ARTICLE | Article in Press

Enhancing weld defect detection in low-resolution environments via selective aggregation transformer super-resolution

Seyed Hamid Safavi

Received: 9 September 2026 / Accepted: 19 November 2026
© The Author(s) 2026

Abstract
Reliable weld defect detection is crucial for automated industrial inspection, yet real-world conditions often yield low-resolution (LR) and degraded images that severely compromise the accuracy of object detection models. To bridge this gap, this study proposes the Selective Aggregation Transformer for Super-Resolution (SATSR) to reconstruct high-fidelity weld images prior to detection. We systematically evaluate the impact of various super-resolution (SR) techniques–including standard convolutional and GAN-based architectures–on a YOLOv11 detection pipeline under both synthetic and realistic degradation scenarios (up to $\times 4$ scale). Quantitative results demonstrate that integrating SR preprocessing aids in recovering degraded defect features, yielding absolute improvements of up to 0.03 in mAP-50 (e.g., from 0.49 to 0.52 under $\times 4$ bicubic degradation) and up to 0.02 in mAP-50-95 compared to directly detecting on low-resolution inputs. Furthermore, comparative evaluations across image quality, model complexity, and inference latency show that while heavier architectures like EDSR or GAN-based models achieve marginally higher detection metrics in isolated settings, SATSR delivers a highly competitive balance–requiring only $\sim 1\text{M}$ parameters and under 11 GFLOPs while suppressing over-smoothing and spurious artifacts. In challenging cross-domain conditions (training on synthetic SR and testing on real-world LR), qualitative and quantitative analyses reveal that preserving topological defect boundaries provides a more stable representation under severe domain shift. Ultimately, this study establishes that while lightweight SR mitigates resolution-induced performance drops, coupling image restoration with explicit domain-alignment strategies is imperative for deploying robust industrial inspection systems.

Keywords Weld defect detection · Super-resolution · YOLO

Introduction

Detecting weld defects is essential in industrial production, especially where safety and reliability matter most. Defects such as cracks, pores, lack of fusion, undercut, and slag inclusions can weaken welded joints and, if overlooked, may cause serious failures. Traditional inspection methods, such as manual checks, contact physical sensing, and X-ray-based non-destructive testing, have drawbacks, including operator dependence, high costs,

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



https://doi.org/10.1038/s41598-026-82193-3

Published online: 25 November 2026

radiation hazards, and limited scalability for mass production. In this context, optical vision systems complement physical sensing modalities by providing rapid, non-contact, and consistent surface defect detection. These operational advantages have led to the rapid rise of automated weld inspection frameworks powered by machine vision and deep learning architectures.

Deep learning, particularly the application of convolutional neural networks (CNNs), has greatly advanced the field of automated surface defect inspection. Two-stage models like Faster R-CNN[1] achieve high accuracy on X-ray images, but their complexity can make real-time deployment difficult. Consequently, the YOLO (You Only Look Once) family of one-stage detectors[2] has emerged as the dominant paradigm for industrial applications because of its optimal balance between processing speed and detection accuracy[3–5].

Recent studies in this domain have primarily focused on architecturally modifying these baseline models to address particular industrial challenges, such as extreme scale variations and complex background interference. For instance, the SFW-YOLO[6] model integrates a high-resolution P2 layer and a dynamic attention mechanism to improve the recognition of small-sized weld defects. Similarly, the YOLOv8-DPE[5] framework employs a Dilation-Wise Residual (DWR) module, while methods such as DASS-YOLO[7] and WSS-YOLO[4] incorporate dynamic snake convolutions to capture fine, irregular defect structures. In parallel, optimizing feature fusion pathways, loss functions, and structural representation learning has been extensively investigated to stabilize the localization of tiny, low-quality defect instances[3,8,9]. Table 1 summarizes recent progress in deep learning-based weld defect inspection, highlighting this primary focus on internal detector optimizations.

Despite these remarkable architectural advancements, existing methods predominantly assume that input images possess sufficient spatial resolution and clarity[3–7,9]. In practical industrial settings, however, image acquisition is frequently constrained by adverse conditions, sensor costs, and bandwidth limits, resulting in degraded imagery. To address degraded visual quality, existing research generally follows three distinct paradigms, each facing practical limitations: (1) *NDT-specific super-resolution methods*, such as wavelet-guided dynamic transformers[10], incorporate frequency-domain priors tailored for non-destructive testing. While effective at sharpening edges, fixed wavelet or physical priors often amplify high-frequency ringing artifacts when exposed to heavy sensor noise and compression blocking. (2) *Lightweight attention-enhanced detectors and SR-assisted pipelines*, which either embed dynamic attention inside compact detectors (e.g., SFW-YOLO[6], DASS-YOLO[7]) or pair restoration backbones with one-stage detectors (e.g., Mamba-based SR with YOLOv10[11]). However, modifying internal detector attention cannot physically recover fine defect structures (such as micro-pores) once they are lost to ×3 or ×4 downsampling, whereas coupling heavy transformer restorers introduces high latency unsuitable for real-time edge inspection. (3) *Domain-shift adaptation frameworks*, including cross-domain transformers[12] and source-free adaptation methods[13], which align latent feature distributions between synthetic and real domains.

Table 1 Recent studies on weld defect inspection using deep learning-based vision methods.

Ref.	Year	Welding process	Image type	Main algorithm	Dataset	Classes	Key result
3	2024	–	Visual	DsP-YOLO	PCB-DET/NEU-DET/GC10-DET	6/6/10	mAP@50 up to 95.8%
4	2024	–	Visual	WSS-YOLO	NEU-DET/GC10-DET	6/10	mAP@50 up to 82.3%
5	2025	TIG	Visual	YOLOv8-DPE	Own + NEU-DET	4/6	mAP@50 up to 99%
6	2025	TIG/MIG	Visual	SFW-YOLO	Own + PCB-DET	4/6	mAP up to 95%
7	2025	–	Visual	DASS-YOLO	NEU-DET	6	mAP@50 77.4
21	2019	–	Visual	MFN + DDN	NEU-DET	6	mAP 82.3
14	2024	MAG	Visual	YOLOv7	LoHi-WELD	4	mAP 69%
22	2025	GMAW	Visual	YOLOv8	multiple weld datasets	2–4	Precision up to 0.982
23	2025	TIG	Visual	ViT	TIG Aluminium 5083	6	Accuracy 99%
24	2019	–	X-ray	SegNet	GDXRay	2	F1 score 0.80
25	2020	–	X-ray	TL + MobileNet	GDXRay	5	Accuracy 97.69%
26	2021	–	X-ray	TL + VGG16/ResNet50	GDXRay	14	F1 score 0.98
27	2024	–	X-ray	Faster R-CNN	Own X-ray dataset	1	F1 score 0.49

Although feature alignment reduces domain discrepancy, it operates in the latent space and cannot reconstruct missing pixel-level spatial details in severely degraded radiographs.

The impact of raw image quality and the potential of task-driven restoration to recover lost visual fidelity before detection thus remain insufficiently explored. To meet these challenges, this study examines weld defect detection on the LoHi-WELD dataset[14], an industrial benchmark comprising both high-resolution (H-WELD) and low-resolution (L-WELD) subsets. Because LoHi-WELD includes two resolution levels of the same domain, it allows us to test whether image super-resolution can close the gap between low-cost acquisition and high-accuracy recognition.

A key contribution of this work is to treat single-image super-resolution (SISR) as a critical preprocessing step, rather than a universally helpful enhancement. Methodologically, rather than applying a general-purpose natural-image restoration backbone, we formulate the Selective Aggregation Transformer for Super-Resolution (SATSR) based on a task-driven spectral-decoupling principle tailored for low-contrast industrial weld imagery. Specifically, while foundational deformable attention mechanisms rely on local gradient offsets that suffer from stochastic variance and induce geometric boundary jitter under low signal-to-noise ratios, SATSR pairs deterministic shift-invariant spatial average pooling with localized 3×3 convolutional projections. This dual-branch design suppresses high-frequency stochastic noise while strictly preserving translation equivariance and sharp anisotropic edge gradients required for downstream bounding-box regression. We systematically compare several modern SISR models (SRCNN[15], EDSR[16], SRGAN[17], ESRGAN[18], and SwinIR[19]) against the developed SAT-based[20] super-resolution (SATSR) framework for weld defect images in terms of reconstruction quality, computational cost, and their effect on a YOLOv11-nano detector under bicubic and realistic degradation pipelines. In addition, by conducting cross-domain experiments involving high-resolution, synthetic low-resolution, and super-resolved images, we aim to show that robust industrial vision systems must jointly consider detector architecture, image quality, and the chosen restoration strategy.

Methods

Dataset

The LoHi-WELD dataset[14], a publicly available industrial benchmark for weld defect detection and classification, serves as the foundation for this study. Developed to address the scarcity of realistic, annotated weld imagery for deep learning research, the dataset comprises two distinct subsets acquired under different capture conditions. The low-resolution subset (L-WELD) consists of 2,000 weld bead images captured with a 640×480 sensor, while the high-resolution subset (H-WELD) includes 1,022 images captured with a 2048×1080 sensor. Overall, the dataset contains more than 22,000 annotated defect instances, positioning it among the largest publicly available weld-defect collections.

In contrast to paired super-resolution benchmarks, the low-resolution images in LoHi-WELD are not simply downsampled versions of the high-resolution images. The two subsets were acquired independently using different cameras, exposure conditions, and field-of-view settings. Consequently, the low-resolution (LR) and high-resolution (HR) images represent two related but non-aligned domains, rather than one-to-one pairs. This distinction is critical for fair evaluation, as it avoids the assumption that each LR image has a perfectly corresponding HR reference image. The dataset encompasses four defect categories: pore, deposit, discontinuity, and stain. Each image is annotated with axis-aligned bounding boxes for both the weld bead region and the defect instances.

The dual-resolution structure of LoHi-WELD is well-suited for investigating the practical impact of super-resolution in industrial inspection. The L-WELD subset represents a cost-efficient acquisition scenario, with smaller images that reduce storage, bandwidth, and computational demands. In contrast, the H-WELD subset offers greater visual detail and serves as a higher-quality reference for assessing the influence of resolution on defect

detection performance. This configuration establishes the dataset as an effective testbed for evaluating whether super-resolution techniques can bridge the gap between low-cost acquisition and high-accuracy defect detection.

