remote sensing

MDPI

Article

WISE-Net: Wavelet-Guided Illumination Self-Supervised Enhancement for Low-Light Remote Sensing Images

Hong Dai [1,2], Hui Zhao [1,*], Shuai Wang [1], Shaowei Li [1,2], Ze Jiang [1,2] and Xuewu Fan [1]

[1] Space Optics Lab, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710019, China; daihong21@mails.ucas.ac.cn (H.D.); wangshuai3@opt.ac.cn (S.W.); lishaowei24@mails.ucas.ac.cn (S.L.); jiangze22@mails.ucas.ac.cn (Z.J.); fanxuewu@opt.ac.cn (X.F.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: zhaohui@opt.ac.cn

Highlights

What are the main findings?

- WISE-Net separates wavelet-guided low-frequency enhancement from self-supervised refinement without paired clean training targets.
- The evaluated real and external paired datasets provide complementary evidence of visibility recovery and artifact suppression.

What are the implications of the main findings?

- Low-frequency modulation provides an inspectable brightness enhancement path for grayscale low-light imagery.
- Refinement trained on the enhanced distribution addresses noise exposed by brightness lifting; broader sensor transfer remains to be validated.

Abstract

Low-light remote sensing images are often affected by photon-limited exposure, signal-dependent noise, uneven contrast, and weak or missing structures. Where denoising may not correct underexposure, increasing the brightness can reveal noise in high-resolution grayscale imagery without registered normal-light references. We propose WISE-Net (Wavelet-guided Illumination Self-supervised Enhancement Network), a two-stage framework that separates these tasks. Stage 1 predicts a brightness-aware gain for the low-frequency Haar component and reuses the input high-frequency sub-bands during reconstruction. With Stage 1 kept frozen, Stage 2 then performs masked self-supervised blindspot refinement using the distribution of the enhanced image. We evaluate the method on real 16-bit grayscale images from a ground-based time-delay-and-integration intensified complementary metal oxide semiconductor (TDI-ICMOS) system and on two external paired low-light remote sensing test sets. Results from no-reference and reference-based metrics together with local visual comparisons indicate that the two stages support visibility recovery and reduce enhancement-exposed artifacts.

Keywords: low-light remote sensing; self-supervised learning; wavelet transform; image enhancement; TDI-ICMOS imaging

Check for updates

Academic Editor: Fa-Ming Fang

Received: 20 October 2026
Revised: 19 November 2026
Accepted: 23 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Low-light remote sensing supports nighttime Earth observation, weak signal surveillance, and long-range imaging. Such images often contain spatially connected structures

such as roads, building groups, and vegetation boundaries. Photon scarcity and sensor noise can leave these structures underexposed, with weak local contrast and unstable highlights. Increasing brightness may expose noise and blur, while aggressive denoising can remove details needed for interpretation [1–3]. Therefore, enhancement must improve visibility without unnecessarily compromising structure.

Existing methods differ in their supervision as well as their assumptions about image formation. Retinex-based approaches separate illumination from reflectance to model spatially varying brightness [4–9], while supervised enhancement and restoration models use paired targets with autoencoder, RAW-domain, progressive, SNR-aware, or transformer architectures [1,10–15]. Semi-supervised approaches combine paired fidelity learning with perceptual objectives [16]. Although such models can perform well when the training and test data are closely matched, registered normal-light references are difficult to obtain for real remote sensing scenes. Unpaired and adversarial approaches relax the pairing requirement [17]. Zero-reference methods optimize curve, illumination, or spatial constraints from the input itself, and include architecture search and joint enhancement denoising models [18–23]. These lines of work motivate self-supervised enhancement, while leaving open how to coordinate illumination recovery and artifact suppression in real 16-bit grayscale push-broom imagery.

For real nighttime observations, collecting registered normal-light targets is often impractical. Recent work has combined paired synthetic data with real nighttime scenes for reference-based and no-reference evaluation [3], while other studies have explored unpaired enhancement and self-supervised denoising [17,24–26]. These approaches address complementary limitations, as denoising may not correct severe underexposure and unconstrained enhancement may amplify noise; this motivates a framework that coordinates both operations without requiring clean targets.

We propose WISE-Net (Wavelet-guided Illumination Self-supervised Enhancement Network), a fully self-supervised two-stage framework for low-light remote sensing imagery. Stage 1 decomposes the input into Haar sub-bands and modulates only the low-frequency component. Its objective combines dual-view consistency with regularization for relative contrast, dark region gain, bright region protection, gain smoothness, overexposure, and gradient exclusion. With Stage 1 kept frozen, Stage 2 applies self-supervised blindspot refinement (SBR) to the enhanced image's distribution in order to suppress noise and local artifacts exposed by brightness lifting.

The main contributions of this work are summarized as follows:

(1) We propose WISE-Net, a fully self-supervised two-stage framework for real 16-bit grayscale low-light remote sensing images. WISE-Net separates visibility recovery from post-enhancement refinement, and does not require clean reference images or unpaired normal-light images.
(2) We configure a wavelet-guided illumination module that predicts a brightness-aware gain for the low-frequency Haar component and reuses the original high-frequency sub-bands during reconstruction; the contribution lies in this illumination modulation design within the self-supervised pipeline.
(3) We adapt blindspot learning to the Stage 1 enhanced image distribution through SBR. Blindspot learning is an established principle, which we introduce here in order to combine masked pixel reconstruction with structural regularization after illumination enhancement.
(4) We curate an in-house dataset of real 16-bit grayscale low-light TDI-ICMOS images acquired with a ground-based camera-level equivalent push-broom configuration. We use three test subsets that vary in image width. We also evaluate the trained model on

two public paired low-light remote sensing test sets in order to assess cross-dataset performance.

In the remainder of this paper, Section 2 reviews low-light enhancement and self-supervised denoising; Section 3 presents WISE-Net; Section 4 describes the experiments and ablations; and Section 5 concludes the study.

1.1. Related Work

1.1.1. Low-Light Image Enhancement

Low-light enhancement methods differ in their mapping assumptions and supervision. Direct value mapping methods such as histogram equalization, gamma correction, and contrast-limited adaptive histogram equalization redistribute gray levels using fixed functions or histogram statistics. They are computationally simple and can improve contrast without training data, but do not clearly separate illumination from noise; strong remapping can also saturate highlights or expose noise in dark regions.

Retinex methods represent an image through illumination and reflectance. Classical single-scale and multiscale methods estimate illumination by spatial filtering, while learned decomposition methods predict components with data-driven constraints [4–9]. Adjusting illumination can accommodate spatially uneven exposure while retaining reflectance variation. However, the decomposition is not unique, and its assumptions and regularization can affect halo formation, noise amplification, and detail preservation.

Supervised models learn a mapping from low-light images to paired normal-light or long-exposure targets. Representative approaches include LLNet, RAW-domain learning, progressive enhancement, and SNR-aware models, alongside more general restoration architectures such as Restormer [1,10–15]. Semi-supervised methods additionally balance paired fidelity learning with perceptual quality objectives [16]. Paired supervision supplies a direct fidelity target and supports joint brightness and detail restoration. Its practical limitation is the need for registered targets, while performance under a different sensor or degradation distribution depends on transfer from the training data.

Unpaired adversarial methods such as EnlightenGAN match enhanced outputs to a normal-light image distribution without requiring scene-aligned pairs [17]. They reduce acquisition constraints and can learn a useful appearance prior; however, distribution matching does not guarantee scene-specific fidelity or radiometric accuracy. The choice of normal-light training images can also influence the resulting appearance.

Zero-reference methods avoid normal-light targets by constraining properties such as exposure, spatial consistency, or illumination smoothness. Zero-DCE and Zero-DCE++ learn enhancement curves, RUAS uses Retinex-inspired unrolling and architecture search, SCI and SCI++ learn illumination correction, and ZERO-IG coordinates illumination-guided denoising and enhancement [18–23]. These methods reduce dependence on reference data, but their constraints must balance visibility, noise, and structural detail. This balance motivates evaluating them on high-resolution grayscale push-broom images alongside the proposed self-supervised formulation.

1.1.2. Self-Supervised Denoising

Self-supervised restoration is useful when clean references are unavailable. Noise2Noise learns from independent noisy observations, while Noise2Void predicts masked pixels from their surrounding context [24,25]. Neighbor2Neighbor forms training pairs from a single noisy observation [26], and SN2N uses spatial redundancy to construct self-supervised views for microscopy denoising [27]. These methods focus on noise estimation or suppression, but do not address visibility loss in severely underexposed images. WISE-Net uses self-supervised views to optimize Stage 1 and then applies blindspot refinement to the

enhanced distribution; its contribution is this coordination of illumination modulation and refinement, rather than a new blindspot learning principle.

1.1.3. Motivation for WISE-Net

Together, these methods highlight a practical tension: enhancement recovers visibility, but also exposes noise and blur, while denoising can suppress noise without necessarily correcting the underlying underexposure. WISE-Net addresses these two tasks sequentially, first modulating low-frequency illumination and then refining the enhanced result with a blindspot model. This ordering gives each stage a distinct role.

The distinction is the allocation and coordination of established mechanisms, not the invention of wavelet decomposition, spatial resampling, or blindspot learning; Table 1 summarizes these roles.

Table 1. Mechanism-level comparison and scope of the WISE-Net contribution; the entries describe processing roles.

Mechanism or Representative Work	Established Role	Use and Contribution Boundary in WISE-Net
Haar wavelets [28]	Separate coarse intensity variation from directional detail.	Apply brightness-aware learned modulation to the low-frequency component, with measured high-frequency sub-bands reused in reconstruction.
Neighbor2Neighbor and SN2N [26,27]	Derive self-supervised observations from spatial redundancy.	Use diagonal resampling for Stage 1 dual-view training.
Noise2Void [24]	Learn contextual prediction from masked noisy observations.	Apply masked refinement to the frozen Stage-1 output distribution together with structural regularization.
ZERO-IG [23]	Coordinate illumination-guided denoising and enhancement.	Use low-frequency gain modulation followed by separately trained refinement.

2. Materials and Methods

Figure 1 summarizes WISE-Net. Given a noisy low-light remote sensing image I_{noisy}, Stage 1 performs wavelet-guided illumination modulation and Stage 2 refines the enhanced image through self-supervised learning. This ordering separates visibility recovery from suppression of noise and local artifacts exposed by brightness lifting.

The overall pipeline is written as follows:

$$\hat{I}_{\text{out}} = f_{\text{refine}}\left(f_{\text{illum}}(I_{\text{noisy}})\right) \tag{1}$$

where $f_{\text{illum}}(\cdot)$ denotes the Stage 1 wavelet-guided illumination modulation network and $f_{\text{refine}}(\cdot)$ denotes the Stage 2 SBR network. Both stages are trained without clean reference images.

2.1. Stage 1: Wavelet-Guided Illumination Modulation

2.1.1. Haar Wavelet Decomposition

We decompose the input image $I \in \mathbb{R}^{H\times W}$ with a single-level Haar discrete wavelet transform [28]:

$$\text{DWT}(I) = \{I_{LL}, I_{LH}, I_{HL}, I_{HH}\}. \tag{2}$$

The low-frequency approximation I_{LL} represents large-scale intensity variation, while I_{LH}, I_{HL}, and I_{HH} contain directional changes associated with edges, textures, and high-frequency noise. WISE-Net modulates only I_{LL} and reuses the input high-frequency sub-bands in the inverse transform. Thus, the model does not apply a learned gain directly

to those bands; while this design can limit direct band amplification, it cannot recover structures that are absent from the input.

