table 1, the proposed method improves mAP50 from 0.5655 for LR inputs to 0.5899, while the gain over bicubic is marginal. Therefore, the category-level analysis is used to explain which types of objects benefit from detection-oriented super-resolution and which categories remain challenging.

Compared with LR inputs, the proposed method improves AP in most categories, especially those whose detection depends on local contours, compact shapes and target-background transitions. For example, small vehicle AP increases from 0.4082 to 0.5653, large vehicle AP increases from 0.5167 to 0.6772, basketball court AP increases from 0.5111 to 0.6092, harbor AP increases from 0.5773 to 0.6433, bridge AP increases from 0.4051 to 0.4468, and soccer-ball field AP increases from 0.4738 to 0.5314. The ship AP also improves substantially over LR and bicubic inputs, although SwinIR obtains a slightly higher AP for this category. These results indicate that detection-guided optimization and LCEM are particularly useful when the lost high frequency information is closely related to object boundaries, compact geometric shapes and target-background separation.

The improvements, however, are not uniform across all categories. For planes, ships, large vehicles and swimming pools, SwinIR achieves slightly higher or comparable AP values, suggesting that some categories may still benefit from reconstruction-oriented texture recovery. For baseball diamonds and ground track fields, the proposed method does not outperform LR or bicubic inputs. These categories contain larger scene-level structures and may rely more on global layout and contextual consistency than on local edge enhancement alone. This indicates that detection-oriented super-resolution does not simply improve every category by sharpening local details. The helicopter category shows the most obvious degradation. Its AP decreases from 0.4667 for LR inputs to 0.1788 with the proposed method, while the HR reference also achieves only 0.3379. This suggests that the weak performance on helicopters is not solely caused by the proposed super-resolution branch, but is also related to category-specific detection difficulty under the current detector. Helicopters usually appear as small and sparse targets with large appearance variation, and their detection can be sensitive to detector bias, limited category samples and local structural distortions introduced during enhancement. Therefore, the applicability boundary of the proposed method should be recognized: it is more reliable for categories with stable contour and boundary cues, while extremely small, sparse, or visually ambiguous categories may require additional category-aware constraints or stronger detector adaptation.

Table 2 – Category-level Average Precision (AP) comparison on the DOTA dataset.

Category	LR	bicubic [5]	SwinIR [6]	Ours	HR reference
Plane	0.8264	0.8377	0.8425	<u>0.8345</u>	0.8845
Ship	0.5843	0.6174	0.7371	<u>0.7233</u>	0.7980
Storage tank	0.6332	<u>0.6477</u>	0.6282	0.6638	0.7041
Baseball diamond	<u>0.6258</u>	0.6422	0.5835	0.5494	0.6665
Tennis court	<u>0.8655</u>	<u>0.8697</u>	0.8382	0.8717	0.9265
Basketball court	0.5111	<u>0.5513</u>	0.5601	0.6092	0.6627
Ground track field	<u>0.4720</u>	0.5263	0.4016	0.4043	0.4688
Harbor	<u>0.5773</u>	0.6072	<u>0.6293</u>	0.6433	0.6847
Bridge	0.4051	0.4181	<u>0.4451</u>	0.4468	0.4975
Large vehicle	0.5167	0.6037	0.6880	<u>0.6772</u>	0.7455
Small vehicle	0.4082	0.3985	<u>0.4714</u>	0.5653	0.6305
Helicopter	0.4667	<u>0.4588</u>	0.1401	0.1788	0.3379
Roundabout	0.5563	0.5925	0.5017	<u>0.5697</u>	0.6163
Soccer-ball field	0.4738	<u>0.4920</u>	0.4682	0.5314	0.5902
Swimming pool	0.5595	<u>0.5800</u>	0.5998	<u>0.5815</u>	0.5751

4.4 Ablation studies

To isolate the contribution of each detection-oriented component, we conduct ablation experiments on DOTA by progressively introducing LCEM and the detection-guided loss. Table 3 reports both reconstruction metrics and detection performance for different model variants.

The ablation results reveal a clear trade-off between reconstruction-oriented metrics and downstream detection performance. The reconstruction-only baseline, Group 1, achieves the highest PSNR of 37.7794 and the lowest LPIPS of 0.1284, but obtains the lowest mAP50 of 0.5183. This indicates that optimizing image fidelity alone may produce visually consistent or perceptually close results, but it does not necessarily preserve the structures most useful for object localization and classification. Therefore, the best reconstruction metric values do not directly correspond to the best detection performance.

