*remote sensing*

MDPI

*Article*

# LL-CSFormer: A Novel Image Denoiser for Intensified CMOS Sensing Images under a Low Light Environment

**Xin Zhang, Xia Wang * and Changda Yan**

Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing Institute of Technology, Beijing 100081, China; 3120205296@bit.edu.cn (X.Z.); 3120195335@bit.edu.cn (C.Y.)
* Correspondence: angelniuniu@bit.edu.cn

**Abstract:** Intensified complementary metal-oxide semiconductor (ICMOS) sensors can capture images under extremely low-light conditions ($\leq$0.01 lux illumination), but the results exhibit spatially clustered noise that seriously damages the structural information. Existing image-denoising methods mainly focus on simulated noise and real noise from normal CMOS sensors, which can easily mistake the ICMOS noise for the latent image texture. To solve this problem, we propose a low-light cross-scale transformer (LL-CSFormer) that adopts multi-scale and multi-range learning to better distinguish between the noise and signal in ICMOS sensing images. For multi-scale aspects, the proposed LL-CSFormer designs parallel multi-scale streams and ensures information exchange across different scales to maintain high-resolution spatial information and low-resolution contextual information. For multi-range learning, the network contains both convolutions and transformer blocks, which are able to extract noise-wise local features and signal-wise global features. To enable this, we establish a novel ICMOS image dataset of still noisy bursts under different illumination levels. We also designed a two-stream noise-to-noise training strategy for interactive learning and data augmentation. Experiments were conducted on our proposed ICMOS image dataset, and the results demonstrate that our method is able to effectively remove ICMOS image noise compared with other image-denoising methods using objective and subjective metrics.

**Keywords:** low light; intensified CMOS image; image denoising; cross-scale transformer; two-stream noise-to-noise



**Citation:** Zhang, X.; Wang, X.; Yan, C. LL-CSFormer: A Novel Image Denoiser for Intensified CMOS Sensing Images under a Low Light Environment. *Remote Sens.* **2023**, *15*, 2483. https://doi.org/10.3390/rs15102483

Academic Editor: Benoit Vozel

Received: 13 March 2023
Revised: 4 May 2023
Accepted: 6 May 2023
Published: 9 May 2023



## 1. Introduction

Night vision technology uses optoelectronics to capture images under low light conditions. Owing to the limitation of the human eye, people cannot accurately identify the detailed features of objects under extremely low illumination. Intensified charge-coupled devices (ICCDs) and intensified complementary metal-oxide semiconductor (ICMOS) devices combine a CCD or CMOS sensor with an image intensifier tube. They can acquire images under extremely low-light conditions, requiring less power than other types of night vision devices, and they are relatively inexpensive.

The primary drawback of ICMOS devices is that the image intensifier amplifies the intensity of noise while enhancing the signal, resulting in obvious scintillation noise in the acquired image. Owing to the crosstalk effect of microchannel plates, ICMOS image noise is not independent and identically distributed (IID), but spatially clustered noise [1,2], which is more complex than that of normal CMOS images. This kind of noise seriously destroys the original structural features of the image and greatly increases the difficulty of image denoising and noise modeling.

Many approaches have been proposed for image denoising. Existing denoising methods can be categorized into spatial-domain, transform-domain, sparse-representation, and deep learning-based methods. Spatial-domain methods are mainly aimed at the independent and identical distribution of natural image noise. They adopt filters to remove noise,

such as mean filtering, bilateral filtering [3], or non-local means (NLM) [4]. However, these methods are not suitable for spatially clustered noise in ICMOS images. Transform-domain methods firstly apply a specific transformation to the noisy image, and then process the transform coefficients according to the difference between noise and signal in the transform domain. Examples of transform-domain methods include the Fourier transform, wavelet transform [5], and block-matching and 3D filtering (BM3D) [6]. These methods usually consider noise as high-frequency components and image signals as low-frequency components. However, spatially clustered noise includes both low- and high-frequency components, so it is difficult for transform-domain methods to process ICMOS images. Sparse representation methods sparsely represent a noisy image through an overcomplete atomic library, separating the image from the noise by sparsity using techniques such as principal component analysis (PCA) [7], learned simultaneous sparse coding (LSSC) [8], multi-scale weighted group sparse coding model (MS-WGSC) [9] or latent low-rank representation (LatLRR) [10]. However, these methods are computationally expensive and inefficient for denoising low-light grayscale images.

Deep learning methods separate noise and image signals by learning the distribution characteristics of noise from a large number of noisy data samples. These methods have been widely used for image denoising in recent years [11–17]. Many studies demonstrate that learning-based methods outperform traditional methods for image denoising, especially for real image noise from different imaging devices. DnCNN [11] proposed a residual end-to-end denoising network for additive Gaussian noise. CBDnet [12] designed a convolutional blind denoising network for real photographs. VDN [13] integrated image denoising into a Bayesian framework to remove non-IID noise in real scenarios. DANet [14] could simultaneously deal with both the noise removal and noise generation tasks. GR2R [16] could obtain the noise model through a residual image and a random mask. DeamNet [17] incorporated adaptive consistency into the network design. However, to the best of our knowledge, no learning-based denoising method has been reported for ICMOS sensing images, and existing deep learning-based denoising methods only consider independent and identically distributed image noise.