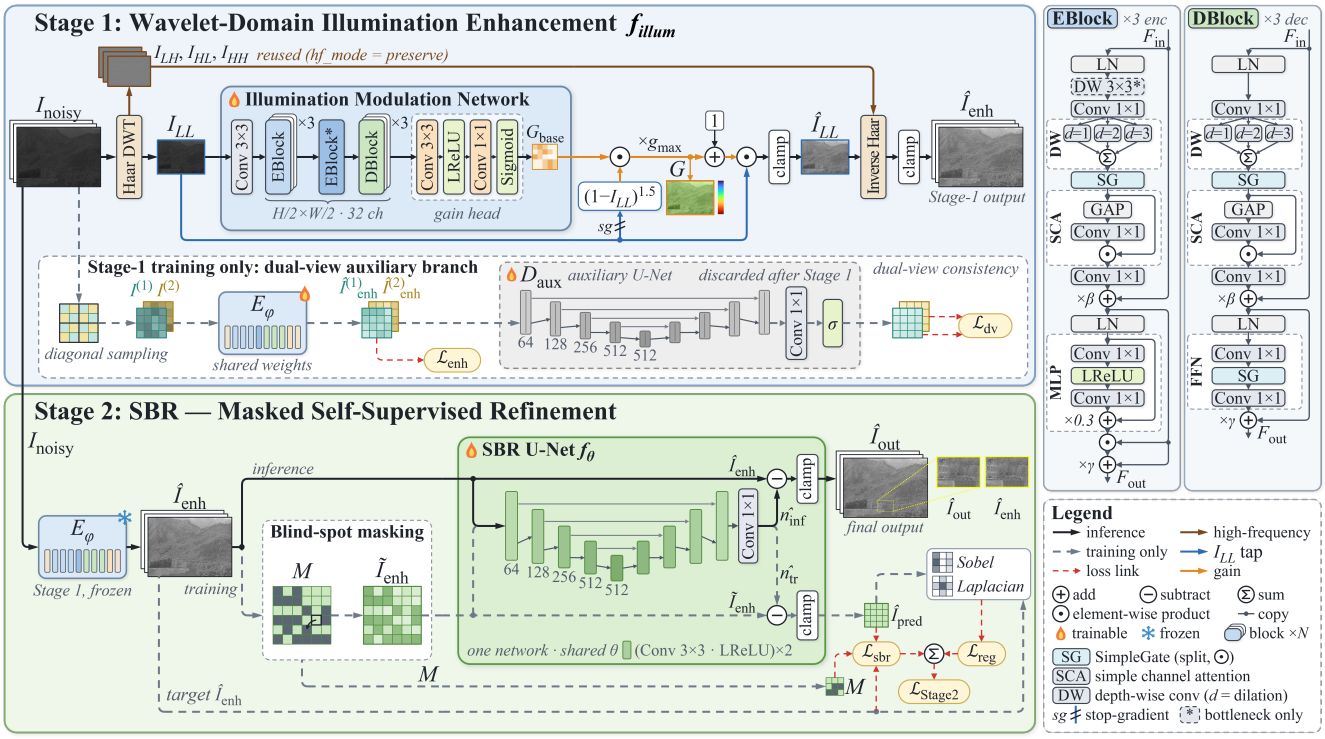

Figure 1. Overall architecture and training/inference paths of WISE-Net. Stage 1 performs wavelet-guided illumination modulation. The low-frequency component is enhanced, while the original high-frequency sub-bands are reused in the inverse Haar transform without direct modulation. The auxiliary denoising branch is used only during Stage 1 training, and is discarded after optimization. Stage 2 freezes Stage 1. During training, random masking and neighbor replacement produce the masked enhanced input for the SBR network, while the original enhanced image remains the target. During inference, masking is disabled and the complete Stage 1 output is processed by the same SBR parameters.

2.1.2. Low-Frequency Illumination Modulation Network

The Stage 1 illumination module is illustrated in Figure 1 (Stage 1).

The Stage 1 network takes I_{LL} as input and predicts the enhanced component $\hat{I}_{LL}$ and a spatially varying gain map G. Its encoder–bottleneck–decoder uses multi-scale dilated depthwise convolutions and channel-mixing blocks. A lightweight head maps the decoded features to a base gain map $G_{\text{base}} \in [0, 1]$.

The base gain G_{base} is combined with a brightness-aware term that assigns stronger enhancement to darker regions:

$$G = G_{\text{base}} \cdot (1 - I_{LL})^{p} \cdot G_{\max} \tag{3}$$

where p controls the sensitivity to input brightness and $G_{\max}$ limits the admissible gain. Both are fixed throughout training. The enhanced low-frequency component is then obtained by

$$\hat{I}_{LL} = \text{clamp}(I_{LL} \cdot (1 + G), 0, 1). \tag{4}$$

This formulation increases gain in darker regions while limiting unnecessary amplification in already bright areas.

2.1.3. Wavelet Reconstruction

After low-frequency enhancement, the final Stage 1 output is reconstructed using the inverse Haar wavelet transform:

$$\hat{I}_{\text{enh}} = \text{IDWT}(\hat{I}_{LL}, I_{LH}, I_{HL}, I_{HH}). \tag{5}$$

Reusing the input high-frequency sub-bands separates the learned illumination update from direct high-frequency modulation. The reconstructed image is still clipped to the valid intensity range, which describes the reconstruction path.

The implemented Haar analysis uses a factor of $1/4$ for each sum or signed sum over a 2×2 block; synthesis uses the corresponding signed sums without further scaling. Let Y denote the reconstruction before output clipping and let $\delta_{LL} = \hat{I}_{LL} - I_{LL}$. Exact arithmetic then gives

$$Y = I + U(\delta_{LL}), \qquad \|Y - I\|_2^2 = 4\|\delta_{LL}\|_2^2, \tag{6}$$

where U repeats each coefficient over its corresponding 2×2 block. Thus, changing only the low-frequency coefficients creates a block-wise constant update, and does not introduce any additional high-frequency coefficient changes before clipping. This is an algebraic property of the implemented transform, not a guarantee of unbiased restoration relative to an unknown clean image. With $C(Y) = \min(1, \max(0, Y))$ and clipping residual $r = C(Y) - Y$, the output satisfies

$$\|C(Y) - Y\|_2 = \|r\|_2, \qquad \sum_{b \in \{LH,HL,HH\}} \|\text{DWT}_b(C(Y)) - I_b\|_2^2 \leq \tfrac{1}{4}\|r\|_2^2, \tag{7}$$

where the bound follows from the mutually orthogonal signed-sum analysis rows, each with squared norm $1/4$, to ensure even image dimensions. Unchanged coefficients retain input high-frequency noise as well as measured structure. The clipping residual and disagreement with an unknown clean high-frequency signal are distinct sources of error, and must not be interpreted as numerical inverse-transform error.

2.1.4. Self-Supervised Dual-View Construction for Stage 1

Stage 1 is trained without clean targets. We generate two content-consistent noisy views from each raw image by diagonal resampling: within every 2×2 pixel unit, the top-left and bottom-right pixels are averaged for one view and the top-right and bottom-left pixels for the other. The views exploit local scene similarity and different pixel samples; independence of their noise is an approximation that can be affected by spatially correlated sensor noise. Bilinear interpolation restores each view to the training patch size; we denote them by I_L and I_R, respectively.

This construction draws on spatially redundant self-supervised denoising and supplies the two views used to optimize Stage 1. The contribution of WISE-Net is the enhancement framework built around these views.

2.2. Stage 1 Loss Functions

The Stage 1 objective combines dual-view self-supervision with enhancement regularization:

$$\mathcal{L}_{\text{Stage1}} = \alpha \mathcal{L}_{\text{dv}} + \beta \mathcal{L}_{\text{enh}} \tag{8}$$

where $\mathcal{L}_{\text{dv}}$ is the dual-view self-supervised consistency term and $\mathcal{L}_{\text{enh}}$ is the enhancement regularization term. The coefficients α and β balance self-supervised consistency and illumination regularization.

Dual-view self-supervised consistency. During Stage 1 training, both views are passed through the shared Stage 1 enhancement network. The resulting enhanced views are then processed by a training-only auxiliary denoising U-Net. Its purpose is to provide a stable denoising representation for the dual-view consistency objective; it is not retained as a third inference output. We enforce consistency between the two final outputs by

$$\mathcal{L}_{\mathrm{dv}} = \left\| \hat{I}_L^{\mathrm{final}} - \hat{I}_R^{\mathrm{final}} \right\|_1. \tag{9}$$

The views provide a consistency signal through approximately shared local content and different pixel samples. Resampling and interpolation do not establish independent noise realizations. Spatially correlated noise can remain shared between the views, while fine structures can produce content differences; the consistency term is a training constraint.

Enhancement regularization. The regularizer is defined as

$$\mathcal{L}_{\mathrm{enh}} = w_{\mathrm{rc}}\mathcal{L}_{\mathrm{rc}} + w_{\mathrm{mg}}\mathcal{L}_{\mathrm{mg}} + w_{\mathrm{bp}}\mathcal{L}_{\mathrm{bp}} + w_{\mathrm{sm}}\mathcal{L}_{\mathrm{sm}} + w_{\mathrm{ov}}\mathcal{L}_{\mathrm{ov}} + w_{\mathrm{ex}}\mathcal{L}_{\mathrm{ex}}, \tag{10}$$

where the weights w_{rc}, w_{mg}, w_{bp}, w_{sm}, w_{ov}, and w_{ex} are fixed hyperparameters that balance the six terms. The cumulative ablation disables the indicated terms while keeping the remaining weights unchanged.

The six terms play the following roles:

$$\begin{aligned} \mathcal{L}_{\mathrm{rc}} = {} & \left\| \nabla_x \log(\hat{I}_{LL} + \epsilon) - \nabla_x \log(I_{LL} + \epsilon) \right\|_1 \\ & + \left\| \nabla_y \log(\hat{I}_{LL} + \epsilon) - \nabla_y \log(I_{LL} + \epsilon) \right\|_1. \end{aligned} \tag{11}$$

Preservation of relative contrast maintains local brightness relationships after enhancement:

$$\mathcal{L}_{\mathrm{mg}} = \frac{\sum M_{\mathrm{dark}} \cdot \max\left(0, r_{\min} - \frac{\hat{I}_{LL}}{I_{LL}+\epsilon}\right)}{\sum M_{\mathrm{dark}} + \epsilon} \tag{12}$$

where $M_{\mathrm{dark}} = \sigma(k(\tau_{\mathrm{dark}} - I_{LL}))$ is a soft mask for the dark region, $r_{\min}$ denotes the required minimum relative gain, and k controls mask softness. The penalty discourages insufficient low-frequency gain:

$$\mathcal{L}_{\mathrm{bp}} = \mathrm{mean}\left(M_{\mathrm{bright}} \cdot \left| \frac{\hat{I}_{LL}}{I_{LL} + \epsilon} - 1 \right| \right) \tag{13}$$

where $M_{\mathrm{bright}} = \sigma\Big(k(I_{LL} - \tau_{\mathrm{bright}})\Big)$ is a soft mask for the already-visible regions. Protecting the bright region suppresses unnecessary gain departures from unity in these regions:

$$\mathcal{L}_{\mathrm{sm}} = \mathrm{mean}(|\nabla_x G|) + \mathrm{mean}(|\nabla_y G|). \tag{14}$$

Smoothness of the gain map suppresses abrupt spatial changes that would otherwise introduce artifacts:

$$\begin{aligned} \mathcal{L}_{\mathrm{ov}} = {} & \mathrm{mean}\Big(\max(0, \hat{I}_{\mathrm{enh}} - \tau_{\mathrm{sat}})^2\Big) \\ & + \mathrm{mean}\Big(\max(0, \hat{I}_{LL} - \tau_{\mathrm{sat}})^2\Big) \end{aligned} \tag{15}$$

where τ_{sat} denotes a fixed saturation threshold. The overexposure penalty discourages saturation in both the reconstructed image and the enhanced low-frequency component:

$$\mathcal{L}_{\mathrm{ex}} = \mathrm{Exclusion}(I_{LL}, G). \tag{16}$$

Following the exclusion principle used in unsupervised nighttime enhancement [29], the exclusion loss discourages coincident gradients in I_{LL} and G, thereby reducing gain transitions that track strong luminance edges.

More precisely, let A_s be average pooling downsampling to scale s, let D_d be the forward difference in direction $d \in \{x, y\}$, and set $q(t) = 2\sigma(t) - 1 = \tanh(t/2)$. The three-scale implementation is

$$\mathcal{L}_{\text{ex}} = \frac{1}{6} \sum_{s=0}^{2} \sum_{d \in \{x,y\}} \left[\text{mean} \left(q(D_d A_s I_{LL})^2 q(D_d A_s G)^2 \right) \right]^{1/4}. \tag{17}$$

Each summand lies in $[0, 1]$. For small differences, $q(t) \approx t/2$, so the penalty increases with spatial overlap of luminance and gain gradients. For a spatially constant input component, it vanishes irrespective of the gain gradient; therefore, it penalizes coincident transitions rather than enforcing vector orthogonality, hard gradient thresholding, or suppression of every edge. Its weight sets a soft trade-off with brightness and consistency objectives.