LCEM and the detection-guided loss play different roles in this framework. LCEM mainly serves as a structural refinement module. When LCEM is added alone in Group 2, SSIM increases from 0.7681 to 0.8023 and mAP50 slightly improves from 0.5183 to 0.5264. This suggests that edge-aware feature enhancement can improve structural consistency and provide limited detection benefits. However, without task-level supervision from the detector, LCEM alone is insufficient to strongly align the super-resolved output with downstream detection requirements.

In contrast, the detection-guided loss provides direct task-level gradients from the frozen YOLOv11 detector. When the detection-guided loss is introduced alone in

Group 3, mAP50 increases substantially from 0.5183 to 0.5807, although PSNR, SSIM and LPIPS become worse than those of the reconstruction-only baseline. This result shows that detector-guided supervision is the main source of downstream detection improvement in the proposed framework. It encourages the super-resolution branch to recover detector-relevant structures, even when this optimization does not lead to the best image-level reconstruction metrics.

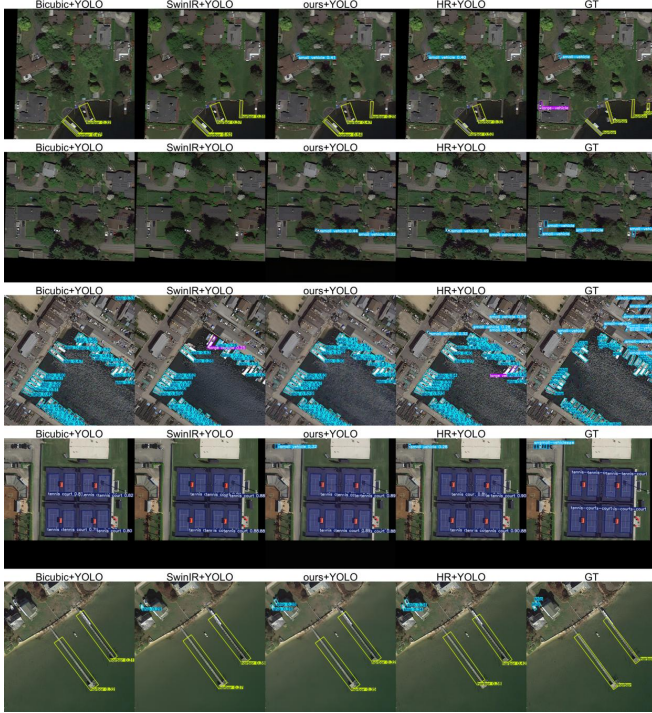
When LCEM and the detection-guided loss are combined in Group 4, the model achieves the best mAP50 of 0.5899. Compared with Group 3, LCEM further improves mAP50 from 0.5807 to 0.5899, indicating that contour-sensitive structural refinement complements the task-level supervision provided by the frozen detector. Meanwhile, Group 4 also recovers better PSNR and SSIM than Group 3, suggesting that LCEM can help improve structural quality while supporting detection-oriented optimization. Overall, the ablation study shows that the proposed framework does not simply maximize reconstruction fidelity. Instead, detection-guided loss determines the main task-aware optimization direction, while LCEM provides complementary edge-aware structural support.

4.5 Qualitative analysis

Fig. 3 provides qualitative examples of how different super-resolution strategies affect object detection under low-resolution conditions. Bicubic upsampling can recover the image size but still suffers from blurred boundaries and aliasing artefacts. SwinIR improves visual sharpness in some regions, but the super-resolved textures do not always correspond to structures that are useful for object localization and classification. In

Table 3 – Ablation study of detection-oriented components on the DOTA dataset.

Setting	LCEM	Detection-guided loss	PSNR (dB) $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	mAP50 $\uparrow$
Group 1	No	No	37.7794	0.7681	0.1284	0.5183
Group 2	Yes	No	<u>36.6545</u>	0.8023	<u>0.1592</u>	0.5264
Group 3	No	Yes	<u>33.2184</u>	0.7220	<u>0.2113</u>	<u>0.5807</u>
Group 4 (Ours)	Yes	Yes	35.8972	<u>0.7754</u>	0.1977	0.5899

**Figure 3** – Qualitative detection comparison on DOTA across different methods.

comparison, the proposed method tends to preserve clearer object contours and target-background transitions in the selected examples, which helps the detector produce more stable localization results.