To guarantee rigorous cross-evaluation and prevent data leakage, dataset partitioning was strictly executed at the source-image level prior to any degradation or super-resolution processing. Specifically, the original weld imagery was partitioned into independent training (80%), validation (10%), and test (10%) subsets. Subsequently, all derivative representations—including synthetic bicubic downsampling, realistic degradation, and single-image super-resolution reconstructions—were generated strictly within their assigned subsets. Consequently, no corresponding HR, LR, or SR variants of any single physical image appear across both training and evaluation splits, ensuring absolute domain separation and unbiased generalization assessment.

Generation of synthetic low-resolution weld images

The sensitivity of weld defect detection to image resolution and degradation was investigated by generating synthetic low-resolution (LR) images from the original high-resolution (HR) LoHiWELD samples using two distinct pipelines: bicubic downsampling and a realistic degradation process. In the bicubic setting, HR images were downsampled by scale factors of x3 and x4 using bicubic interpolation to obtain the corresponding LR images.

In the realistic setting, common image artifacts observed in real inspection systems were approximated prior to downsampling. With a probability of 0.8, HR images were blurred using a Gaussian filter with a radius sampled uniformly from the interval [0.3, 1.2]. The blurred images were then downsampled at scales of x3 and x4 using bicubic interpolation. To simulate compression artifacts, the resulting LR images were optionally re-encoded as JPEG with quality sampled uniformly between 35 and 75, with a probability of 0.7. Finally, with a probability of 0.6, zero-mean Gaussian noise with a standard deviation uniformly sampled from [2, 8] in the pixel-intensity domain was added, followed by clipping to the valid intensity range. The overall realistic image degradation process can be mathematically formulated as:

$$I_{LR} = (I_{HR} \circledast k) \downarrow_s + n \tag{1}$$

where I_{HR} represents the high-resolution image, $\circledast$ denotes the convolution operation, k is the Gaussian blur kernel, $\downarrow_s$ signifies the bicubic downsampling operation with a scale factor of s, and n represents the additive Gaussian noise. These two pipelines produced synthetic bicubic LR and realistic LR subsets, which were subsequently used in the super-resolution and detection experiments.

The parameter ranges and probabilities in this degradation pipeline were selected to reflect common imaging conditions in industrial welding environments without completely destroying the defect signals. Specifically, the Gaussian blur radius $r \in [0.3, 1.2]$ accounts for minor optical defocus across curved weld seams and slight mechanical vibrations on the production line; radii below 0.3 have minimal effect, whereas values above 1.2 prior to $\times 3$ or $\times 4$ downsampling would completely smooth out small micro-pores. The JPEG quality range $Q \in [35, 75]$ represents standard image compression used for data transmission across bandwidth-limited industrial networks, covering moderate compression ($Q = 75$) down to noticeable 8×8 blocking artifacts ($Q = 35$). Similarly, the additive Gaussian noise standard deviation $\sigma \in [2, 8]$ reflects typical sensor noise found in standard industrial cameras operating under factory lighting conditions (corresponding to signal-to-noise ratios of roughly 30–42 dB). Rather than testing at a single fixed degradation level, uniform sampling across these ranges evaluates the models over a broad mix of mild, moderate, and severe conditions (with a $0.8 \times 0.7 \times 0.6 = 33.6\%$ chance that all three degradations occur together). Regarding the impact of these parameters on the reported results, moving toward the more severe end of the ranges ($r \to 1.2$, $Q \to 35$, $\sigma \to 8$) affects different model architectures in distinct ways: GAN-based methods typically react to higher noise (σ) and stronger JPEG compression (Q) by turning noise and block edges into false surface textures, whereas standard CNNs tend to over-smooth fine defect boundaries as the blur radius (r) increases.

Super-resolution models

To reconstruct HR-like images from degraded inputs, we employed a suite of state-of-the-art single-image super-resolution (SISR) architectures spanning classical convolutional, residual, generative-adversarial, and transformer-based paradigms: SRCNN[15], EDSR[16], SRGAN[17], ESRGAN[18], SwinIR[19], and the developed SATSR. To eliminate domain shift stemming from external natural-image benchmarks (e.g., DIV2K), all evaluated architectures were trained from scratch exclusively on paired single-channel grayscale sub-image patches cropped dynamically from the dedicated LoHi-WELD training split (80%), ensuring that feature extraction and restoration filters are tailored specifically to industrial metallic textures.

For EDSR, SwinIR, and SATSR, models were trained on low-resolution patches of size 32×32 pixels (corresponding to 96×96 and 128×128 high-resolution targets at $\times 3$ and $\times 4$ scales, respectively) with a mini-batch size of 4. SRGAN and ESRGAN were trained on 24×24 low-resolution patches with a batch size of 4, whereas SRCNN extracted 33×33 patches with a stride of 14 and a batch size of 16. Optimization was performed up to 400 epochs (200 epochs for ESRGAN) governed by an early stopping mechanism with a patience of 20 epochs monitoring validation PSNR. SATSR was optimized using AdamW ($\mathrm{LR} = 2 \times 10^{-4}$, weight decay $= 10^{-4}$), SwinIR using Adam ($\mathrm{LR} = 2 \times 10^{-4}$), while EDSR, SRCNN, SRGAN, and ESRGAN utilized Adam ($\mathrm{LR} = 1 \times 10^{-4}$). Regarding objective functions, EDSR, SwinIR, and SATSR were supervised via pixel-level $\mathcal{L}_1$ loss; SRCNN was trained using Mean Squared Error (MSE); SRGAN combined $\mathcal{L}_1$ reconstruction loss, frozen VGG-19 perceptual feature loss, and binary cross-entropy adversarial loss with weighting factors of 1.0, 1.0, and 10^{-3}, respectively; and ESRGAN incorporated relativistic adversarial loss with weighting parameters of 1.0 (pixel $\mathcal{L}_1$), 1.0 (perceptual), and 5×10^{-3} (adversarial). Hyperparameter configurations (such as patch dimensions, learning rates, and loss weights) for each baseline architecture were deliberately chosen to reflect the canonical implementations and stability constraints established in their respective foundational works, ensuring that each model operates within its empirically validated optimal operating regime.

Each trained SR model was subsequently evaluated on low-resolution images produced by the bicubic and realistic pipelines at different magnification factors ($\times 3$ and $\times 4$), generating reconstructed images that share the spatial grid of the high-resolution references. Within these synthetic degradation experiments, bounding-box coordinates and defect labels from the HR images were directly reused for the corresponding synthetic LR and super-resolved synthetic images (leveraging scale-invariant normalized coordinates), ensuring consistent cross-model comparison without re-annotation. In contrast, for all experiments involving real LR imagery (L-WELD) and their super-resolved counterparts, the native, independently acquired L-WELD annotations were strictly employed.

Selective aggregation transformer for super-resolution (SATSR)

The Selective Aggregation Transformer for Super-Resolution (SATSR) serves as the primary image restoration architecture in this study, building upon the foundational SAT model introduced by Tran et al.[20]. While the original SAT architecture was designed for general natural image super-resolution[20], our proposed SATSR is specifically extended and optimized for the unique visual characteristics of industrial weld defect datasets. In contrast to traditional Vision Transformers, which exhibit quadratic computational complexity relative to image resolution[20], SATSR utilizes a highly efficient Selective Aggregation Attention mechanism.

A significant architectural difference between our method and the original SAT lies in the token compression strategy. The baseline SAT model employs a Density-driven Token Aggregation (DTA) algorithm, which involves density-guided center selection and similarity-weighted aggregation. To streamline the network for fast, high-resolution industrial inspection and reduce computational overhead, our SATSR replaces the complex DTA algorithm with a straightforward spatial reduction technique. As shown in Fig. 1 (c), for an input feature map $X \in \mathbb{R}^{C \times H \times W}$, the Query ($Q$) is projected directly from the flattened tokens, whereas the Key (K) and Value (V) are obtained from a spatially reduced feature map produced through average pooling:

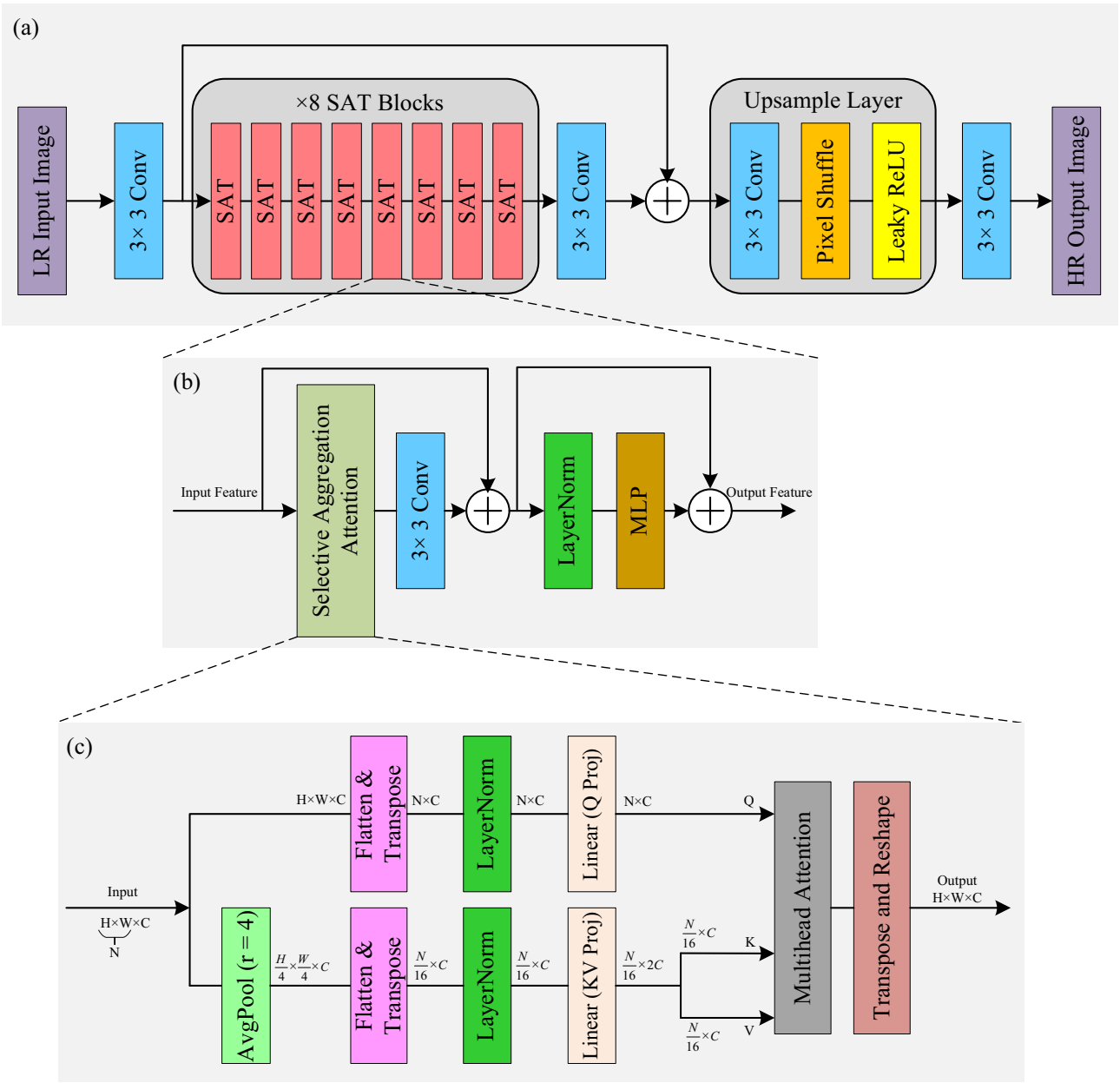


Fig. 1 The detailed architecture of the developed model: (**a**) Overall network architecture for super-resolution, consisting of shallow feature extraction, a deep feature extraction backbone with cascaded SAT blocks, and a pixel-shuffle reconstruction module; (**b**) The internal structure of the Selective Aggregation Transformer Block (SATB); (**c**) Detailed data flow of the Selective Aggregation Attention (SAA) module with spatial downsampling for Key and Value projections.

$$K, V = W_{kv}(\text{AvgPool}(X)), \quad Q = W_q(X) \tag{2}$$

Here, W_{kv} and W_q denote linear projection weights. By reducing the spatial dimensions of K and V by a factor of r (empirically set to 4), the attention complexity is significantly decreased, making it highly well-suited for fast industrial image processing while avoiding clustering bottlenecks.

Furthermore, we introduce key structural modifications to the SAT block itself to better suit the weld defect detection task. Whereas the original SAT architecture utilizes a standard transformer block design (Pre-Norm followed by attention) and alternates between Local Transformer Blocks (LTB) and Selective Aggregation

Transformer Blocks (SATB)[20], as shown in Fig. 1 (a), our network core simplifies the macro-architecture by employing eight consecutive SAT blocks while enhancing their internal micro-architecture. Specifically, as illustrated in Fig. 1 (b), we inject a 3×3 convolutional layer immediately after the Selective Aggregation Attention module within each block. This architectural modification is crucial for industrial inspection, as the convolution operation captures local inductive biases and preserves high-frequency structural defect boundaries that pure self-attention might otherwise smooth over. Following this, a Multi-Layer Perceptron (MLP) with Gaussian Error Linear Unit (GELU) activation is applied to enhance non-linear feature representations. Global and local residual connections are strategically placed to facilitate gradient flow and preserve low-frequency background information. Finally, the upsampling stage employs a 3×3 convolution, sub-pixel convolution (PixelShuffle), and a Leaky ReLU activation to efficiently reconstruct the high-resolution output without substantial computational overhead.

From a theoretical and signal-processing perspective, these architectural modifications establish a task-driven spectral decomposition tailored for low-contrast industrial weld imagery rather than representing a mere heuristic simplification. In the foundational SAT architecture, Deformable Token Aggregation (DTA) predicts dynamic sampling offsets $\Delta p_k = \mathcal{F}_{\text{offset}}(X)$ conditioned on local spatial gradients. While effective on high-contrast natural scenes, weld surfaces under severe $\times 3$ and $\times 4$ degradations exhibit low signal-to-noise ratios (SNR), where sensor noise and compression artifacts dominate local intensity gradients. Consequently, dynamic offset estimation suffers from high stochastic variance, inducing non-uniform geometric warping (topological jitter) along delicate defect boundaries. Although such sub-pixel warping may not heavily penalize pixel-wise fidelity metrics (e.g., PSNR), it severely disrupts the spatial translation equivariance required by downstream object detectors for precise bounding-box regression ($\mathcal{L}_{\text{box}}$ and $\mathcal{L}_{\text{dfl}}$). By replacing DTA with deterministic spatial average pooling, SATSR introduces a shift-invariant, zero-phase low-pass spatial regularizer that suppresses high-frequency stochastic variance by a factor of $1/K^2$ while strictly preserving the geometric centroid of minute anomalies. Complementarily, because transformer self-attention inherently acts as a global low-pass operator prone to over-smoothing sharp transitions, the injected 3×3 convolutional projection explicitly restores localized high-frequency inductive bias and translation equivariance. Together, the 3×3 convolution and spatial average pooling form a synergistic frequency-decoupling mechanism—simultaneously capturing sharp anisotropic defect contours (e.g., discontinuities) and stable isotropic regional representations (e.g., pores and stains) without the optimization instability of deformable offsets.

Object detector architecture and training

Weld defect detection utilized a YOLO-based one-stage object detector configured for 1280-pixel input images. The total loss function $\mathcal{L}_{total}$ comprised three components: bounding-box regression loss ($\mathcal{L}_{box}$), classification loss ($\mathcal{L}_{cls}$), and distribution focal loss ($\mathcal{L}_{dfl}$). The mathematical formulation is as follows:

$$\mathcal{L}_{total} = \lambda_{box}\mathcal{L}_{box} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dfl}\mathcal{L}_{dfl} \tag{3}$$

where the relative contributions of each loss component were controlled by scalar weights, empirically set to $\lambda_{box} = 7.5$, $\lambda_{cls} = 0.5$, and $\lambda_{dfl} = 1.5$, respectively. This configuration achieved a balanced trade-off between localization accuracy and class discrimination across the four weld defect categories.

To rigorously assess domain-specific representations without conflating resolution shifts with model transfer discrepancies, separate YOLOv11-nano detector instances were independently fine-tuned for each training domain (e.g., native HR, synthetic LR regimes, specific SR reconstructions, and real LR). Starting from official pre-trained COCO checkpoints, all detector instances were trained for 400 epochs with a batch size of 8 and an early stopping patience of 30 epochs on a single GPU. Maintaining identical architectural configurations, loss parameters, and optimization schedules across all instances guarantees that downstream performance variations directly reflect domain-specific feature fidelity rather than optimization discrepancies.

Experimental scenarios

Multiple detection scenarios were designed to decouple the effects of resolution, degradation, and super-resolution on weld defect recognition. In the baseline scenario, the YOLO detector was trained and evaluated on original HR images, providing a reference for subsequent experiments. In the synthetic LR scenario, the detector was evaluated on bicubic and realistic LR images generated from HR data, which enabled quantification of performance degradation attributable solely to reduced resolution and artifacts. In the super-resolution scenario, the same detector was evaluated on super-resolved versions of the LR images produced by SRCNN[15], EDSR[16], SRGAN[17], ESRGAN[18], SwinIR[19], and SATSR. This approach facilitated analysis of how different reconstruction models mediate the trade-off between perceptual quality and detection accuracy across bicubic and realistic degradation settings.

In addition to these core experiments, further configurations were evaluated in which the detector was trained on HR images and tested on synthetic LR, trained on synthetic LR and tested on real LR, and trained on super-resolved synthetic LR and tested on super-resolved real LR images. These extended scenarios were intended to emulate practical deployment conditions where the training and deployment image distributions may differ due to variations in acquisition hardware, compression, and noise characteristics.

Evaluation metrics

Detection performance was evaluated using established object detection metrics. The primary components, Precision and Recall, are defined as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{4}$$

Here, *TP*, *FP*, and *FN* represent the number of true positives, false positives, and false negatives, respectively. A prediction is considered a true positive if its Intersection over Union (IoU) with the ground truth bounding box exceeds a specified threshold. The IoU is computed as:

$$IoU = \frac{Area(B_{pred} \cap B_{gt})}{Area(B_{pred} \cup B_{gt})} \tag{5}$$

In this equation, B_{pred} and B_{gt} denote the predicted and ground-truth bounding boxes, respectively. Based on these calculations, the mean Average Precision (mAP) is reported, which averages the Average Precision (AP) across all *N* defect classes:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{6}$$

Specifically, mAP is reported at an IoU threshold of 0.5 (mAP50) and as an average over IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05 (mAP50–95). Per-class AP50 and AP50–95 were also computed for pores, deposits, discontinuities, and stains to provide a detailed assessment of detection robustness.

For super-resolution evaluation, image-quality metrics such as PSNR, SSIM, LPIPS[28], DISTS[29], and NIQE[30] were employed. The Peak Signal-to-Noise Ratio (PSNR) quantifies absolute pixel-level errors and is defined as:

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \tag{7}$$

To assess structural distortions, which are essential for characterizing defect morphology, the Structural Similarity Index Measure (SSIM) is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

In this context, μ and σ denote the mean and variance of the image patches x (original HR) and y (reconstructed SR), σ_{xy} is their covariance, and C_1, C_2 are small constants introduced to stabilize the division. In addition to these perceptual metrics, computational efficiency was evaluated using the number of parameters, FLOPs, and inference time per image, facilitating a comprehensive analysis of the trade-off between reconstruction quality and downstream detection performance.

Experiments

Experimental design

All experiments were performed on a workstation featuring an NVIDIA GeForce RTX 4070 Ti SUPER GPU (16 GB VRAM), an Intel Core i9-14900KF CPU (3.20 GHz), and 64 GB of system RAM. The software environment included Python 3.9, PyTorch 2.5.1, and CUDA 12.1. Model implementation and training were conducted using the Ultralytics framework.