2.3. Stage 2: Enhancement-Oriented Self-Supervised Refinement

2.3.1. Architecture Overview

Figure 1 (Stage 2) shows the Stage 2 path. The trained Stage 1 network is frozen and used as a deterministic front end. Stage 2 then suppresses noise and local artifacts exposed by brightness lifting while retaining the recovered visibility. We refer to this component as self-supervised blindspot refinement (SBR). It receives a masked enhanced image $\tilde{I}_{\text{enh}}$ during training, and receives the complete Stage 1 output $\hat{I}_{\text{enh}}$ during inference.

2.3.2. SBR Network

SBR uses a U-Net-style encoder–decoder with residual prediction [30]. The encoder aggregates context at progressively coarser scales, while the decoder uses bilinear upsampling and skip connections to recover spatial detail. During training, the network predicts a residual $\hat{n}_{\text{tr}}$ from the masked image and forms the training prediction:

$$\hat{I}_{\text{pred}} = \tilde{I}_{\text{enh}} - \hat{n}_{\text{tr}}, \qquad \hat{n}_{\text{tr}} = f_\theta(\tilde{I}_{\text{enh}}). \tag{18}$$

During inference, masking is disabled and the complete Stage 1 output is refined directly:

$$\hat{I}_{\text{out}} = \hat{I}_{\text{enh}} - \hat{n}_{\text{inf}}, \qquad \hat{n}_{\text{inf}} = f_\theta(\hat{I}_{\text{enh}}), \tag{19}$$

with the same SBR network parameters used in both the training and inference paths.

2.3.3. Blindspot Masking Strategy

For training, a random mask selects pixels for replacement by sampled neighborhood values. The implementation samples nonzero coordinate offsets and clips the resulting coordinates to the image boundary. Clipping can map an offset back to the original boundary pixel, so replacement does not guarantee that every selected target value is excluded. SBR receives the corrupted image, and the reconstruction loss is evaluated only at selected positions. This training procedure encourages contextual prediction, but does not provide a strict architectural blindspot or guarantee noise-independent targets. At inference, masking is disabled and SBR processes the complete enhanced image.

2.3.4. Stage 2 Loss Function

Writing $X = \hat{I}_{\text{enh}}$, the masking operation is $\tilde{X}_i = (1 - M_i)X_i + M_i X_{\pi(i)}$, where $M_i \sim \text{Bernoulli}(p)$ and $\pi(i)$ is the clipped sampled neighbor coordinate. For N spatial

positions, $\mathbb{E}[\sum_i M_i] = pN$. Under mask draws independent of the replacement sampling, $\mathbb{E}\|\tilde{X} - X\|_2^2 = p\sum_i \mathbb{E}|X_{\pi(i)} - X_i|^2$. Increasing p increases both the expected number of supervised locations and the expected input perturbation. This can reduce available context for fine structures even though the loss normalizes by the number of selected pixels. While the relation explains this trade-off, it does not imply monotonic quality improvement or identify an optimal masking ratio.

The Stage 2 objective is

$$\mathcal{L}_{\text{Stage2}} = w_{\text{sbr}}\mathcal{L}_{\text{sbr}} + w_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{20}$$

where w_{sbr} and w_{reg} balance masked pixel reconstruction and structural preservation; both are fixed whenever Stage 2 is included.

The blindspot reconstruction term compares the training prediction with the unmasked Stage-1 output:

$$\mathcal{L}_{\text{sbr}} = \frac{\sum M \odot \left|\hat{I}_{\text{pred}} - \hat{I}_{\text{enh}}\right|}{\sum M + \epsilon} \tag{21}$$

where M is the blindspot mask.

To limit over-smoothing, let $E(\cdot)$ denote the Sobel gradient magnitude and let M_e select locations where $E(\hat{I}_{\text{enh}})$ exceeds its spatial mean. The edge term is

$$\mathcal{L}_{\text{edge}} = \left\| M_e \odot \left(E(\hat{I}_{\text{pred}}) - E(\hat{I}_{\text{enh}}) \right) \right\|_1. \tag{22}$$

Using the Laplacian operator Δ, the retained high-frequency ratio is

$$\rho_{\text{hf}} = \frac{\text{mean}\,|\Delta\hat{I}_{\text{pred}}| + \epsilon}{\text{mean}\,|\Delta\hat{I}_{\text{enh}}| + \epsilon}. \tag{23}$$

The structural regularizer used in the implementation is

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \lambda_{\text{hf}}\max(0, \tau_{\text{hf}} - \rho_{\text{hf}}), \tag{24}$$

where λ_{edge} and λ_{hf} balance the two structural terms, while τ_{hf} specifies the minimum retained high-frequency ratio. These parameters are fixed whenever Stage 2 is included.

Algorithm 1 summarizes the training and inference procedure.

Algorithm 1 Training and inference procedure of WISE-Net.

Input: Low-light grayscale image I.
Stage 1 training:
1: Decompose I into Haar wavelet sub-bands.
2: Predict the brightness-aware gain map G from the low-frequency component.
3: Modulate the low-frequency component with G and reconstruct $\hat{I}_{\text{enh}}$.
4: Optimize the Stage-1 network using the dual-view self-supervised objective and enhancement regularization.
Stage 2 training:
5: Freeze the trained Stage-1 network and generate $\hat{I}_{\text{enh}}$.
6: Randomly mask pixels in $\hat{I}_{\text{enh}}$ and replace them with neighboring pixels to obtain $\tilde{I}_{\text{enh}}$.
7: Feed $\tilde{I}_{\text{enh}}$ to the SBR network and predict the residual $\hat{n}_{\text{tr}}$.
8: Form $\hat{I}_{\text{pred}} = \tilde{I}_{\text{enh}} - \hat{n}_{\text{tr}}$ and optimize masked-pixel reconstruction together with structural regularization.
Inference:
9: Feed the complete Stage-1 output $\hat{I}_{\text{enh}}$ to the trained SBR network without masking.
10: Predict $\hat{n}_{\text{inf}}$ and output $\hat{I}_{\text{out}} = \hat{I}_{\text{enh}} - \hat{n}_{\text{inf}}$.

2.4. Training Strategy

Training proceeds sequentially in two stages.

Stage 1. We generate two diagonally resampled views online from each raw training image. The illumination modulation network is trained for 50 epochs on 128×128 patches with a batch size of 32. The training cascade includes the auxiliary denoising branch and dual-view consistency loss. After training, Stage 1 supplies enhanced images to Stage 2.

Stage 2. We freeze the Stage 1 network and train only SBR for 50 epochs on the resulting enhanced image's distribution, using a batch size of 64. This exposes SBR to the same image statistics it encounters at inference while keeping visibility recovery separate from refinement.

We optimize the two stages separately. Stage 1 uses Adam, while Stage 2 uses AdamW; both start with a learning rate of 10^{-4} and use cosine annealing. The auxiliary denoising U-Net is instantiated only in the Stage 1 training cascade. It processes the two enhanced views for the dual-view consistency objective shown in Figure 1. Neither its parameters nor its output is used in the Stage 2 inference path. During Stage 2 training, the frozen Stage 1 output serves as the unmasked target; random masking and neighbor replacement produce the input to f_θ.

Figure 2 compares the input and the two stage outputs with their Haar high-frequency responses. The aligned crops show how weak textures and edges appear after enhancement and refinement.

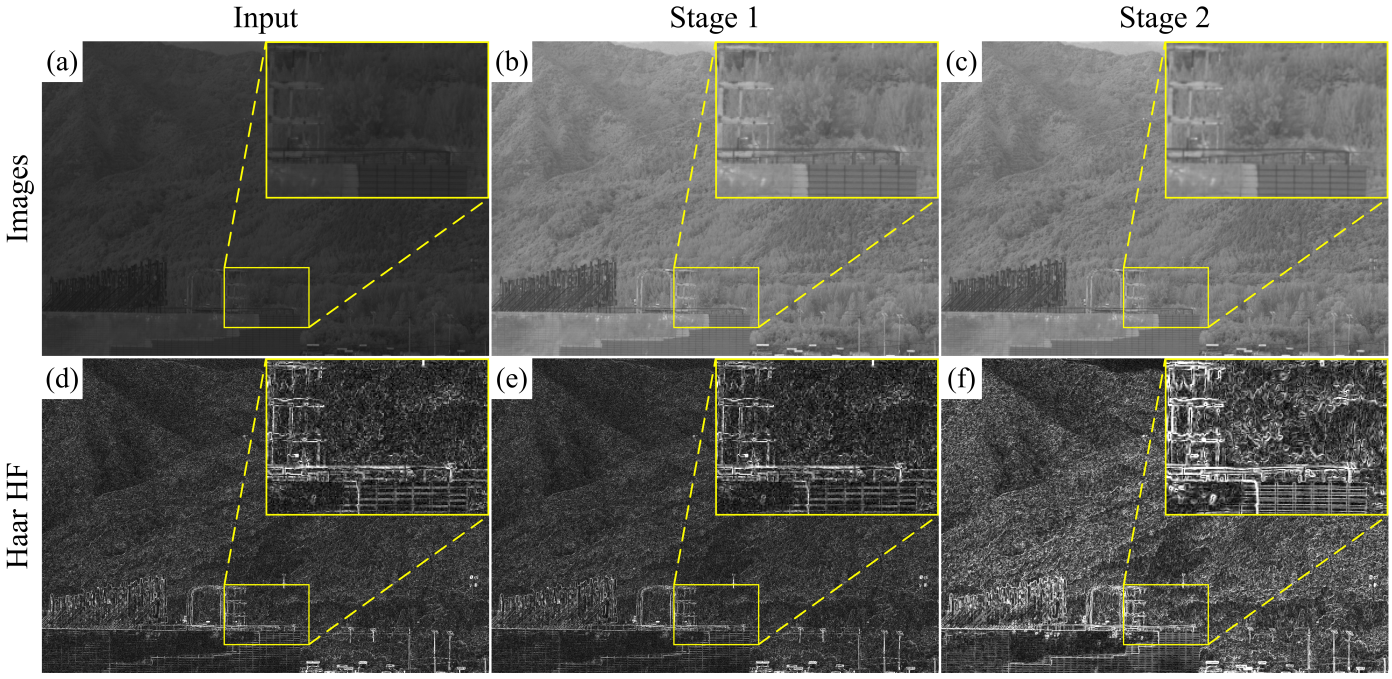

Figure 2. Visual evidence of illumination enhancement and high-frequency structure preservation. From left to right, the first row shows the input, the Stage 1 output, and the Stage 2 output; the second row shows the Haar high-frequency response of each. The yellow boxes mark the regions enlarged in the insets, and the dashed lines connect the regions to their insets. Stage 1 improves the visibility of dark regions while maintaining the spatial continuity of high-frequency structures; Stage 2 further suppresses residual noise and local artifacts while preserving these structures.

2.5. Evaluation Protocol

We evaluate WISE-Net on real low-light images acquired with the target TDI-ICMOS system. This section describes the data and evaluation protocol, compares representative unsupervised baselines, and examines the two-stage design.

2.5.1. Dataset and Settings

We collected the data using a long-range ground-based camera-level equivalent push-broom configuration of a time-delay-and-integration intensified complementary metal oxide semiconductor (TDI-ICMOS) camera. Figure 3 shows the nighttime setup, while Figure 4 compares its camera-level scanning geometry with air-to-ground imaging. The system retains line scanning, motion-synchronized TDI integration, an image intensifier-coupled optical chain, and 16-bit grayscale readout. It reproduces relative motion, line sampling, and camera-level charge accumulation, but not the atmospheric path, viewing geometry, platform dynamics, or radiometry of an airborne or orbital acquisition. Therefore, we treat these data as an in-house ground-based validation set, not as a public benchmark or a substitute for airborne data. The collection contains 1500 real single-channel 16-bit images at 2048 × 1440, consisting of 1460 images for training and 40 for test1_2048 × 1440. The 1460 training images and 40 Test 1 images came from the same acquisition collection. The sets are disjoint, and contain no adjacent scan frames across the split. The split was not grouped by acquisition location. Test 1 evaluates enhancement under the same acquisition conditions. Two additional subsets acquired with the same camera at different image widths are used to assess format generalization: test2_4096 × 1440 has 107 images, while test3_1536 × 1440 has 93.