These qualitative observations are consistent with the quantitative detection results in tables 1 and 2, as well as the ablation trend in Table 3.

5. DISCUSSION

The experimental results show that reconstruction fidelity alone is not a reliable indicator of downstream detection performance in remote sensing super-resolution. Although SwinIR obtains strong PSNR, SSIM and LPIPS scores, its mAP50 is lower than that of bicubic. This suggests that visually plausible super-resolution may not preserve the local structures required by an object detector, especially for small and densely distributed aerial targets.

The ablation study further clarifies the role of the proposed components. Detection-guided optimization brings the largest improvement in mAP50, showing that detector feedback provides useful training signals for recovering task-relevant structures. LCCEM offers a smaller but complementary gain by enhancing edge-sensitive features in the SR branch. These results support the central premise of this work: super-resolution for autonomous remote sensing should be assessed not only by visual fidelity, but also by its effect on downstream perception.

Several limitations should be acknowledged. First, the task alignment is learned using a frozen YOLOv11 detector; its generalization to other detector architectures remains to be investigated. Second, despite the use of the skip-based inference strategy, diffusion-based super-resolution remains more computationally demanding than interpolation or lightweight CNN-based SR, which may limit direct deployment on resource-constrained platforms. Future work will therefore consider cross-detector evaluation, more efficient super-resolution backbones, and extension to other remote sensing modalities such as SAR and hyperspectral imagery.

6. CONCLUSION

This paper presents a detection-oriented framework for residual diffusion-based remote sensing image super-resolution. The proposed method integrates a frozen YOLOv11 detector into SR training and introduces LCCEM to enhance contour-sensitive features. Experiments on the DOTA benchmark show that reconstruction metrics and detection accuracy are not always consistent: methods with stronger visual fidelity do not necessarily yield better detection results. The ablation study further shows that detection-guided supervision is the main source of the mAP50 improvement, while LCCEM provides complementary structural enhancement. These findings highlight the importance of evaluating remote sensing image enhancement in the context of downstream perception tasks.

REFERENCES

- [1] Y. Qi, M. Lou, Y. Liu, et al. "Advancing image super-resolution techniques in remote sensing: A comprehensive survey". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 231 (2026), pp. 68–100.
- [2] Y. Wang, S. M. A. Bashir, M. Khan, et al. "Remote sensing image super-resolution and object detection: Benchmark and state of the art". In: *Expert Systems with Applications* 197 (2022), p. 116793.
- [3] G.-S. Xia, X. Bai, J. Ding, et al. "DOTA: A large-scale dataset for object detection in aerial images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 3974–3983.
- [4] J. Shermeyer and A. Van Etten. "The effects of super-resolution on object detection performance in satellite imagery". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019.
- [5] R. Keys. "Cubic convolution interpolation for digital image processing". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.6 (1981), pp. 1153–1160.
- [6] J. Liang, J. Cao, G. Sun, et al. "SwinIR: Image restoration using Swin Transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 1833–1844.
- [7] X. Wang, L. Xie, C. Dong, et al. "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 1905–1914.
- [8] X. Chen, X. Wang, J. Zhou, et al. "Activating more pixels in image super-resolution transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 22367–22377.
- [9] J. Wang, B. Wang, X. Wang, et al. "Hybrid attention-based U-shaped network for remote sensing image super-resolution". In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–15.
- [10] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 6840–6851.
- [11] C. Saharia, J. Ho, W. Chan, et al. "Image super-resolution via iterative refinement". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2022), pp. 4713–4726.
- [12] Z. Yue, J. Wang, and C. C. Loy. "ResShift: Efficient diffusion model for image super-resolution by residual shifting". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36. 2023, pp. 13294–13307.
- [13] Y. Wang, W. Yang, X. Chen, et al. "SinSR: Diffusion-based image super-resolution in a single step". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 25796–25805.
- [14] F. Meng, Y. Chen, H. Jing, et al. "A conditional diffusion model with fast sampling strategy for remote sensing image super-resolution". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024). doi: [10.1109/TGRS.2024.3458009](https://doi.org/10.1109/TGRS.2024.3458009).
- [15] Y. Xiao, Q. Yuan, K. Jiang, et al. "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2023), pp. 1–14.
- [16] J. Zhang, J. Lei, W. Xie, et al. "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–15.
- [17] J. Sui, X. Ma, X. Zhang, M. O. Pun, and H. Wu. "Adaptive semantic-enhanced denoising diffusion probabilistic model for remote sensing image super-resolution". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18 (2024), pp. 892–906.
- [18] X. Lu, J. Zhang, R. Yang, et al. "Effective variance attention-enhanced diffusion model for crop field aerial image super resolution". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 218 (2024), pp. 50–68.
- [19] Z. Xu, Y. Tang, B. Xu, and Q. Li. "NeurOp-Diff: Continuous remote sensing image super-resolution via neural operator diffusion". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025, pp. 12491–12501.
- [20] X. Ma, G. Fang, and X. Wang. "DeepCache: Accelerating diffusion models for free". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 15762–15772.
- [21] Z. Wang, A. C. Bovik, H. R. Sheikh, et al. "Image quality assessment: From error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.
- [22] R. Zhang, P. Isola, A. A. Efros, et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 586–595.
- [23] J. Kim, J. Oh, and K. M. Lee. "Beyond image super-resolution for image recognition with task-driven perceptual loss". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 2651–2661.
- [24] J. Liu, J. Zhang, Y. Ni, W. Chi, and Z. Qi. "Small-object detection in remote sensing images with super-resolution perception". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024), pp. 15721–15734.
- [25] G. Jocher, J. Qiu, and A. Chaurasia. *Ultralytics YOLO*. Version 11.0.0. 2024.
- [26] J. Canny. "A computational approach to edge detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1986), pp. 679–698.