To evaluate the impact of super-resolution (SR) on weld defect detection, eight complementary scenarios were designed to isolate the effects of image quality, degradation type, and domain shift on a YOLO-based detector. YOLOv11-nano was selected as the detection backbone because of its efficient design and favorable balance between accuracy and computational cost. Its compact architecture is particularly suitable for real-time industrial inspection scenarios where lightweight deployment and low inference latency are critical.

The experimental protocol begins with a high-resolution (HR) benchmark, proceeds to synthetic low-resolution (LR) and real LR settings, and then extends to cross-domain experiments in which the detector is trained and tested across different resolution domains. This approach enables determination of not only whether SR is beneficial, but also under which conditions it provides improvements.

The experiments addressed two primary questions. First, to what extent does detection performance decline when the detector is evaluated on degraded images compared to HR images? Second, can SR reconstruction recover discriminative details sufficiently to restore detection accuracy under both synthetic and real degradation? To address these questions, overall detection metrics and class-wise results are reported across eight distinct operational scenarios. Because H-WELD and L-WELD represent non-aligned image collections acquired independently using distinct optical sensors, their ground-truth annotations are strictly domain-specific. Under the normalized YOLO bounding-box coordinate scheme ($x_c, y_c, w, h \in [0,1]$), spatial scaling naturally preserves relative topological alignment. For all in-domain experiments involving HR, synthetic LR, and super-resolved synthetic images (Scenarios 1–3 and 6), ground-truth annotations derive entirely from the verified H-WELD label set. Conversely, all experiments involving real low-resolution images and their super-resolved outputs (Scenarios 4 and 5) rely strictly on the independent ground-truth annotations provided specifically for the L-WELD subset. For cross-domain evaluations (Scenarios 7 and 8), models trained on H-WELD-derived inputs are evaluated exclusively against the independent L-WELD annotations, ensuring zero label leakage across sensor domains. A comprehensive overview of the input modalities, training domains, and corresponding annotation protocols across all eight scenarios is summarized in Table 2.

Evaluation protocol

All detection experiments utilized an identical training configuration to ensure fair comparison across scenarios. Unless otherwise specified, the detector was trained for 400 epochs with an input image size of 1280, a batch size of 8, and a patience parameter of 30. Optimization settings were held constant across experiments with box = 7.5,

Table 2 Detailed experimental workflow, data flow, input modalities, detector models, and annotation provenance across all evaluated scenarios.

Scenario	Detector Model	Training Domain	Validation Domain	Testing Domain	SR Model	Annotation Source
1	Trained on HR	HR (H-WELD)	HR (H-WELD)	HR (H-WELD)	None	Native H-WELD
2	Retrained on Syn-LR	Synthetic LR	Synthetic LR	Synthetic LR	None	H-WELD (Normalized)
3	Retrained on SR-Syn	SR Synthetic LR	SR Synthetic LR	SR Synthetic LR	Specific SR model	Native H-WELD
4	Retrained on Real-LR	Real LR (L-WELD)	Real LR (L-WELD)	Real LR (L-WELD)	None	Native L-WELD
5	Retrained on SR-Real	SR Real LR	SR Real LR	SR Real LR	Specific SR model	L-WELD (Normalized)
6	Reused from Scen. 1	HR (H-WELD)	HR (H-WELD)	Synthetic LR	None	H-WELD (Normalized)
7	Reused from Scen. 2	Synthetic LR	Synthetic LR	Real LR (L-WELD)	None	Native L-WELD (Indep.)
8	Reused from Scen. 3	SR Synthetic LR	SR Synthetic LR	SR Real LR	Specific SR model	Native L-WELD (Indep.)

Table 3 Detection performance of the baseline model on high-resolution (HR) images.

Class	**Precision**	**Recall**	**AP50**	**AP50-95**
All	0.68	0.59	0.66	0.30
Pore	0.65	0.46	0.58	0.26
Deposit	0.73	0.60	0.66	0.32
Discontinuity	0.70	0.74	0.75	0.32
Stain	0.64	0.54	0.63	0.30

Table 4 Impact of synthetic low-resolution degradation (bicubic and realistic) on detection performance.

Degradation type	Scale	Precision	Recall	mAP50	mAP50-95
Bicubic LR	×3	0.62	0.55	0.57	0.24
	×4	0.62	0.48	0.49	0.20
Realistic LR	×3	0.64	0.50	0.54	0.23
	×4	0.57	0.48	0.48	0.19

cls = 0.5, and dfl = 1.5. In the real-LR setting, an additional input size of 640 was evaluated to assess the detector's dependence on spatial resolution.

Performance was evaluated using Precision, Recall, mAP50, and mAP50-95. These metrics were selected as they collectively capture localization quality, class discrimination, and overall detection robustness. Additionally, per-class scores were reported to determine whether specific defect types benefit more from super-resolution (SR) than others.

Scenario 1: Detection on high-resolution images

The HR benchmark defines the detector's upper performance limit under optimal imaging conditions. As shown in Table 3, the detector achieved a Precision of 0.68, Recall of 0.59, mAP50 of 0.66, and mAP50-95 of 0.30 on HR images. Of the four defect classes, discontinuity was detected with the highest reliability (AP50 of 0.75), whereas pore was the most challenging (AP50 of 0.58). These findings demonstrate that the detector can effectively learn the weld-defect distribution when sufficient visual detail is present.

Scenario 2: Detection on synthetic low-resolution images

Table 4 demonstrates that when the same detector was evaluated on synthetic LR inputs, performance decreased across all degradation settings. As the downsampling factor increased to ×3 and ×4, performance declined consistently, indicating a progressive loss of fine-grained defect cues. The realistic degradation settings proved more

challenging than the bicubic ones at the same scale, suggesting that non-ideal degradation more accurately reflects the severity of real inspection conditions.

Scenario 3: Detection on super-resolved synthetic low-resolution images

To examine whether SR methods can recover information lost due to synthetic degradation, we evaluated several reconstruction models in terms of visual quality, computational complexity, and downstream detection performance. As summarized in Table 5, the results indicate that SR can partially restore detection accuracy, although its effectiveness depends strongly on the SR architecture and the type of degradation.

As presented in Table 5, under the bicubic LR ×3 setting, EDSR achieved the highest mAP50 at 0.60, followed by SwinIR at 0.59 and SATSR at 0.58. In the more challenging bicubic LR ×4 scenario, performance improvements were limited, with SRCNN and SATSR attaining mAP50 values of approximately 0.51 to 0.52. For realistic degradations, detection gains were generally less consistent across all models, suggesting that visually sharper reconstructions do not necessarily yield proportional improvements in detection accuracy.

As shown in Table 6, evaluating models across reconstruction fidelity, computational cost, and downstream detection performance highlights the distinct trade-offs offered by each architecture. While larger models like EDSR achieve slightly higher detection accuracy in certain settings (e.g., Bicubic ×3 in Table 5), they incur substantial computational costs (exceeding 350 GFLOPs). Conversely, SATSR delivers competitive detection accuracy and reconstruction quality with significantly lower overhead (approximately $\sim 1M$ parameters, under 11 GFLOPs, and an inference latency under 35 ms). This favorable balance between task performance and efficiency positions SATSR as an attractive, viable alternative for resource-constrained industrial deployments, even though it does not uniformly outperform heavier models across all metrics.

To ensure reproducible profiling in Table 6, execution times were benchmarked on a dedicated workstation (NVIDIA RTX 4070 Ti SUPER GPU, 16 GB VRAM; Intel Core i9-14900KF CPU) using batch size 1 across all

Table 5 Detection performance on super-resolved synthetic low-resolution images.

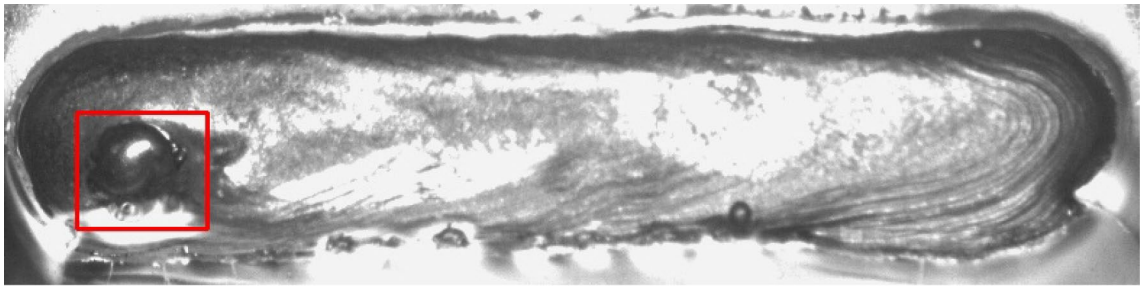
Model	Degradation type	Scale	Precision	Recall	mAP50	mAP50-95
SRCNN[15]	Bicubic LR	×3	0.69	0.53	0.58	0.25
		×4	0.59	0.52	0.52	0.22
	Realistic LR	×3	0.62	0.53	0.55	0.24
		×4	0.68	0.45	0.50	0.21
EDSR[16]	Bicubic LR	×3	0.68	0.55	0.60	0.25
		×4	0.60	0.50	0.52	0.21
	Realistic LR	×3	0.67	0.50	0.55	0.23
		×4	0.62	0.45	0.47	0.20
SRGAN[17]	Bicubic LR	×3	0.69	0.55	0.57	0.25
		×4	0.60	0.49	0.51	0.21
	Realistic LR	×3	0.58	0.52	0.52	0.22
		×4	0.65	0.45	0.46	0.19
ESRGAN[18]	Bicubic LR	×3	0.64	0.52	0.55	0.24
		×4	0.59	0.48	0.50	0.20
	Realistic LR	×3	0.63	0.49	0.52	0.22
		×4	0.63	0.45	0.48	0.19
SwinIR[19]	Bicubic LR	×3	0.68	0.56	0.59	0.26
		×4	0.65	0.49	0.51	0.21
	Realistic LR	×3	0.64	0.49	0.51	0.22
		×4	0.68	0.48	0.51	0.21
SATSR (Ours)	Bicubic LR	×3	0.68	0.52	0.58	0.26
		×4	0.66	0.49	0.51	0.21
	Realistic LR	×3	0.63	0.51	0.52	0.22
		×4	0.62	0.47	0.50	0.20

Table 6 Super-Resolution Performance across different models and degradation settings.