Figure 3. Nighttime ground-based acquisition setup used to collect the in-house low-light images. The TDI-ICMOS camera is fitted with a 144-mm focal-length lens and rigidly mounted at the center of a turntable for scanning of distant targets.

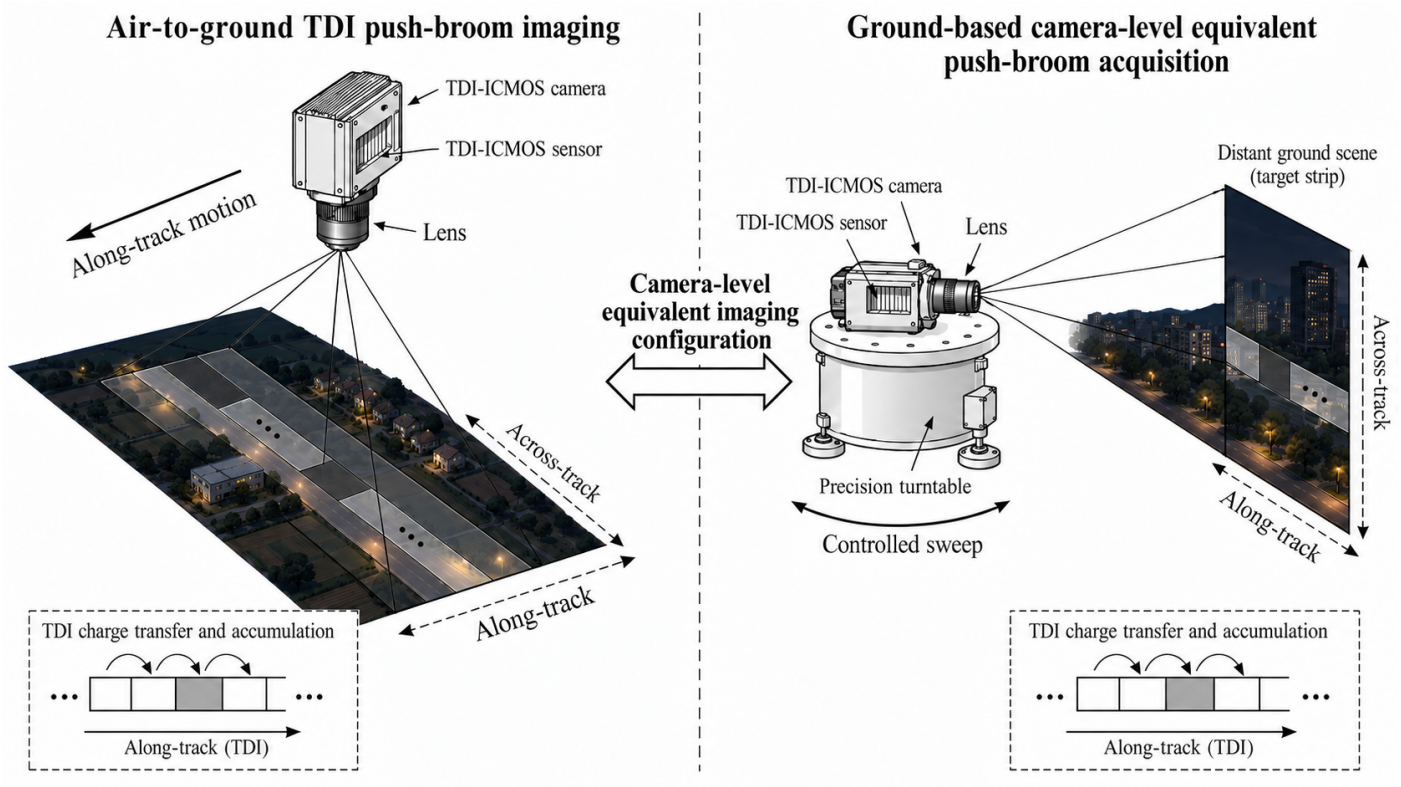

Figure 4. Air-to-ground TDI push-broom imaging and the ground-based camera-level equivalent validation configuration. The left panel illustrates along-track motion, cross-track coverage, and TDI charge transfer and accumulation in an air-to-ground acquisition. The right panel shows the ground-based setup used in this study, in which a precision turntable produces controlled relative motion between the TDI-ICMOS camera and a distant target strip. This setup reproduces the camera-level line scanning and TDI accumulation mechanism.

The camera used a 144 mm focal-length lens and observed targets at ranges of several hundred to several thousand meters. These ranges are much greater than the focal length, so the target can be treated as a distant object for the following paraxial small-angle sampling model. The TDI-ICMOS camera was rigidly fixed at the turntable center, making the turntable mechanical angular velocity w equal to the camera's scan angular velocity, with no intervening transmission ratio. Let u denote the pixel pitch, F the lens focal length, and r the TDI line rate. During one line period $t = 1/r$, the angular scan is wt. Matching the image displacement to one pixel gives

$$\frac{u}{F} = wt = \frac{w}{r}, \qquad w = \left(\frac{u}{F}\right) r. \tag{25}$$

Thus, for fixed pixel pitch and focal length, the ideal scan rate is proportional to the TDI line rate. This relation describes the camera-level motion and line sampling synchronization used in the ground experiment. The long-range geometry approximates distant-target push-broom imaging, but the setup does not reproduce orbital motion, platform attitude dynamics, the atmospheric path, or the radiometric conditions of satellite acquisition.

The source data were decoded and retained as 16-bit grayscale images, without prior 8-bit quantization; normalized floating-point values were used for network computation. WISE-Net is optimized sequentially: Stage 1 learns illumination modulation from image patches, after which it is frozen while Stage 2 learns SBR on the enhanced distribution. The proposed method and all compared baselines were implemented on the same computer using PyTorch and an NVIDIA GPU. The WISE-Net environment used PyTorch 2.11.0+cu128. The dataset composition is summarized in Table 2.

Table 2. Statistics of the training and evaluation data used for WISE-Net.

Split	Images	Resolution	Format	Bit Depth	Role
Training	1460	2048 × 1440	TIFF	16-bit	Training
test1_2048 × 1440	40	2048 × 1440	TIFF	16-bit	Evaluation
test2_4096 × 1440	107	4096 × 1440	TIFF	16-bit	Generalization
test3_1536 × 1440	93	1536 × 1440	TIFF	16-bit	Generalization

2.5.2. Evaluation Metrics

The RAW camera data were decoded to 16-bit grayscale TIFF. Training and inference used 16-bit TIFF inputs, with normalized floating-point values for network computation, and restored images were saved as 16-bit TIFF. For no-reference metric calculation, the evaluated images were converted to 8-bit grayscale by linear mapping over a common fixed intensity range. The same mapping was applied to all evaluated methods without per-image rescaling.

Because the real TDI-ICMOS images do not have paired normal-light ground truth, the primary in-house evaluation uses no-reference image-quality measures. The Natural Image Quality Evaluator (NIQE) models the deviation of locally normalized image statistics from a learned corpus of undistorted natural images; because it is a distance measure, lower values indicate greater statistical naturalness [31]. For fitted feature means μ_0, μ_I and covariance matrices Σ_0, Σ_I of the pristine model and evaluated image, respectively, its distance is $D_{\mathrm{NIQE}} = [(\mu_0 - \mu_I)^{\mathsf{T}}((\Sigma_0 + \Sigma_I)/2)^{-1}(\mu_0 - \mu_I)]^{1/2}$. The Perception-based Image Quality Evaluator (PIQE) estimates spatially localized distortion using block-level statistics and perceptual impairment features; lower values indicate fewer estimated distortions [32]. The Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) uses spatial-domain natural scene statistics and a learned quality regressor to predict perceptual degradation without a reference image; lower values indicate better predicted quality [33]. NIQE compares fitted natural-scene statistics with a pristine-image model, PIQE aggregates block-level distortion estimates without a trained quality regressor, and BRISQUE predicts quality using a trained regressor, meaning that their scores follow different assessment procedures rather than a shared fidelity formula. For the external paired datasets, PSNR, SSIM, and VIFp are reference-based measures computed from the restored image and its registered target.

For an image with normalized gray-level histogram $p(k)$, the entropy is defined as

$$H(I) = -\sum_{k} p(k) \log_2 p(k), \tag{26}$$

where a larger value indicates a broader gray-level distribution and potentially richer intensity variation. Entropy is not a fidelity measure, since noise and artificial contrast can also increase it. NIQE, PIQE, and BRISQUE were not calibrated here against perceptual judgments for 16-bit TDI-ICMOS imagery; their scores depend on the intensity representation and the statistics assessed by each implementation, and retaining 16-bit source files does not establish that all sensor precision is represented by these scores. A lower score indicates agreement with the metric criterion. We interpret these measures together with visual observations and the external reference-based evaluation. The directions used in the tables are NIQE ↓, PIQE ↓, BRISQUE ↓, and Entropy ↑.

3. Results

3.1. Quantitative and Qualitative Comparisons

The quantitative and qualitative results are organized by test subset so that each metric table can be read together with its corresponding visual comparison. Bold and underlined entries denote the best and second-best values, respectively.

3.1.1. Results on test1_2048 × 1440

On test1_2048 × 1440, WISE-Net obtains the lowest NIQE and PIQE and the highest entropy, while SCI achieves the lowest BRISQUE (Table 3). Figure 5 compares two scenes and enlarged regions around vegetation boundaries and rooftop structures. Zero-DCE++ appears pale and compresses gray-level variation. RUAS over-enhances broad areas and obscures local detail. SCI retains much of the texture, but push-broom jitter and blur remain visible. CoLIE provides limited brightness recovery and weak separation between vegetation layers. ZeroIG and SCI++ introduce a gray cast and low local contrast, with the effect less pronounced for SCI++. Li et al. is excessively bright, rendering fine vegetation as diffuse clumps. Di-Retinex remains dark with weak tonal separation. WISE-Net produces a more balanced brightness distribution and clearer vegetation transitions, roof tiles, parapets, and thin metal structures; its SBR stage also suppresses residual noise and local artifacts. These comparisons do not independently establish correction of physical scan jitter or motion blur.

Table 3. Quantitative comparison on test1_2048 × 1440. Best values are shown in bold, second-best values are underlined.

Method	NIQE ↓	PIQE ↓	BRISQUE ↓	Entropy ↑
Zero-DCE++ (TPAMI 2021) [19]	7.4854	24.0223	43.1332	5.2962
RUAS (CVPR 2021) [20]	6.4797	46.0139	52.6413	<u>5.8234</u>
SCI (CVPR 2022) [21]	7.1531	22.8046	**41.6851**	4.5980
CoLIE (ECCV 2024) [34]	7.3754	21.0661	42.2756	5.1808
ZeroIG (CVPR 2024) [23]	6.6634	37.8086	54.0565	5.4621
SCI++ (TPAMI 2025) [22]	7.2950	26.5848	<u>41.8575</u>	5.0347
Li et al. (ICLR 2025) [35]	<u>5.8307</u>	<u>18.0590</u>	46.2133	5.5287
Di-Retinex (IJCV 2025) [36]	8.4895	28.3053	53.8734	3.6802
WISE-Net (Our)	**5.5657**	**13.6800**	43.2031	**5.9167**

3.1.2. Results on test2_4096 × 1440

On test2_4096 × 1440, WISE-Net obtains the best NIQE, PIQE, and entropy scores and the second-best BRISQUE score (Table 4). Figure 6 shows a scene with enlarged regions around a window and its hardware, a vegetation boundary, eaves, and a metal utility pole. Zero-DCE++ appears pale, making the window handle difficult to distinguish and weakening separation between vegetation layers. RUAS overexposes the handle and merges much of the mountain vegetation into a bright region with poor texture. SCI recovers illumination and texture, although jitter and blur remain. CoLIE provides weaker brightness recovery and limited tonal separation. ZeroIG strongly whitens the scene, bringing vegetation close in intensity to the sky; SCI++ shows this effect less strongly. Li et al. is too bright and gray, with diffuse vegetation texture and less distinct eaves and walls. Di-Retinex remains dark with limited local contrast. WISE-Net retains clearer window hardware, vegetation boundaries, eaves, and pole structures, while refinement suppresses residual noise and local artifacts.