AUTHORS



MINGMING XIE is the main author. She received the B.E. degree in aircraft control and information engineering from Sichuan University, Chengdu, China, in 2024. She is currently working towards a Ph.D. degree with the School of Aeronautics and Astronautics, Zhejiang University, Hangzhou, China. Her research interests include deep learning, computer vision and remote sensing, with a particular focus on remote sensing change detection and visual semantic understanding.



HONGWEI LI received his PhD degree in surveying and mapping science and technology from Tongji University in 2024. Following his doctoral studies, he joined the College of Information Science and Electronic Engineering at Zhejiang University, where he is currently conducting a postdoctoral research

program. His research interests span photogrammetry and remote sensing, computer vision, large-scale satellite-borne LiDAR point cloud processing, deep learning and intelligent spatial information analysis, edge-intelligent algorithms, multimodal data processing and UAV-based 3D reconstruction. Over the past few years, he has published more than 20 papers in international journals and conferences and holds five granted invention patents.



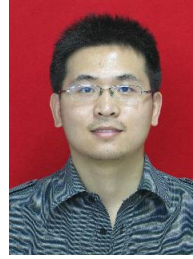
MENGNI XU received the B.E. degree in aerospace engineering from the School of Aeronautics and Astronautics, Zhejiang University, Hangzhou, China, in 2023 and an M.S. degree in aerospace engineering from the same university in 2026. Her research interests include remote

sensing vision and diffusion models, with a particular focus on remote sensing image generation and visual interpretation.



CHAOJIE ZHANG received his bachelor's degree in communication engineering from Zhejiang University in 2004 and his PhD degree in electronic science and technology from Zhejiang University in 2009. After completing his doctorate, he carried out a postdoctoral

fellowship at Zhejiang University from 2009 to 2012. In 2012, he joined Zhejiang University as a faculty member and was promoted to associate research fellow in 2016. He was a visiting scholar at Nanyang Technological University from 2012 to 2013. His research interests include software-defined radio technology, satellite communication technology and miniaturized TT&C transponder technology. He has published over 20 SCI/EI-indexed papers and holds eight granted invention patents.



CHUNHUI WANG received his bachelor's degree in communication engineering from Zhejiang University in 2004 and his PhD degree in electronic science and technology from Zhejiang University in 2009. After completing his doctorate, he carried out a postdoctoral

fellowship at Zhejiang University from 2009 to 2011. In 2011, he joined Zhejiang University as a faculty member and was promoted to associate research fellow in 2015. Over the past few years, he has been engaged in research on micro-satellite TT&C and power technology, as well as satellite formation measurement and communication technology. He has published several SCI/EI-indexed papers and holds four granted invention patents.