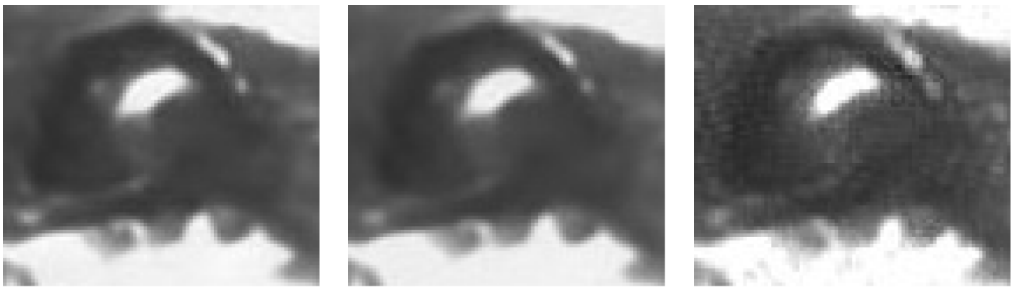
Model	Degradation	PSNR	SSIM	LPIPS	DISTS	NIQE	Params(M)	Time(ms)	FLOPs(G)
SRCNN[15]	Bicubic LR ×3	19.91	0.74	0.34	0.22	7.01	0.057281	3.79	4.22
	Bicubic LR ×4	18.58	0.65	0.42	0.24	6.16	0.057281	3.58	7.50
	Realistic LR ×3	18.83	0.71	0.39	0.24	7.45	0.057281	3.73	4.22
	Realistic LR ×4	18.83	0.64	0.47	0.26	7.23	0.057281	3.61	7.50
EDSR[16]	Bicubic LR ×3	27.13	0.80	0.32	0.19	6.71	43.670790	75.29	357.74
	Bicubic LR ×4	25.72	0.74	0.42	0.22	7.54	43.080705	56.02	411.03
	Realistic LR ×3	26.32	0.76	0.37	0.21	8.39	43.670790	76.00	357.74
	Realistic LR ×4	24.90	0.71	0.47	0.24	7.76	43.080705	56.06	411.03
SRGAN[17]	Bicubic LR ×3	24.24	0.72	0.16	0.13	5.45	1.565651	9.84	13.19
	Bicubic LR ×4	23.02	0.64	0.22	0.16	6.43	1.528724	9.23	16.82
	Realistic LR ×3	23.88	0.68	0.22	0.17	4.20	1.565651	9.86	13.19
	Realistic LR ×4	22.43	0.58	0.24	0.18	5.32	1.528724	9.17	16.82
ESRGAN[18]	Bicubic LR ×3	24.32	0.68	0.18	0.17	3.73	9.040513	36.46	76.45
	Bicubic LR ×4	22.27	0.63	0.26	0.22	7.97	9.003585	23.11	81.92
	Realistic LR ×3	24.86	0.71	0.18	0.15	4.41	9.040513	36.38	76.45
	Realistic LR ×4	22.78	0.62	0.25	0.21	7.70	9.003585	23.10	81.92
SwinIR[19]	Bicubic LR ×3	27.09	0.80	0.30	0.18	7.04	1.053761	19277.21	6.49
	Bicubic LR ×4	25.48	0.74	0.40	0.21	7.51	1.016833	232.86	9.85
	Realistic LR ×3	26.24	0.76	0.35	0.20	7.43	1.053761	19280.51	6.49
	Realistic LR ×4	24.80	0.71	0.45	0.24	8.15	1.016833	232.97	9.85
SATSR (Ours)	Bicubic LR ×3	27.11	0.79	0.30	0.18	7.13	1.033537	34.86	6.89
	Bicubic LR ×4	24.98	0.72	0.43	0.22	7.64	0.996609	12.64	10.25
	Realistic LR ×3	26.27	0.76	0.37	0.20	7.16	1.033537	34.99	6.89
	Realistic LR ×4	24.41	0.70	0.46	0.23	7.76	0.996609	12.53	10.25

103 full-resolution test images (2048×1080 native dimensions), preceded by 20 warm-up runs via synchronized CUDA timers excluding disk I/O. Notably, while theoretical FLOPs are normalized over fixed input patches, full-image inference requires models under the ×3 scale to process 1.78× more spatial tokens (683×360) than under the ×4 scale (512×270). Consequently, deep feature extraction over larger input feature maps results in longer runtimes across all evaluated architectures at ×3 compared to ×4. Furthermore, for SwinIR, the spatial dimensions under ×3 are non-divisible by its fixed window partition (8×8), triggering memory-conserving tiled inference with excessive window padding and patch stitching that drastically elevates its latency to ≈ 19.28 seconds. In contrast, SATSR natively processes full frames without windowing or tiling bottlenecks, maintaining rapid inference (34.86 ms at ×3 and 12.64 ms at ×4).

To further illustrate the impact of different super-resolution algorithms on defect morphology, a qualitative visual comparison is presented in Figure 2. As observed in the zoomed-in patches of a pore defect under realistic degradation (×3 scale), shallow convolutional networks such as SRCNN yield overly smoothed edges, blurring the distinct boundaries required for accurate bounding box regression. Furthermore, while GAN-based approaches (SRGAN and ESRGAN) enhance global contrast, they introduce noticeable high-frequency artifacts and hallucinated textures that distort the true topology of the weld surface. In contrast, our proposed Selective Aggregation Transformer for Super-Resolution (SATSR) successfully balances noise suppression with structural preservation. By leveraging the efficient spatial-reduction attention mechanism and local convolutions, SATSR effectively recovers sharp, well-defined defect boundaries and realistic metal surface reflections, closely approximating the original high-resolution reference. This structural preservation illustrates why SATSR maintains competitive downstream detection accuracy (mAP) while preserving key topological features, presenting a favorable trade-off against both over-smoothing and artifact-inducing models within a compact footprint. To evaluate the visual reconstruction fidelity across all defect geometries, extended qualitative comparisons covering pore, deposit, discontinuity, and stain instances are provided in the Appendix (Figure 4).



(a) High-Resolution (HR) Reference Image



(b) SRCNN (c) EDSR (d) SRGAN



(e) ESRGAN (f) SwinIR (g) SATSR (Ours)

Fig. 2 Qualitative visual comparison of different super-resolution methods under the realistic degradation setting (×3 scale). The red box indicates the zoomed-in region. Our proposed SATSR effectively recovers sharp defect boundaries and structural details compared to convolution-based and GAN-based models.

Table 7 Detection performance evaluated on real low-resolution (LR) weld images at different input scales.

Input size	Precision	Recall	mAP50	mAP50-95
1280	0.60	0.58	0.59	0.23
640	0.61	0.59	0.61	0.24

Scenario 4: Detection on real low-resolution images

As shown in Table 7, the real LR experiments demonstrate that the detector is more sensitive to practical image degradation than to synthetic downsampling alone. With an input size of 1280, the detector achieved an overall mAP50 of 0.59, whereas reducing the input size to 640 increased it slightly to 0.61. This outcome is attributable to the native resolution of the real LR images, which is 640×480 pixels. When these images are resized to a 1280 input size, the YOLO preprocessing pipeline artificially upsamples them, typically using bicubic or bilinear interpolation. This upsampling process introduces blurring and scaling artifacts without recovering genuine high-frequency details, thereby degrading feature representations, which helps explain why the native 640 input scale yields marginally better detection performance.

Scenario 5: Detection on super-resolved real low-resolution images

Table 8 shows that, for real LR inputs, the benefit of SR was weaker and less consistent than in the synthetic scenario. Bicubic upsampling alone achieved an mAP50 near 0.54, while several SR methods yielded only modest gains. SRCNN produced the highest result on Bicubic LR ×3 with an mAP50 of 0.573, and also achieved the best performance on Realistic LR ×4 with an mAP50 of 0.580. In contrast, some GAN-based methods caused fluctuations across degradation types, indicating that perceptual enhancement does not necessarily translate to improved task-oriented detection. These findings indicate that real LR detection remains difficult after reconstruction, especially when degradation is severe or highly non-uniform.

Statistical Reliability and Variance Analysis: To substantiate that the performance margins observed under realistic degradation (Scenario 5) reflect genuine architectural advantages rather than stochastic training artifacts, the principal detection experiments were repeated across five independent runs ($N = 5$) with distinct random seeds. As tabulated in Table 9, the proposed SATSR architecture demonstrates exceptional stability, yielding the lowest variance (2.55×10^{-6}) and a standard deviation of only ± 0.0016 in mAP-50–95.

Table 8 shows that, for real LR inputs, the benefit of SR was weaker and less consistent than in the synthetic scenario. Bicubic upsampling alone achieved an mAP50 near 0.54, while several SR methods yielded only modest gains. SRCNN produced the highest result on Bicubic LR ×3 with an mAP50 of 0.573, and also achieved the best performance on Realistic LR ×4 with an mAP50 of 0.580. In contrast, some GAN-based methods caused fluctuations across degradation types, indicating that perceptual enhancement does not necessarily translate to improved task-oriented detection. These findings indicate that real LR detection remains difficult after reconstruction, especially when degradation is severe or highly non-uniform. Furthermore, to substantiate that these performance margins reflect genuine architectural behavior rather than stochastic training artifacts, the principal

Table 8 Detection performance on super-resolved real low-resolution images.