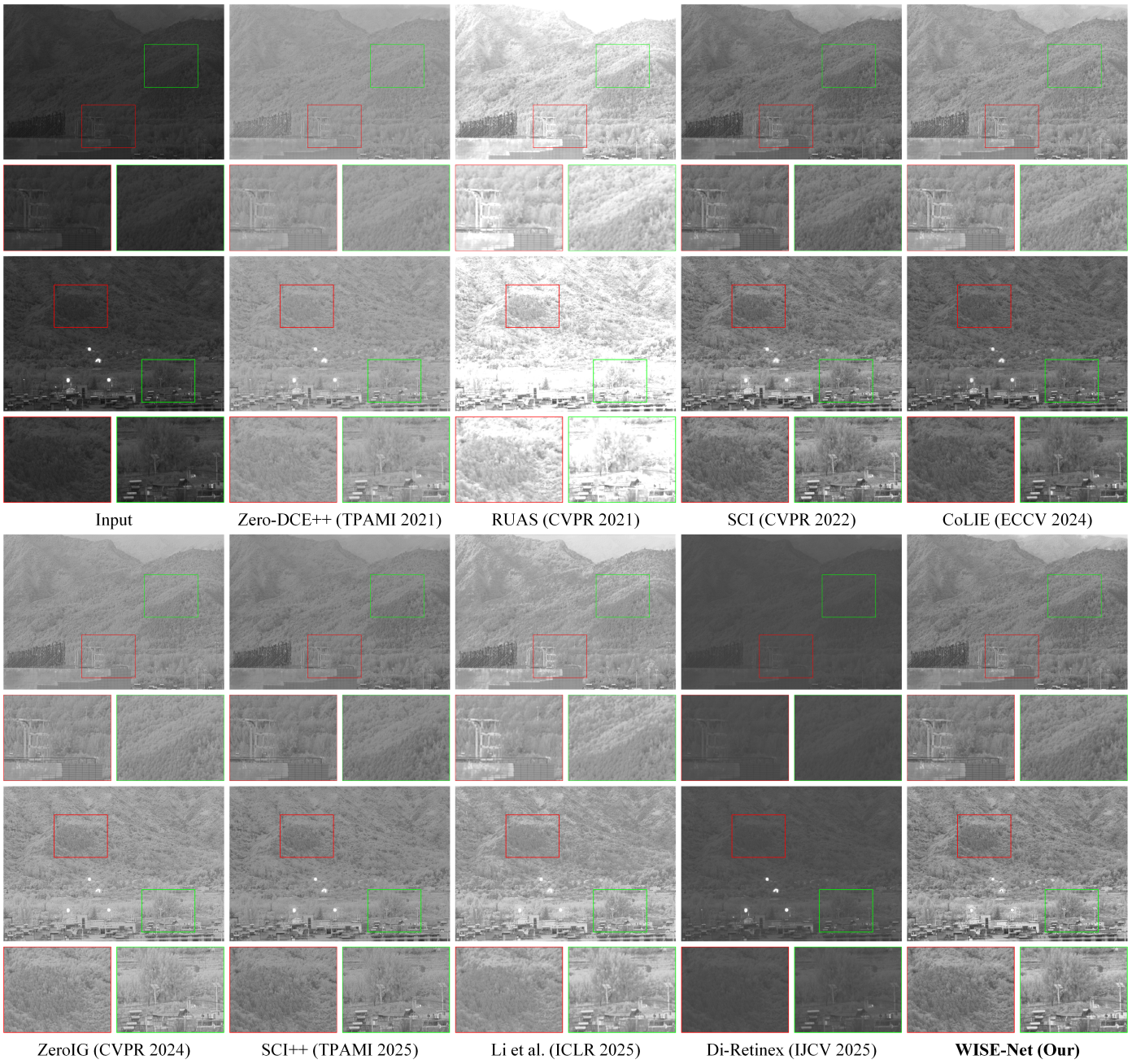

Figure 5. Qualitative comparison on test1_2048 × 1440. WISE-Net improves global illumination while preserving weak structures and suppressing enhancement-amplified artifacts. Matching colors indicate corresponding crops [19–23,34–36].

Table 4. Quantitative comparison on test2_4096 × 1440. Best values are shown in bold, second-best values are underlined.

Method	NIQE ↓	PIQE ↓	BRISQUE ↓	Entropy ↑
Zero-DCE++ (TPAMI 2021) [19]	7.6608	31.1089	43.2841	5.4586
RUAS (CVPR 2021) [20]	6.6339	37.9125	52.3296	<u>5.6375</u>
SCI (CVPR 2022) [21]	7.3265	24.3116	42.2891	4.7171
CoLIE (ECCV 2024) [34]	7.4923	22.9593	42.7686	5.2443
ZeroIG (CVPR 2024) [23]	6.8635	29.1717	53.6837	5.4854
SCI++ (TPAMI 2025) [22]	7.4810	30.5791	**41.5197**	5.0911
Li et al. (ICLR 2025) [35]	<u>5.7755</u>	<u>14.6675</u>	43.1569	5.5761
Di-Retinex (IJCV 2025) [36]	8.7859	20.6568	52.2714	3.7175
WISE-Net (Our)	**5.6513**	**10.0385**	<u>42.0130</u>	**5.9698**

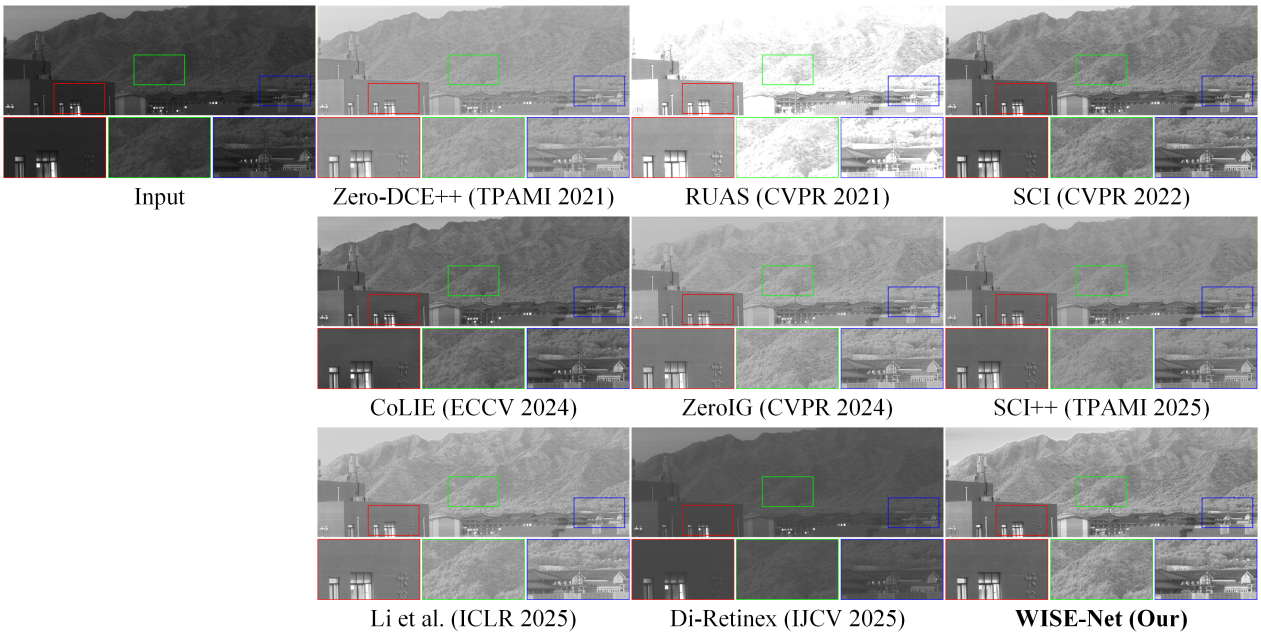

Figure 6. Qualitative comparison on the wider-format generalization subset test2_4096 × 1440. WISE-Net maintains structural detail and stable local contrast after brightness lifting. Matching colors indicate corresponding crops [19–23,34–36].

3.1.3. Results on test3_1536 × 1440

On test3_1536 × 1440, WISE-Net obtains the lowest NIQE and highest entropy and ranks second in PIQE; SCI and CoLIE obtain the two lowest BRISQUE scores (Table 5). Figure 7 shows two urban scenes. The first row contains Input, Zero-DCE++, RUAS, SCI, and CoLIE; the second contains ZeroIG, SCI++, Li et al., Di-Retinex, and WISE-Net. The enlarged regions cover a mid-rise building, the lower floors of a high-rise beside trees, and the upper floors and roof of another residence near a utility pole. These crops expose facade texture, roof boundaries, vegetation, and linear structures.

Table 5. Quantitative comparison on test3_1536 × 1440. Best values are shown in bold, second-best values are underlined.

Method	NIQE ↓	PIQE ↓	BRISQUE ↓	Entropy ↑
Zero-DCE++ (TPAMI 2021) [19]	7.2763	**15.7598**	37.5953	5.7655
RUAS (CVPR 2021) [20]	5.9977	46.4076	54.4205	<u>6.3648</u>
SCI (CVPR 2022) [21]	6.6631	20.2412	**32.4015**	5.4162
CoLIE (ECCV 2024) [34]	7.0021	17.4580	<u>33.3426</u>	5.7561
ZeroIG (CVPR 2024) [23]	5.9598	45.0477	52.1824	6.1888
SCI++ (TPAMI 2025) [22]	6.8288	21.3105	35.4894	5.7982
Li et al. (ICLR 2025) [35]	<u>5.8545</u>	20.8049	42.3929	6.2841
Di-Retinex (IJCV 2025) [36]	7.6053	36.7662	46.6907	4.5505
WISE-Net (Our)	**5.0743**	<u>17.0166</u>	36.4186	**6.3746**

Zero-DCE++ appears pale and reduces tonal separation between the high-rise facade and nearby trees. RUAS overexposes the facade and sky, obscuring building detail. SCI recovers illumination while retaining texture, but structural edges remain blurred by jitter. CoLIE leaves both scenes relatively dark with weak local contrast. ZeroIG whitens and smooths the high-rise facade; SCI++ shows a milder gray shift. Li et al. is overly bright and gray, reducing separation on the facade and roof. Di-Retinex remains dark with limited contrast. WISE-Net improves visibility and separates buildings from vegetation; its refined output retains facade and roof texture and reduces visible artifacts around building edges and the utility pole. The crops indicate improved local edge definition after SBR.

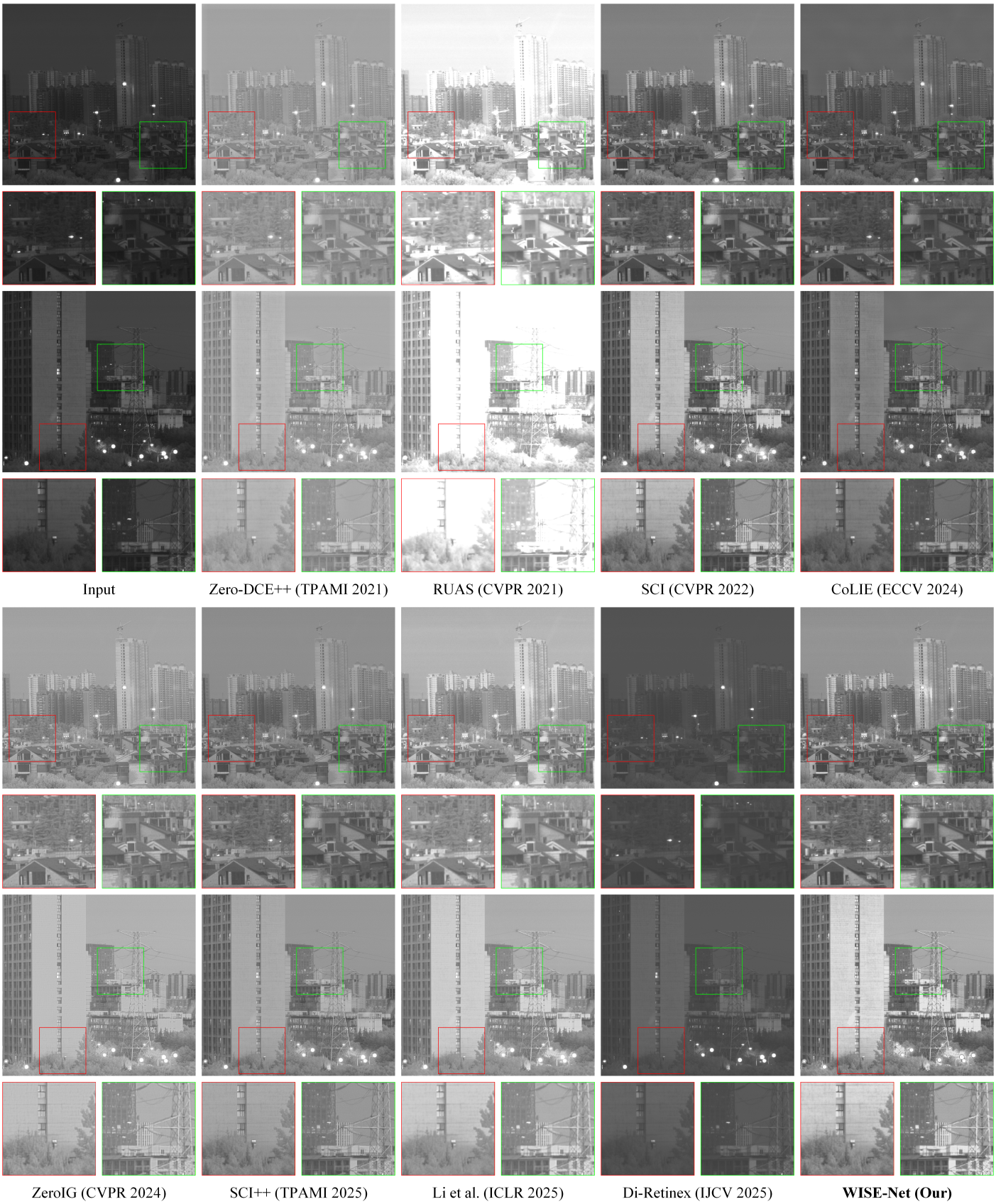

Figure 7. Qualitative comparison on the narrower-format generalization subset test3_1536 × 1440. WISE-Net recovers visibility while retaining detail and limiting residual noise. Matching colors indicate corresponding crops [19–23,34–36].

To complement the no-reference evaluation on the in-house camera data, we report reference-based results on two public paired low-light remote sensing test sets released with the DFFN study [3]. The DFFN data construction starts from 751 high-resolution images selected from the iSAID test set, creates low-light/normal-light pairs through random cropping and low-light simulation, and additionally provides a high-pixel test subset for high-resolution evaluation. In the versions used here, *iSAID2_val* consists of paired 512 × 512 images, whereas *iSAID-high-pixel* consists of paired 1920 × 1080 images. Table 6 lists the PSNR, SSIM, and VIFp for both subsets, computed against their registered targets. We apply the weights learned from the 1460-image in-house training set directly to these two external sets; no external images are used for retraining, fine-tuning, or parameter selection.

Table 6. Reference-based evaluation on the two public paired low-light remote-sensing test sets. Best values are shown in bold, second-best values are underlined.

Method	*iSAID2_val*			*iSAID-high-pixel*		
	PSNR ↑	**SSIM ↑**	**VIFp ↑**	**PSNR ↑**	**SSIM ↑**	**VIFp ↑**
Zero-DCE++ (TPAMI 2021) [19]	14.0800	0.3203	0.1025	12.5472	0.2560	0.0864
RUAS (CVPR 2021) [20]	12.1421	0.3150	0.0904	10.3706	0.2299	0.0736
SCI (CVPR 2022) [21]	14.2234	0.3061	0.1039	15.1587	0.2639	0.0863
CoLIE (ECCV 2024) [34]	14.1994	0.3261	0.1127	16.0232	0.2945	0.0956
ZeroIG (CVPR 2024) [23]	<u>17.0671</u>	<u>0.5114</u>	0.1549	<u>17.6513</u>	<u>0.4446</u>	0.1290
SCI++ (TPAMI 2025) [22]	14.5281	0.3038	0.0980	15.0944	0.2554	0.0808
Li et al. (ICLR 2025) [35]	15.4837	0.4497	0.1373	13.1876	0.3871	0.1164
Di-Retinex (IJCV 2025) [36]	14.6775	0.4519	<u>0.1684</u>	16.9810	0.4345	<u>0.1409</u>
WISE-Net (Our)	**23.5101**	**0.7762**	**0.3202**	**19.9138**	**0.7183**	**0.2762**

This protocol tests transfer under a joint change in scene content, image format, and degradation distribution. Differences in sensor response, noise correlation, and low-light simulation are plausible sources of mismatch; however, the present experiment does not isolate their individual effects. The paired test results establish performance on these particular datasets, but do not identify a causal explanation for cross-sensor generalization.

3.1.4. Results on *iSAID2_val*

On *iSAID2_val*, WISE-Net achieves the highest PSNR, SSIM, and VIFp among the compared methods. Figure 8 shows scenes containing parked vehicles, marked court lines, and tree–ground boundaries. RUAS overexposes broad regions and loses tonal separation, whereas several other methods either remain dark or introduce gray whitening. WISE-Net produces a brightness level closer to the reference. The red crops preserve vehicle contours and nearby linear structures, while the green crops retain court boundaries and vegetation texture. These visual observations are consistent with the reference-based scores, and were obtained without retraining.

3.1.5. Results on *iSAID-high-pixel*

The *iSAID-high-pixel* subset tests the same model at a substantially larger spatial format. WISE-Net again obtains the highest PSNR, SSIM, and VIFp among the compared methods (Table 6). As shown in Figure 9, the scene contains multiple courts, parked vehicles, trees, and broad low-contrast surfaces. RUAS produces strong overexposure, whereas the other methods differ in brightness and local sharpness. WISE-Net maintains vehicle shapes in the red crops and court lines and the tree boundary in the green crops. Its output is closer to the reference in both overall tone and local structure. Nevertheless, the 19.9138 PSNR, 0.7183 SSIM, and 0.2762 VIFp obtained at this larger format should be interpreted as evidence of transfer under a changed image size rather

than as saturation-level fidelity. The absolute values leave room for improvement in large-format generalization, especially under differences between the public paired data and the in-house TDI-ICMOS data.

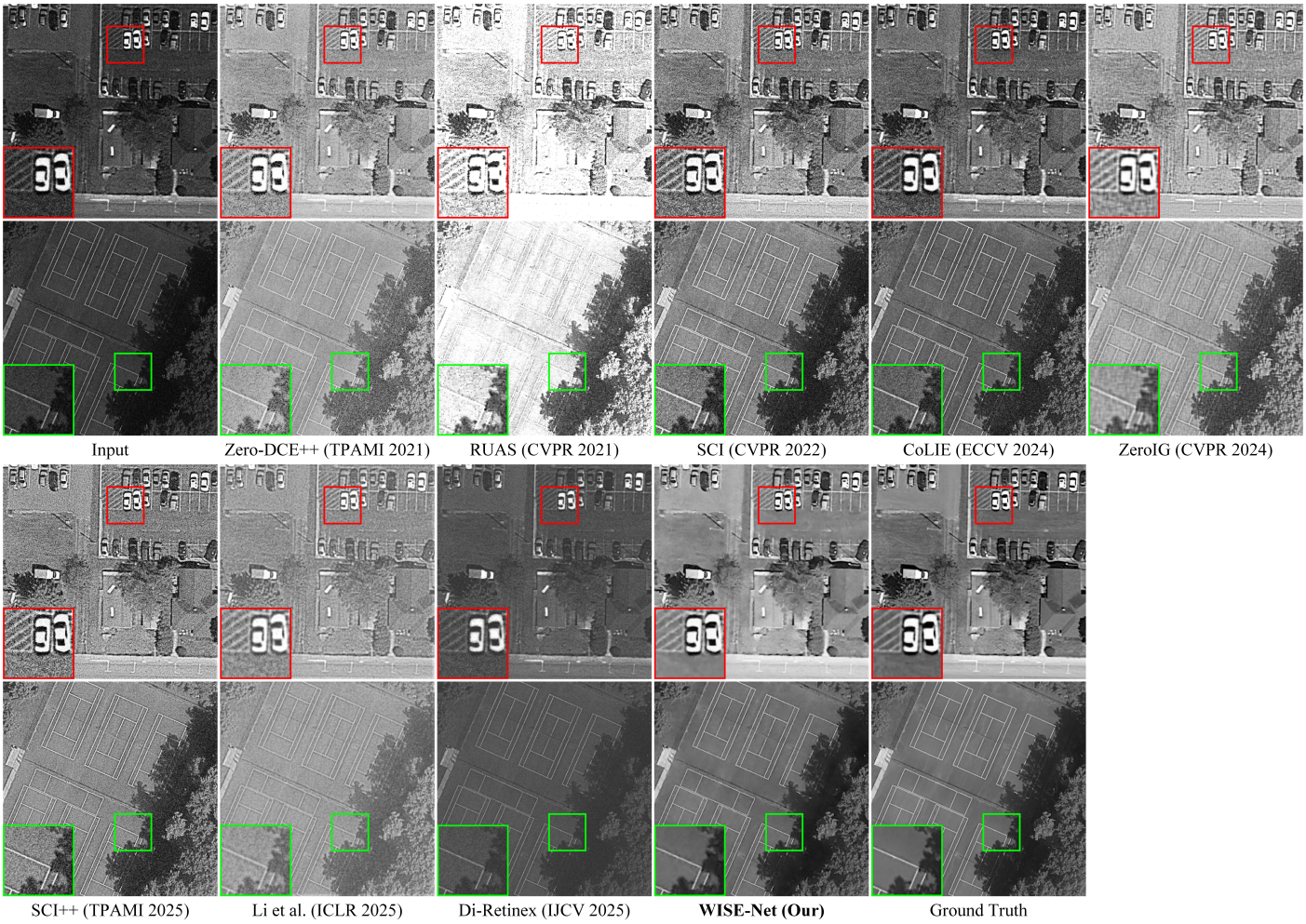

Figure 8. Visual comparison on the external paired *iSAID2_val* low-light remote sensing test set. The boxes and enlarged regions mark the same spatial locations across the input, reference, and enhanced outputs. Matching colors indicate corresponding crops [19–23,34–36].

3.2. Cumulative and Single-Component Ablation

Table 7 combines two ablation designs on the 40 images of test1_2048 × 1440. The cumulative baseline retains the Stage 1 architecture, dual-view training with the auxiliary denoising branch, and $\mathcal{L}_{rc}$, $\mathcal{L}_{mg}$, $\mathcal{L}_{ov}$, and $\mathcal{L}_{ex}$. We successively add gain map smoothness, bright region protection, and SBR. The last two rows individually remove the smoothness and bright region protection losses from full WISE-Net. The third cumulative configuration is full Stage 1, and adding SBR gives full WISE-Net.

Along the cumulative sequence, adding $\mathcal{L}_{sm}$ and then $\mathcal{L}_{bp}$ lowers PIQE and BRISQUE at each step. Adding SBR improves all four reported values along this sequence. Figure 10 illustrates all six configurations using the same crop locations. Removing $\mathcal{L}_{bp}$ worsens all three quality scores and lowers entropy relative to full WISE-Net. Removing $\mathcal{L}_{sm}$ slightly lowers NIQE but raises PIQE and BRISQUE while also lowering entropy. The full model has the lowest PIQE and BRISQUE and the highest entropy, but does not reach the lowest NIQE across the combined table. These comparisons do not establish order-independent effects, nor do they separate the effects of masked reconstruction and structural regularization within SBR.