Model	Degradation type	Scale	Precision	Recall	mAP50	mAP50-95
Baseline	Bicubic upsampling	×3	0.540	0.556	0.541	0.203
		×4	0.547	0.542	0.540	0.198
SRCNN[15]	Bicubic LR	×3	0.594	0.580	0.573	0.229
		×4	0.575	0.579	0.565	0.226
	Realistic LR	×3	0.528	0.587	0.552	0.222
		×4	0.573	0.591	0.580	0.232
EDSR[16]	Bicubic LR	×3	0.536	0.549	0.533	0.203
		×4	0.530	0.522	0.516	0.196
	Realistic LR	×3	0.516	0.545	0.516	0.198
		×4	0.571	0.522	0.526	0.196
SRGAN[17]	Bicubic LR	×3	0.589	0.531	0.529	0.201
		×4	0.443	0.513	0.483	0.181
	Realistic LR	×3	0.582	0.507	0.526	0.203
		×4	0.491	0.502	0.486	0.182
ESRGAN[18]	Bicubic LR	×3	0.534	0.523	0.524	0.196
		×4	0.558	0.493	0.519	0.196
	Realistic LR	×3	0.573	0.546	0.524	0.199
		×4	0.552	0.515	0.509	0.189
SwinIR[19]	Bicubic LR	×3	0.538	0.549	0.532	0.201
		×4	0.546	0.521	0.515	0.188
	Realistic LR	×3	0.545	0.538	0.528	0.200
		×4	0.533	0.506	0.509	0.187
SATSR (Ours)	Bicubic LR	×3	0.547	0.521	0.530	0.205
		×4	0.506	0.564	0.533	0.192
	Realistic LR	×3	0.521	0.556	0.511	0.196
		×4	0.509	0.541	0.524	0.194

Table 9 Statistical reliability evaluation over five independent runs ($N = 5$) under Scenario 5 (Realistic LR degradation). Metrics are reported as Mean ± Std, accompanied by sample variance, 95% confidence interval (CI), and paired *t*-test against SATSR.

Method	mAP@0.50	mAP@0.50:0.95	Variance ($\times 10^{-5}$)	95% CI (mAP50-95)	*p*-value
SRCNN	0.5640 ± 0.0093	0.2268 ± 0.0046	2.09	[0.2211, 0.2325]	< 0.001
EDSR	0.5146 ± 0.0068	0.1908 ± 0.0031	0.98	[0.1869, 0.1947]	0.4759
SRGAN	0.4884 ± 0.0088	0.1812 ± 0.0049	2.40	[0.1751, 0.1873]	0.0092
ESRGAN	0.5112 ± 0.0061	0.1876 ± 0.0032	1.04	[0.1836, 0.1916]	0.0810
SwinIR	0.5121 ± 0.0105	0.1924 ± 0.0057	3.30	[0.1853, 0.1996]	0.8287
SATSR (Ours)	**0.5222 ± 0.0061**	**0.1919 ± 0.0016**	**0.25**	**[0.1900, 0.1939]**	—

Table 10 Cross-resolution detection performance: models trained on high-resolution images and evaluated on synthetic low-resolution images.

Input size	Degradation type	Scale	Precision	Recall	mAP50	mAP50-95
1280	Bicubic LR	×3	0.53	0.44	0.45	0.18
		×4	0.48	0.35	0.34	0.13
	Realistic LR	×3	0.51	0.38	0.39	0.15
		×4	0.43	0.30	0.27	0.10
640	Bicubic LR	×3	0.61	0.45	0.48	0.19
		×4	0.55	0.39	0.37	0.13
	Realistic LR	×3	0.56	0.41	0.41	0.16
		×4	0.45	0.32	0.30	0.11

detection experiments were repeated across five independent runs ($N = 5$) with distinct random seeds. As tabulated in Table 9, the proposed SATSR architecture demonstrates exceptional statistical reliability, achieving an mAP50 of 0.5222 ± 0.0061 and the lowest variance (2.55×10^{-6}, Std = ± 0.0016) in mAP50-95.

Scenario 6: Train on HR and test on synthetic LR

Table 10 demonstrates that the cross-resolution setting highlights the detector's strong dependence on the training domain. Training on HR images and testing directly on synthetic low-resolution (LR) images resulted in a substantial performance decline, particularly at higher levels of degradation. The highest performance was observed for Bicubic LR ×3, achieving an mAP50 of 0.45 at a 1280 input size, while Bicubic LR ×4 decreased to 0.34. Realistic degradations also posed challenges, with mAP50 values ranging from 0.39 to 0.27.

Scenario 7: Train on synthetic LR and test on real LR

Table 11 demonstrates that training on synthetic LR data and testing on real LR data substantially reduces detection quality. This outcome indicates that synthetic degradations do not fully represent the complexity of real capture conditions. In the split-real-LR evaluation, performance remained low, with mAP50 values ranging from 0.17 to 0.24. When the entire real-LR set was used for testing, the highest result achieved was an mAP50 of 0.23. These results indicate that detectors trained exclusively on synthetic LR images generalize poorly to real degradations, particularly for small and visually ambiguous defect classes.

Scenario 8: Train on SR synthetic and test on SR real low-resolution images

Table 12 presents the evaluation for Scenario 8, which examines whether applying super-resolution to both synthetic training data and real testing data can mitigate cross-domain discrepancies. Detection performance across all evaluated frameworks indicates a pronounced synthetic-to-real domain shift, with mAP50 scores ranging

Table 11 Cross-domain detection performance of models trained on synthetic low-resolution data and evaluated on real low-resolution images.

Testing protocol	Degradation type	Scale	Precision	Recall	mAP50	mAP50-95
Split real LR	Bicubic LR	×3	0.28	0.24	0.19	0.05
		×4	0.31	0.28	0.24	0.07
	Realistic LR	×3	0.28	0.25	0.19	0.05
		×4	0.27	0.22	0.17	0.04
Whole real LR	Bicubic LR	×3	0.30	0.25	0.19	0.05
		×4	0.34	0.29	0.23	0.07
	Realistic LR	×3	0.28	0.25	0.19	0.05
		×4	0.28	0.22	0.15	0.04

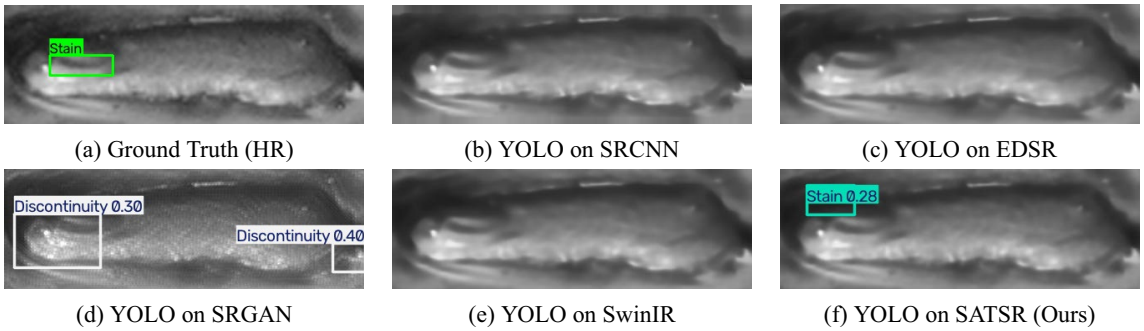
Table 12 Cross-domain detection performance: models trained on super-resolved synthetic images and tested on super-resolved real low-resolution images.

Model	Degradation type	Scale	Precision	Recall	mAP50	mAP50-95
SRCNN[15]	Bicubic LR	×3	0.197	0.146	0.095	0.023
		×4	0.225	0.195	0.123	0.031
	Realistic LR	×3	0.163	0.143	0.078	0.020
		×4	0.290	0.210	0.151	0.037
EDSR[16]	Bicubic LR	×3	0.259	0.215	0.154	0.042
		×4	0.252	0.197	0.158	0.043
	Realistic LR	×3	0.257	0.198	0.136	0.037
		×4	0.287	0.205	0.155	0.041
SRGAN[17]	Bicubic LR	×3	0.260	0.190	0.144	0.038
		×4	0.223	0.197	0.119	0.029
	Realistic LR	×3	0.256	0.229	0.144	0.039
		×4	0.219	0.180	0.114	0.027
ESRGAN[18]	Bicubic LR	×3	0.345	0.253	0.209	0.060
		×4	0.251	0.187	0.148	0.040
	Realistic LR	×3	0.244	0.172	0.141	0.035
		×4	0.220	0.174	0.116	0.030
SwinIR[19]	Bicubic LR	×3	0.246	0.225	0.163	0.044
		×4	0.318	0.224	0.195	0.050
	Realistic LR	×3	0.241	0.214	0.147	0.039
		×4	0.254	0.166	0.113	0.028
SATSR (Ours)	Bicubic LR	×3	0.283	0.218	0.179	0.048
		×4	0.271	0.219	0.181	0.047
	Realistic LR	×3	0.239	0.192	0.146	0.040
		×4	0.280	0.197	0.160	0.040

from 0.11 to 0.21. While ESRGAN peaks at mAP50=0.209 at Bicubic LR ×3 and SwinIR at 0.195 at ×4, both architectures incur significant computational overhead and slow inference times that impede real-time industrial deployment.

In contrast, the proposed SATSR approach demonstrates a competitive and balanced profile, maintaining consistent detection accuracy (mAP-50 between 0.146 and 0.181) while operating with substantially lower computational overhead ($\sim 1M$ parameters, < 11 GFLOPs) and minimal latency. Although it does not achieve the highest absolute detection scores in every degradation setting compared to ESRGAN or SwinIR, it achieves a practical compromise between downstream accuracy and computational budget.

These results indicate that isolated super-resolution preprocessing cannot fully resolve domain discrepancies in real-world weld inspection; however, model selection remains crucial for practical deployment. In contrast to parameter-heavy models that generate spurious textures or introduce significant computational bottlenecks, the proposed **SATSR** approach preserves essential structural information with minimal latency, making it a



(a) Ground Truth (HR) (b) YOLO on SRCNN (c) YOLO on EDSR

(d) YOLO on SRGAN (e) YOLO on SwinIR (f) YOLO on SATSR (Ours)

Fig. 3 Qualitative detection results illustrate the effect of super-resolution on YOLOv11 performance under severe domain shift (Scenario 8). Standard and GAN-based SR methods (**b-e**) produce false negatives due to over-smoothing or spurious artifacts. In contrast, the proposed SATSR (**f**) maintains structural integrity, allowing the detector to accurately localize the defect, consistent with the high-resolution reference (**a**).

Table 13 Ablation study of the architectural components in SATSR, showing super-resolution reconstruction metrics and downstream weld defect detection performance.