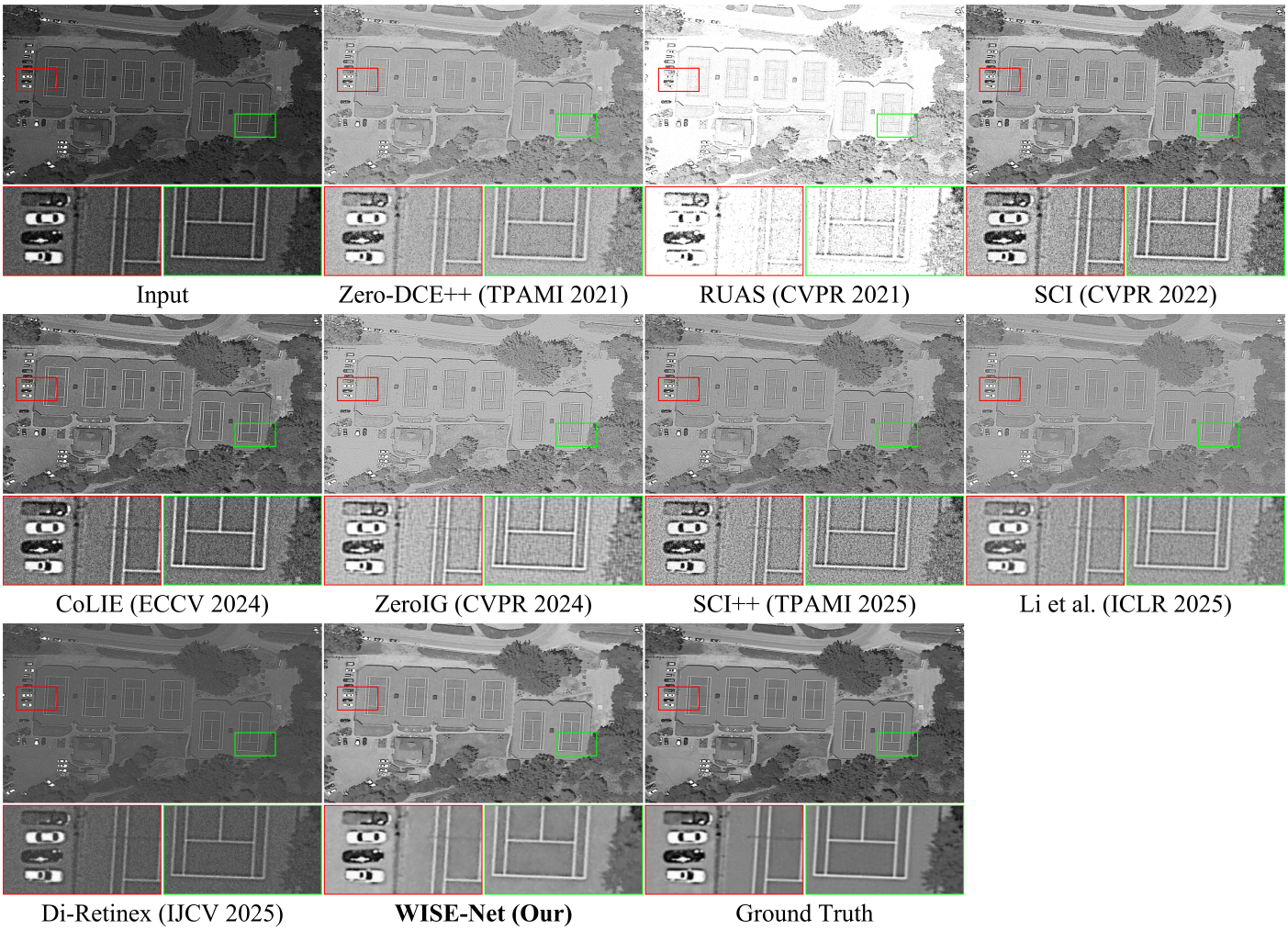

Figure 9. Visual comparison on the external paired *iSAID-high-pixel* test set. The high-pixel images test whether the enhancement behavior remains stable when the spatial format differs from the in-house training images. Matching colors indicate corresponding crops [19–23,34–36].

Table 7. Cumulative and single-component removal results on test1_2048 × 1440. Bold and underlining indicate the best and second-best values in each column, respectively.

Configuration	NIQE ↓	PIQE ↓	BRISQUE ↓	Entropy ↑
Baseline	6.3107	33.6922	50.1445	5.5272
Baseline + $\mathcal{L}_{sm}$	6.2994	32.1537	50.0666	<u>5.5417</u>
Baseline + $\mathcal{L}_{sm}$ + $\mathcal{L}_{bp}$	6.3124	30.7389	47.1781	5.5287
Baseline + $\mathcal{L}_{sm}$ + $\mathcal{L}_{bp}$ + SBR (WISE-Net)	<u>5.5657</u>	**13.6800**	**43.2031**	**5.9167**
WISE-Net w/o $\mathcal{L}_{sm}$	**5.4919**	29.1536	<u>44.8655</u>	5.5199
WISE-Net w/o $\mathcal{L}_{bp}$	5.6443	<u>28.8372</u>	45.3539	5.5293

Figure 11 summarizes the four no-reference metrics on test1_2048 × 1440; their exact values and rankings are reported in Table 3.

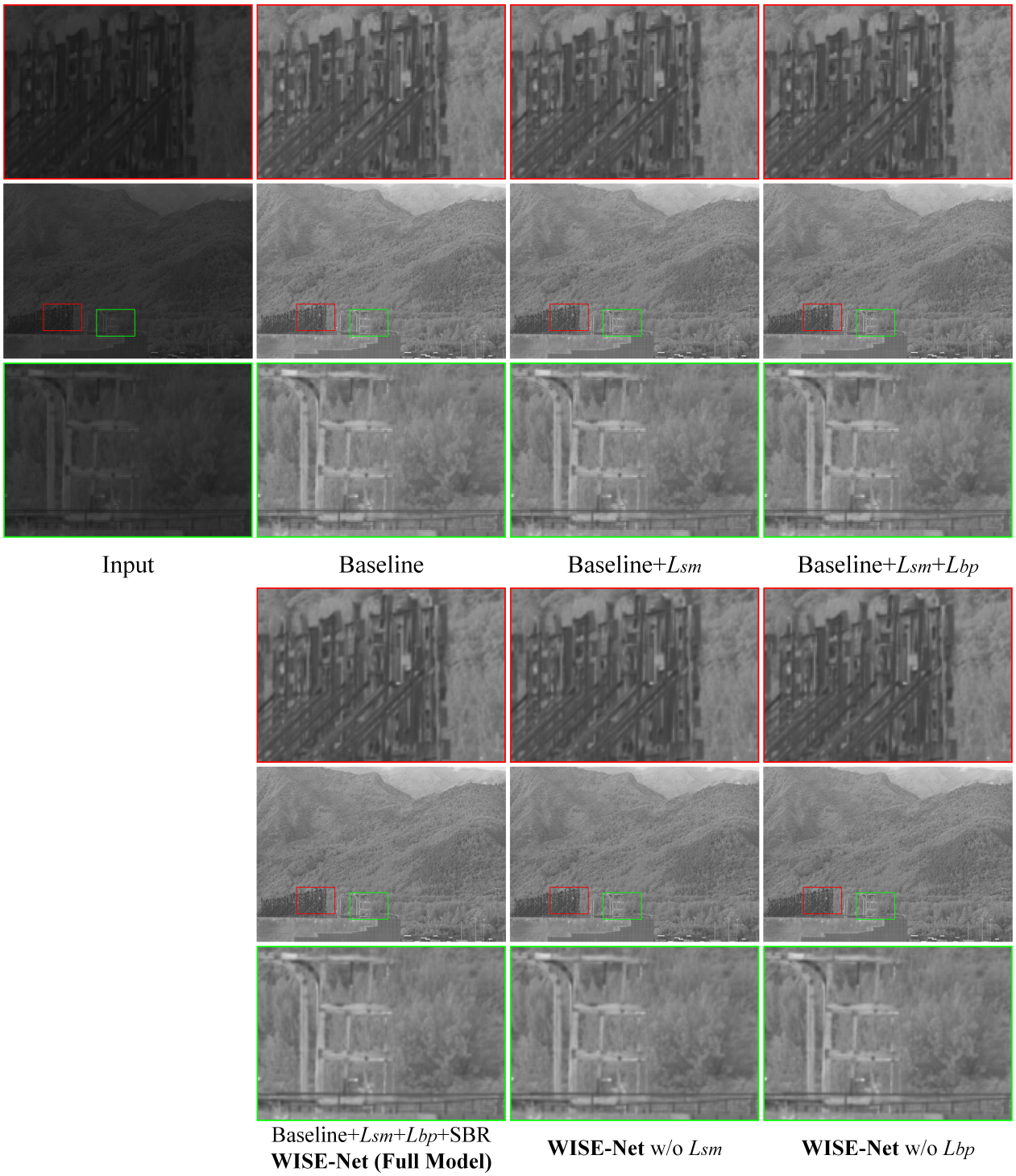

Figure 10. Visual comparison of the cumulative and single-component ablations. The upper block, from left to right, shows Input, Baseline, Baseline+$\mathcal{L}_{\mathrm{sm}}$, and Baseline+$\mathcal{L}_{\mathrm{sm}}$+$\mathcal{L}_{\mathrm{bp}}$, while the lower block shows full WISE-Net after adding SBR, WISE-Net without $\mathcal{L}_{\mathrm{sm}}$, and WISE-Net without $\mathcal{L}_{\mathrm{bp}}$. Within each block, the middle row shows the full scene; the upper and lower rows show the red and green crops. All configurations use the same crop locations.

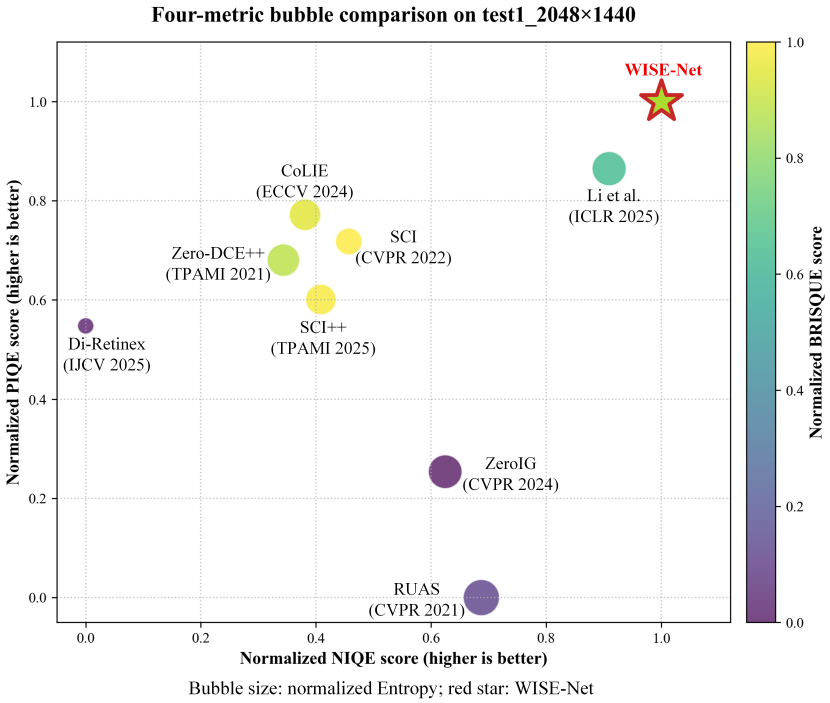

Figure 11. Four-metric bubble comparison on test1_2048 × 1440. The chart summarizes the relative NIQE, PIQE, BRISQUE, and entropy values of the nine compared methods [19–23,34–36].

3.3. Sensitivity to Exclusion Weight and Masking Ratio

We evaluated two parameter sweeps on the same 40 images from test 1 (Table 8). In Stage 1, we varied the exclusion weight and evaluated the enhanced output. In Stage 2, we varied the masking ratio while keeping the original Stage 1 checkpoint frozen, then evaluated the refined output. Each sweep includes a fresh evaluation of the corresponding original checkpoint at the default setting of 0.50. Test 1 inputs and restored outputs were stored as 16-bit TIFF. The no-reference scores were computed on the 8-bit representation described in the evaluation protocol.

Table 8. Parameter sensitivity on the 40 images from test 1. Each entry is the mean across images. The default setting is 0.50 in each sweep.

Stage	Parameter Value	NIQE ↓	PIQE ↓	BRISQUE ↓	Entropy ↑
Stage 1	$w_{ex} = 0.25$	6.2419	35.6273	46.3420	5.5351
	$w_{ex} = 0.50$	6.3124	30.7389	47.1781	5.5287
	$w_{ex} = 1.00$	6.3737	46.7303	51.5111	5.5218
Stage 2	Mask ratio 0.25	5.7037	13.7141	43.0944	5.9351
	Mask ratio 0.50	5.5657	13.6800	43.2031	5.9167
	Mask ratio 0.75	6.5901	13.7399	43.5995	5.8987

Increasing the Stage 1 exclusion weight from 0.25 to 0.50 lowers PIQE from 35.6273 to 30.7389, while NIQE and BRISQUE rise slightly. At 1.00, all three scores worsen relative to 0.50, with the largest numerical increase in PIQE. Entropy decreases slightly as the weight increases. With Stage 1 frozen, a Stage 2 mask ratio of 0.50 yields the lowest NIQE and PIQE, while 0.25 yields the lowest BRISQUE and the highest entropy. Increasing the ratio to 0.75 raises NIQE from 5.5657 to 6.5901, with smaller changes in PIQE and BRISQUE; entropy decreases across the three ratios. We retain 0.50 for both parameters based on these trade-offs on test 1. Each setting is represented by one trained checkpoint.

4. Discussion

The two-stage design assigns clear but limited roles to the constraints. Low-frequency modulation changes brightness without learning a separate gain for each high-frequency sub-band, while SBR learns contextual refinement on the frozen enhanced distribution. Reusing measured high-frequency coefficients preserves a structural prior before clipping, meaning that it does not provide a clean-detail reference. Residual high-frequency noise and saturated values remain potential failure modes.

Both training objectives are bounded below by zero, but the neural network parameterizations and interacting regularizers make optimization nonconvex. A lower bound does not prove convergence of the parameter sequence, monotonic stochastic loss reduction, or attainment of a global optimum. Sequential training fixes the enhancement front end during refinement, thereby removing joint parameter updates across stages; however, it does not eliminate local optima. While training loss trajectories can document empirical optimization behavior, stability and sensitivity over repeated runs require separate experimental evidence.

Spatial dependence is a shared limitation of the self-supervised constraints. Diagonal resampling does not remove sensor correlation, and neighbor replacement does not ensure that target noise is independent of the visible context; consequently, these losses express useful consistency and contextual prediction assumptions rather than the likelihood of calibrated sensor noise. The present image-level evaluation does not establish a Poisson–Gaussian noise calibration or quantify the effect of acquisition sequence dependence.

The cumulative ablation measures conditional effects along its specified sequence, but does not identify order-independent contributions or establish the optimality of the fixed loss weights. Likewise, external paired testing measures transfer under the joint changes present in those datasets as opposed to isolating sensor response, noise correlation, and scene content effects. While these distinctions limit causal and cross-sensor claims, they do not change the reported evaluation protocol.

5. Conclusions

This paper presents WISE-Net, a fully self-supervised two-stage framework for low-light remote sensing image enhancement. Stage 1 applies a learned brightness-adaptive gain to the low-frequency Haar component and reuses the original high-frequency sub-bands during reconstruction. Stage 2 applies SBR to the enhanced-image distribution to suppress amplified noise and local artifacts without paired clean targets.

The Stage 1 objective combines dual-view consistency with constraints on relative contrast, minimum dark region gain, bright region gain, spatial smoothness, overexposure, and gradient exclusion. WISE-Net achieves the lowest NIQE on the three in-house subsets. In the cumulative ablation of test 1, adding SBR gives the lowest NIQE, while the remaining metrics show trade-offs.

The evaluation has a defined scope. The in-house grayscale data come from one TDI-ICMOS camera and a ground-based camera-level equivalent configuration, so they do not represent the full variability of airborne or orbital observations. The paired public tests provide cross-dataset evidence, but the scores on larger formats (PSNR 19.9138, SSIM 0.7183, and VIFp 0.2762) leave room for improvement. These tests do not replace validation with additional sensors or TDI conditions. Reusing the measured high-frequency sub-bands limits direct high-frequency amplification, but cannot recover structure absent from the input. The no-reference metrics are complementary indicators rather than absolute fidelity measures. Future work will extend validation to additional sensors, and where data can be released, to airborne or orbital acquisitions and downstream remote sensing tasks.

Author Contributions: Conceptualization, H.D.; methodology, H.D. and H.Z.; software, H.D.; validation, H.D. and Z.J.; formal analysis, H.D., H.Z. and S.W.; investigation, S.L. and Z.J.; resources, H.Z.; data curation, H.D.; writing—original draft preparation, H.D.; writing—review and editing, H.D., H.Z., S.W. and X.F.; visualization, H.D.; supervision, H.Z. and S.W.; project administration, H.Z. and S.W.; funding acquisition, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62403476).

Data Availability Statement: The in-house camera data are not publicly available due to restrictions on their release. The public evaluation datasets are identified in the manuscript.

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

References

1. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to See in the Dark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3291–3300. https://doi.org/10.1109/cvpr.2018.00347.
2. Wei, K.; Fu, Y.; Yang, J.; Huang, H. A Physics-Based Noise Formation Model for Extreme Low-Light Raw Denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 2755–2764. https://doi.org/10.1109/cvpr42600.2020.00283.
3. Yao, Z.; Fan, G.; Fan, J.; Gan, M.; Chen, C.L.P. Spatial–Frequency Dual-Domain Feature Fusion Network for Low-Light Remote Sensing Image Enhancement. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4706516. https://doi.org/10.1109/tgrs.2024.3434416.
4. Land, E.H.; McCann, J.J. Lightness and Retinex Theory. *J. Opt. Soc. Am.* **1971**, *61*, 1–11. https://doi.org/10.1364/josa.61.000001.
5. Jobson, D.J.; Rahman, Z.; Woodell, G.A. Properties and performance of a center/surround retinex. *IEEE Trans. Image Process.* **1997**, *6*, 451–462. https://doi.org/10.1109/83.557356.
6. Jobson, D.J.; Rahman, Z.; Woodell, G.A. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans. Image Process.* **1997**, *6*, 965–976. https://doi.org/10.1109/83.597272.
7. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. *arXiv* **2018**, arXiv:1808.04560. https://doi.org/10.48550/arXiv.1808.04560.
8. Zhang, Y.; Zhang, J.; Guo, X. Kindling the Darkness: A Practical Low-Light Image Enhancer. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1632–1640. https://doi.org/10.1145/3343031.3350926.
9. Yang, W.; Wang, W.; Huang, H.; Wang, S.; Liu, J. Sparse Gradient Regularized Deep Retinex Network for Robust Low-Light Image Enhancement. *IEEE Trans. Image Process.* **2021**, *30*, 2072–2086. https://doi.org/10.1109/tip.2021.3050850.
10. Lore, K.G.; Akintayo, A.; Sarkar, S. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognit.* **2017**, *61*, 650–662. https://doi.org/10.1016/j.patcog.2016.06.008.
11. Li, R.-K.; Li, M.-H.; Chen, S.-Q.; Chen, Y.-T.; Xu, Z.-H. Dark2Light: Multi-stage progressive learning model for low-light image enhancement. *Opt. Express* **2023**, *31*, 42887–42900. https://doi.org/10.1364/oe.507966.
12. Wang, T.; Zhang, K.; Shen, T.; Luo, W.; Stenger, B.; Lu, T. Ultra-High-Definition Low-Light Image Enhancement: A Benchmark and Transformer-Based Method. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 2654–2662. https://doi.org/10.1609/aaai.v37i3.25364.
13. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5718–5729. https://doi.org/10.1109/cvpr52688.2022.00564.
14. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H.; Shao, L. Learning Enriched Features for Fast Image Restoration and Enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1934–1948. https://doi.org/10.1109/tpami.2022.3167175.
15. Xu, X.; Wang, R.; Fu, C.-W.; Jia, J. SNR-Aware Low-light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17693–17703. https://doi.org/10.1109/cvpr52688.2022.01719.
16. Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; Liu, J. From Fidelity to Perceptual Quality: A Semi-Supervised Approach for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3060–3069. https://doi.org/10.1109/cvpr42600.2020.00313.
17. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z. EnlightenGAN: Deep Light Enhancement Without Paired Supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. https://doi.org/10.1109/tip.2021.3051462.

18. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1777–1786. https://doi.org/10.1109/cvpr42600.2020.00185.
19. Li, C.; Guo, C.; Loy, C.C. Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4225–4238. https://doi.org/10.1109/tpami.2021.3063604.
20. Liu, R.; Ma, L.; Zhang, J.; Fan, X.; Luo, Z. Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10556–10565. https://doi.org/10.1109/cvpr46437.2021.01042.
21. Ma, L.; Ma, T.; Liu, R.; Fan, X.; Luo, Z. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5627–5636. https://doi.org/10.1109/cvpr52688.2022.00555.
22. Ma, L.; Ma, T.; Xu, C.; Liu, J.; Fan, X.; Luo, Z.; Liu, R. Learning With Self-Calibrator for Fast and Robust Low-Light Image Enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 9095–9112. https://doi.org/10.1109/tpami.2025.3586712.
23. Shi, Y.; Liu, D.; Zhang, L.; Tian, Y.; Xia, X.; Fu, X. ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 3015–3024. https://doi.org/10.1109/cvpr52733.2024.00291.
24. Krull, A.; Buchholz, T.-O.; Jug, F. Noise2Void - Learning Denoising From Single Noisy Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2124–2132. https://doi.org/10.1109/cvpr.2019.00223.
25. Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; Aila, T. Noise2Noise: Learning Image Restoration without Clean Data. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Proceedings of Machine Learning Research; Volume 80, pp. 2965–2974. https://proceedings.mlr.press/v80/lehtinen18a.html.
26. Huang, T.; Li, S.; Jia, X.; Lu, H.; Liu, J. Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14776–14785. https://doi.org/10.1109/cvpr46437.2021.01454.
27. Qu, L.; Zhao, S.; Huang, Y.; Ye, X.; Wang, K.; Liu, Y.; Liu, X.; Mao, H.; Hu, G.; Chen, W.; Guo, C.; He, J.; Tan, J.; Li, H.; Chen, L.; Zhao, W. Self-inspired learning for denoising live-cell super-resolution microscopy. *Nat. Methods* **2024**, *21*, 1895–1908. https://doi.org/10.1038/s41592-024-02400-9.
28. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. https://doi.org/10.1109/34.192463.
29. Jin, Y.; Yang, W.; Tan, R.T. Unsupervised Night Image Enhancement: When Layer Decomposition Meets Light-Effects Suppression. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 404–421. https://doi.org/10.1007/978-3-031-19836-6_23.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
31. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. https://doi.org/10.1109/lsp.2012.2227726.
32. Venkatanath, N.; Praneeth, D.; Bh, M.C.; Channappayya, S.S.; Medasani, S.S. Blind image quality evaluation using perception based features. In Proceedings of the 21st National Conference on Communications (NCC), Mumbai, India, 27 February–1 March 2015; pp. 1–6. https://doi.org/10.1109/ncc.2015.7084843.
33. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. https://doi.org/10.1109/tip.2012.2214050.
34. Chobola, T.; Liu, Y.; Zhang, H.; Schnabel, J.A.; Peng, T. Fast Context-Based Low-Light Image Enhancement via Neural Implicit Representations. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 413–430. https://doi.org/10.1007/978-3-031-73016-0_24.
35. Li, H.; Hu, X.; Wang, H. Interpretable Unsupervised Joint Denoising and Enhancement for Real-World Low-Light Scenarios. In Proceedings of the 13th International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025. Available online: https://openreview.net/forum?id=PVHoELf5UN (accessed on 20 October 2026).
36. Sun, S.; Ren, W.; Peng, J.; Song, F.; Cao, X. DI-Retinex: Digital-Imaging Retinex Model for Low-Light Image Enhancement. *Int. J. Comput. Vis.* **2025**, *133*, 8293–8314. https://doi.org/10.1007/s11263-025-02542-z.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.