Configuration	Blocks	3×3 Conv	AvgPool	Params (M)	PSNR (dB)	SSIM	mAP_{50}	$mAP_{50:95}$
Ablation (4 Blocks)	4	✓	✓	0.665	24.70	0.704	0.4823	0.1974
Ablation (12 Blocks)	12	✓	✓	1.328	24.37	0.690	0.5051	0.2040
Ablation (w/o AvgPool)	8	✓	×	0.997	24.84	0.709	0.4599	0.1827
Ablation (w/o 3×3 Conv)	8	×	✓	0.701	24.57	0.692	0.4670	0.1956
SATSR (Default)	**8**	✓	✓	**0.997**	**24.40**	**0.692**	**0.5028**	**0.2029**

well-suited candidate among the evaluated models for resource-constrained pipelines where computational efficiency and structural stability are prioritized. Addressing the residual domain discrepancy under severe degradation conditions underscores the potential of integrating super-resolution and detection within an end-to-end, domain-adaptive framework.

Figure 3 presents a representative challenging defect instance from the cross-domain evaluation to qualitatively illustrate how restoration artifacts influence downstream localization. Standard regression-based super-resolution (SR) methods, such as SRCNN, EDSR, and SwinIR, tend to over-smooth the low-contrast "Stain" defect, leading to a false negative detection by YOLOv11. Meanwhile, SRGAN introduces spurious high-frequency artifacts that mislead the detector into predicting an incorrect class ("Discontinuity"). In this specific instance, SATSR preserves critical boundary contours without hallucinatory textures, enabling the detector to correctly recognize and classify the defect, albeit with a slight bounding-box shift caused by the domain gap. While these visual observations illustrate SATSR's ability to retain delicate structural contours in low-contrast regions, they represent local topological behavior rather than uniform statistical superiority across the entire dataset, which remains constrained by cross-domain discrepancies as evidenced in Table 12.

Ablation study on SATSR architectural components

To investigate the individual contributions of the architectural modifications introduced into SATSR, a comprehensive ablation study was conducted. Specifically, we evaluated the effect of varying the depth of SAT blocks (4, 8, and 12 blocks), removing the localized 3×3 convolutional input projection, and omitting the spatial average pooling operation that replaces Deformable Token Aggregation (DTA). The resulting models were evaluated both on super-resolution reconstruction fidelity and downstream weld defect detection performance, as summarized in Table 13.

The ablation results reveal two critical insights regarding task-driven super-resolution. First, omitting average pooling (*w/o AvgPool*) results in a substantial degradation in downstream detection performance (mAP_{50} drops by

4.29% from 0.5028 to 0.4599, and mAP_{50-95} declines from 0.2029 to 0.1827), despite exhibiting higher pixel-level PSNR. This demonstrates that average pooling suppresses high-frequency noise and stabilizes geometric boundaries, yielding feature representations more conducive to downstream object detection. Second, eliminating the shallow 3×3 convolution (*w/o* 3×3 *Conv*) decreases mAP_{50} by 3.58% (0.4670 vs. 0.5028), confirming that injecting local inductive bias is essential for capturing fine edge gradients before global transformer aggregation. Finally, scaling the architecture to 12 blocks incurs a 33% parameter overhead (1.328 M vs. 0.997 M) while yielding only marginal gains (+0.0023 in mAP_{50}), whereas reducing it to 4 blocks causes a noticeable drop to 0.4823. Consequently, the default configuration with 8 blocks, average pooling, and 3×3 convolutional projection achieves the most favorable Pareto balance between parameter footprint and defect localization accuracy.

Discussion

The findings demonstrate that single-image super-resolution is an effective strategy for mitigating the performance drop caused by low-resolution imaging in weld defect detection, although its effectiveness is closely tied to the nature of the image degradation. As expected, the detector reached its best overall performance on HR images, where fine defect boundaries and weak textures are fully preserved. Once the images were degraded, detection quality dropped, with the magnitude of the drop depending on both the degradation process and the scale factor. Synthetic LR images were easier to process than real LR images, confirming that practical, non-ideal degradations introduce a more complex and less predictable appearance shift.

A second key observation is that improved perceptual quality does not always guarantee improved detection performance. While some SR models reconstructed visually sharper images, their detection gains were occasionally modest. Conversely, models such as EDSR, SwinIR, and particularly SATSR, produced competitive detection results by effectively restoring task-relevant structures. However, this advantage was not preserved uniformly in the real-LR experiments, suggesting that detector performance is governed more by the accurate reconstruction of defect morphology than by perceptual fidelity alone.

The cross-domain experiments further highlight the limitations of training solely within a single resolution regime. A detector trained on HR images struggled when directly applied to LR images, and models trained on synthetic LR images generalized poorly to real LR data. Furthermore, applying SR independently to both domains did not fully bridge this gap. These findings indicate that resolution mismatch and degradation mismatch both contribute heavily to domain shift. For practical inspection systems, this implies that SR should ideally be integrated with domain adaptation, mixed-resolution training, or joint end-to-end learning architectures rather than being employed as an isolated preprocessing step.

Class-wise analysis also reveals that not all defect types respond equally to degradation and SR reconstruction. Discontinuities remained relatively stable across many settings, whereas pores and deposits were highly sensitive to reduced resolution. This pattern aligns with the fact that some defect categories possess strong structural cues, while others rely on finer textures and local contrast. Consequently, SR-based inspection systems must be evaluated on a per-defect basis rather than relying exclusively on aggregated metrics.

Beyond these class-specific variations, a broader inspection of the quantitative findings reveals that overall detection metrics remain moderate across several evaluated regimes, notably with an upper baseline mAP-50 of 0.66 on HR images and constrained performance in cross-domain scenarios. Several interrelated physical and algorithmic factors account for these values. First, weld surface inspection under the LoHi-WELD benchmark entails non-uniform metallic reflections, complex weld seam ripples, and significant inter-class visual overlap. Tiny defects such as pores lack pronounced texture cues, while low-contrast stains share diffuse boundaries with benign thermal discolourations. Second, severely downsampled resolutions fundamentally eliminate high-frequency edge gradients, causing severe localization ambiguity for small defect instances. Third, our study purposely deployed YOLOv11-nano to mirror resource-constrained edge-computing deployments; while offering high throughput and compact memory footprint, lightweight architectures inherently trade parameter capacity

and fine-grained representation capability. From a practical applicability standpoint, these characteristics indicate that standalone super-resolution combined with lightweight detection is best suited as a high-speed preliminary screening tool (coarse defect filtering) within industrial pipelines, alerting operators to suspicious weld regions, rather than serving as a fully autonomous decision-maker without human-in-the-loop verification.

The present investigation adopts a decoupled SR-then-detection framework as a pragmatic baseline to systematically isolate and benchmark the impact of single-image restoration across standardized degradation regimes. While this modular design facilitates independent stage evaluation without inter-task gradient interference, it inherently lacks task-driven feedback during the feature restoration phase. This separation explains why conventional pixel-level or perceptual restoration metrics (PSNR, SSIM, and LPIPS) frequently diverge from downstream detection gains (mAP): generic restoration objectives prioritize global texture consistency and pixel fidelity, often smoothing out the localized high-frequency edge gradients that object detectors rely upon for bounding-box regression. Particularly for thin, elongated, and directional defect morphologies (such as fine cracks, lack of fusion, and undercuts), spatial-domain restoration alone can introduce boundary ambiguity. To transcend these limitations, future research can naturally transition toward end-to-end joint super-resolution and detection architectures, wherein the restoration subnet is supervised directly by task-specific localization losses. Moreover, coupling such joint frameworks with frequency-aware mechanisms (e.g., wavelet-guided transformers[10] for preserving high-frequency subbands) and linear-complexity state-space models (e.g., Mamba backbones[11]) offers a promising paradigm to bridge the persistent synthetic-to-real domain gap without incurring prohibitive computational overhead.

Overall, the experiments validate that SR can successfully recover a significant portion of lost detection performance under controlled synthetic degradations, though the benefits are more constrained for real low-resolution images. The most practical interpretation is that SR is highly valuable when it preserves edge information and defect morphology without amplifying artifacts. To build reliable vision systems for real-world weld inspection, combining efficient SR models like SATSR with resolution-aware training and robust detector designs remains the most promising strategy.

Despite these encouraging findings, several limitations of the present study should be noted. Although our cross-domain evaluations expose the persistent synthetic-to-real gap, the current empirical analysis relies on a single benchmark dataset (LoHi-WELD) and evaluates a single compact detector backbone (YOLOv11-nano). While YOLOv11-nano represents a pragmatic baseline for resource-constrained edge inspection, detector responses to super-resolved features may vary with other one-stage, two-stage, or transformer-based detection heads. Furthermore, practical manufacturing environments introduce diverse welding processes, illumination fluctuations, and sensor variations that extend beyond the conditions captured in current benchmarks. Therefore, conclusions regarding broad industrial deployment should remain qualified until these restoration strategies are systematically validated across multi-sensor setups and heterogeneous detection frameworks.

Conclusion and future works

This study systematically evaluated the integration of single-image super-resolution into a YOLOv11-based weld defect detection pipeline, addressing the challenges of low-resolution and degraded industrial imaging. Through a comprehensive comparison of visual quality, computational complexity, and detection recovery, the proposed Selective Aggregation Transformer for Super-Resolution (SATSR) emerged as an efficient and competitive solution, demonstrating a well-balanced trade-off between detection accuracy and computational overhead, particularly in edge-preserving reconstruction under constrained computational budgets. By leveraging an efficient spatial-reduction attention mechanism, SATSR maintains a highly lightweight profile (approximately $\sim 1M$ parameters) while successfully preserving critical defect morphology–effectively avoiding the over-smoothing inherent to standard convolutional networks and the spurious artifacts common in GAN-based approaches.

Despite these advancements, cross-domain evaluations revealed a substantial performance drop when models trained on synthetically degraded data were applied to real-world low-resolution images. Crucially, while applying SR as an independent preprocessing step could not completely close this domain gap, qualitative observations indicated that SATSR's boundary preservation aids the detector in recovering subtle structural contours under challenging low-contrast conditions. Nevertheless, these findings demonstrate that treating image reconstruction and object detection as strictly decoupled processes is insufficient for perfect generalization, as independent SR networks optimize for perceptual quality rather than task-specific feature alignment.

To address this fundamental limitation, future research will focus on developing a joint super-resolution and detection architecture. By integrating the lightweight SATSR backbone directly with the object detection modules into a unified, end-to-end trainable network, the reconstruction process can be explicitly driven by detection-oriented loss functions rather than solely by perceptual metrics. We anticipate that this joint optimization strategy will encourage robust feature-level domain alignment, bridging the gap between synthetic and real-world degradations and paving the way for highly reliable automated industrial inspection systems.

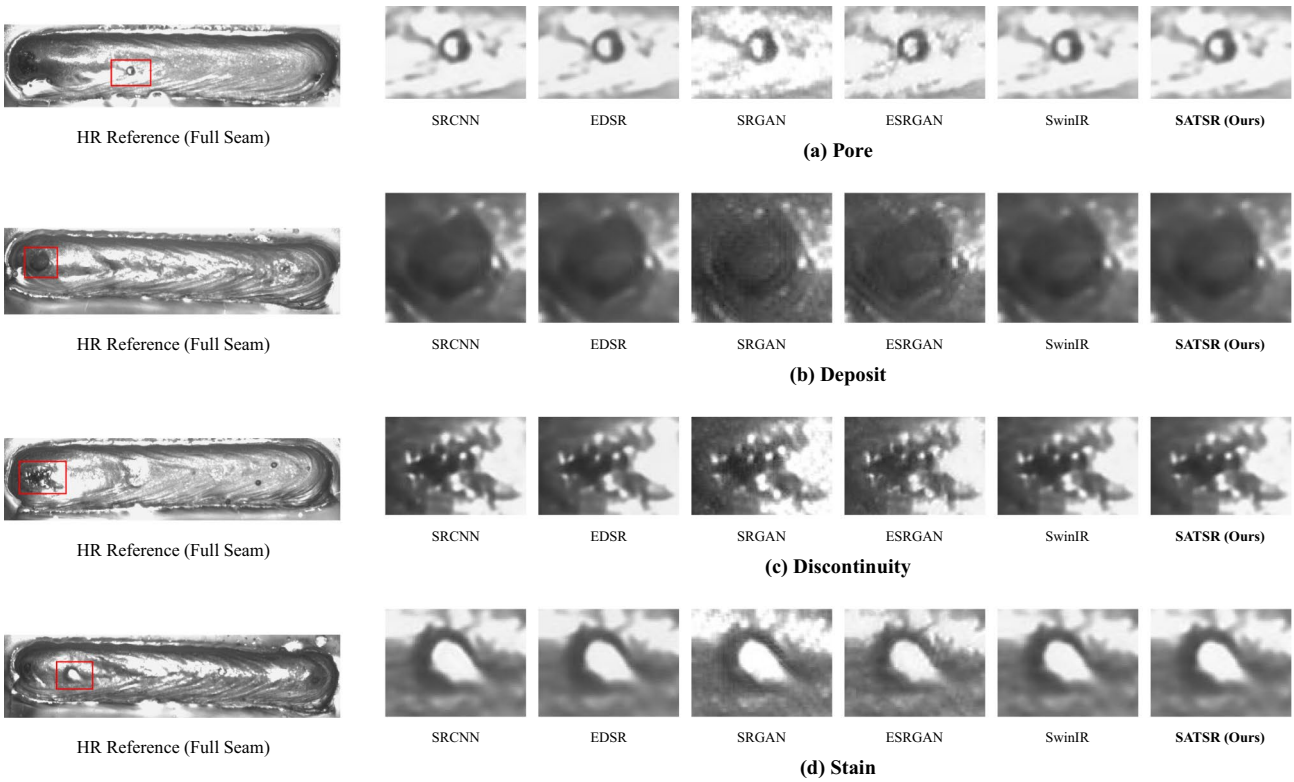


Fig. 4 Qualitative super-resolution comparison across all four weld defect classes. For each category, the left panel presents the complete high-resolution weld bead with a red bounding box marking the evaluated defect region. The right panels display localized zoom-in patches reconstructed by SRCNN, EDSR, SRGAN, ESRGAN, SwinIR, and our proposed SATSR. Across void boundaries (pore), elevated pools (deposit), irregular fractures (discontinuity), and surface discolorations (stain), SATSR demonstrates competitive structural sharpness while preventing edge blurring and hallucination artifacts.

Appendix: Extended qualitative comparison across defect classes

To supplement the visual analysis in Section 3 (Figure 2), this appendix provides visual reconstruction fidelity across all four annotated weld defect categories in the LoHi-WELD dataset: *pore*, *deposit*, *discontinuity*, and *stain*. Figure 4 illustrates the full high-resolution (HR) weld bead alongside zoomed-in patches reconstructed by SRCNN, EDSR, SRGAN, ESRGAN, SwinIR, and our proposed SATSR.

Author contributions S.H.S. performed conceptualization, methodology, investigation, simulation, and interpretation of the results; prepared the original manuscript draft; and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, the authors used Google Gemini in order to refine the academic language, structure, and clarity of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.***39**, 1137–1149 (2016).
2. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,779–788 (2016).
3. Zhang, Y., Zhang, H., Huang, Q., Han, Y. & Zhao, M. DsP-YOLO: An anchor-free network with dspan for small object detection of multiscale defects. *Expert. Syst. with Appl.***241**, 122669 (2024).
4. Lu, M., Sheng, W., Zou, Y., Chen, Y. & Chen, Z. WSS-YOLO: An improved industrial defect detection network for steel surface defects. *Measurement***236**, 115060 (2024).
5. Yang, L., Gu, Y. & Feng, H. YOLOv8-DPE: a novel lightweight object detection network for weld defect detection. *Weld. World*1–21 (2025).
6. Luo, Y. et al. SFW-YOLO: A lightweight multi-scale dynamic attention network for weld defect detection in steel bridge inspection. *Measurement***253**, 117608 (2025).
7. Shi, Y., Zheng, S., Bian, M., Zhang, X. & Yang, L. DASS-YOLO: Improved yolov7-tiny with attention-guided shape awareness and dysnakeconv for spray code defect detection. *Symmetry***17**, 906 (2025).
8. Zhong, Z. et al. Structure-preserving surface defect detection for flexible integrated circuit substrates. *Image Vis. Comput.* 106007 (2026).
9. He, L., Zheng, L. & Xiong, J. FMV-YOLO: a steel surface defect detection algorithm for real-world scenarios. *Electronics***14**, 1143 (2025).
10. Chen, Z., Zhang, H., Fan, M. & Li, N. Wavelet-guided dynamic transformer for high-fidelity super-resolution in terahertz imaging systems. *IEEE Sensors J.*https://doi.org/10.1109/JSEN.2026.3730211 (2026).
11. Zhao, F. et al. Mamba-based super-resolution and semi-supervised yolov10 for freshwater mussel detection using acoustic video camera: A case study at lake izunuma, japan. *Ecol. Informatics***90**, 103324. https://doi.org/10.1016/j.ecoinf.2025.103324 (2025).
12. Xu, T. et al. CDTrans: Cross-domain transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 6687–6697 (2021).
13. Liang, J., Hu, D. & Feng, J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6028–6039 (2020).
14. Block, S. B., da Silva, R. D., Lazzaretti, A. E. & Minetto, R. LoHi-WELD: A novel industrial dataset for weld defect detection and classification, a deep learning study, and future perspectives. *IEEE Access***12**, 77442–77453 (2024).
15. Dong, C., Loy, C. C., He, K. & Tang, X. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis Mmachine Intell.***38**, 295–307 (2015).
16. Lim, B., Son, S., Kim, H., Nah, S. & Lee, K.M. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*,1132–1140 (IEEE, 2017).
17. Ledig, C. et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* 105–114 (IEEE, 2017).
18. Wang, X. et al. Esrgan: Enhanced super-resolution generative adversarial networks. In *European conference on computer vision* 63–79 (Springer, 2018).
19. Liang, J. et al. SwinIR: Image restoration using swin transformer. In *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)* 1833–1844 (IEEE, 2021).
20. Tran, D. P. et al. SAT: Selective Aggregation Transformer for image super-resolution. In *2026 IEEE conference on computer vision and pattern recognition (CVPR)* (IEEE, 2026).
21. He, Y., Song, K., Meng, Q. & Yan, Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Trans. Instrum. Meas.***69**, 1493–1504 (2019).
22. Diaz-Cano, I. et al. Automated fillet weld inspection based on deep learning from 2d images. *Appl. Sci.***15**, 899 (2025).

23. Contreras Ortiz, A., Santiago, R. R., Hernandez, D. E. & Lopez-Montiel, M. Multiclass evaluation of vision transformers for industrial welding defect detection. *Math. Comput. Appl.***30**, 24 (2025).
24. Tokime, R. B., Maldague, X. & Perron, L. Automatic defect detection for x-ray inspection: Identifying defects with deep convolutional network. In *Proceedings of the Canadian Institute for Non-destructive Evaluation (CINDE), Edmonton, AB, Canada* 18–20 (2019).
25. Pan, H., Pang, Z., Wang, Y., Wang, Y. & Chen, L. A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects. *Ieee Access***8**, 119951–119960 (2020).
26. Kumaresan, S., Aultrin, K. J., Kumar, S. & Anand, M. D. Transfer learning with cnn for classification of weld defect. *Ieee Access***9**, 95097–95108 (2021).
27. Ajmi, C., Zapata, J., Elferchichi, S. & Laabidi, K. Advanced faster-RCNN model for automated recognition and detection of weld defects on limited x-ray image dataset. *J. Nondestruct. Eval.***43**, 14 (2024).
28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, 586–595 (IEEE, 2018).
29. Ding, K., Ma, K., Wang, S. & Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine intelligence***44**, 2567–2581 (2020).
30. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.***20**, 209–212 (2012).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Seyed Hamid Safavi[1]

✉ Seyed Hamid Safavi
h.safavi@uma.ac.ir

[1] Faculty of Advanced Technologies, University of Mohaghegh Ardabili, Ardabil, Iran