Recently, transformer and its variants have made significant breakthroughs in computer vision tasks [18–21], demonstrating their ability to model global and long-range dependencies more powerfully than CNNs. Consequently, transformers have also been used for image restoration. For instance, Uformer [22] introduced a novel locally-enhanced window (LeWin) transformer block for window-based self-attention. SwinIR [23] was the first to employ the Swin transformer for image restoration. Restormer [24] proposed several key designs in the building blocks to capture long-range pixel interactions. However, these methods are somewhat weak in obtaining local information. Figure 1 shows an example of comparison with different methods.

To achieve ICMOS image denoising, we propose a novel image denoiser called a low-light cross-scale transformer (LL-CSFormer). Due to the introduction of the microchannel plate, the noise from ICMOS images tends to be spatially clustered and unevenly scaled, which is different from the independent identically distributed (IID) noise in simulated images and real images from normal CMOS sensors. General image-denoising methods do not work well on ICMOS images, because the random scale and pattern of the clustered noise may be mistaken as the true image structure with these methods. To this end, we propose a novel cross-scale transformer network that effectively separates noise and signal through multi-scale and multi-range learning. Firstly, unlike existing deep learning-based methods with only a full-resolution pipeline or progressively low-resolution pipeline, we designed a cross-scale structure to maintain the high-resolution spatial representation and ensure rich semantic information from low-resolution representation. On the one hand, the multi-scale features here can help to extract unevenly scaled noise in ICMOS sensing images, and on the other hand, they can help to discover the latent texture in the image by rich semantic information. Secondly, the proposed network introduces multi-range learning, combining the core mechanisms of short-range model convolution and long-

range model transformer. This combination enables the network to extract both noise-wise local features and signal-wise global features in ICMOS sensing images. Additionally, we established an ICMOS image dataset of still noisy bursts by obtaining images from a direct-coupled camera under different illumination levels and scenes. To enable the training pipeline, we designed a two-stream noise-to-noise strategy that inputs noisy image pairs of the same scene for interactive learning and data augmentation.

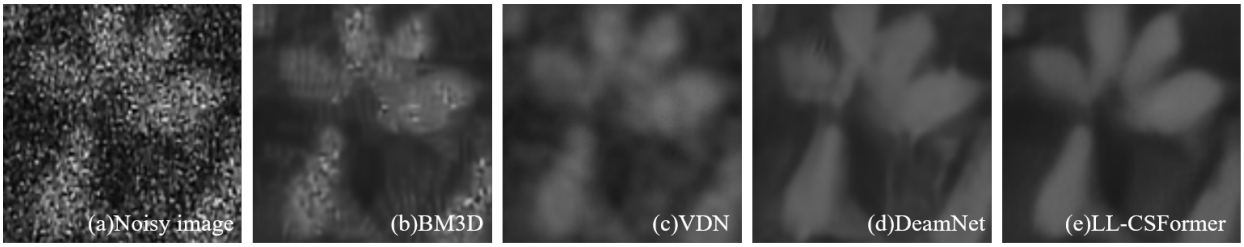



**Figure 1.** An example of an ICMOS noisy image. Compared with other image-denoising methods, our LL-CSFormer performs the best.

Our contributions can be summarized as follows:

- We propose an image denoiser for ICMOS sensing images called a low-light cross-scale transformer (LL-CSFormer).
- Considering the spatially clustered noise from ICMOS sensing images, we designed a cross-scale transformer network to effectively separate the signal and noise by multi-scale and multi-range learning.
- We established a novel ICMOS image dataset of still noisy bursts under different illumination levels and scenes.
- Extensive experiments conducted on our proposed dataset demonstrate that our method outperforms existing state-of-the-art image-denoising methods for ICMOS sensing image denoising.

The remainder of this paper is organized as follows. In Section 2, we introduce the principle of intensified CMOS imaging system. The pipeline and details of our proposed method are given in Section 3. In Section 4, we provide the experimental results and analysis. We further discuss the results and findings in Section 5. Finally, the conclusions are presented in Section 6.

## 2. Intensified CMOS Imaging System

ICMOS is a night vision device composed of an intensifier coupled with a CMOS sensor. Figure 2 shows an example of an image intensifier and a directly-coupled ICMOS camera. This device integrates the high sensitivity of night vision direct-view imaging devices and the camera function of TV imaging devices. The principle of ICMOS imaging is more complicated than normal CMOS imaging. As shown in Figure 3, the intensifier consists of three parts: a photocathode, a microchannel plate, and a phosphor screen. Photons enter the lens, and the weak light signal is converted into an electronic image after photoelectric conversion by the photocathode. The generated electrons are then injected into the microchannel plate to obtain energy multiplication. The electrons are projected onto the fluorescent phosphor screen and converted into an optical image, and finally, the CMOS sensor captures the light signal from the phosphor screen to generate the final image.

Owing to the complexity of the ICMOS structure, the ICMOS noise model is different from that of natural images, and includes components from four sources: the photocathode, MCP, phosphor screen, and CMOS sensor. In general, due to the crosstalk effect caused by a microchannel plate, ICMOS image noise has two main characteristics. (1) The ICMOS noise is not independent and identically distributed (IID) but spatially clustered; (2) The ICMOS noise exhibits a randomly clustered pattern, which varies widely in scale [1]. This kind of noise greatly destroys the structural information of the latent image and even causes

a lot of unreasonable textures. Moreover, ICMOS noise is unevenly distributed in the temporal dimension and has strong randomness, which manifests as serious scintillation noise. As shown in Figure 4, we make a comparison of ICMOS image noise under $10^{-3}$ lx and Gaussian noise with $\sigma = 55$. The two images have the same background and the ICMOS image comes from the proposed dataset. It can be seen that the ICMOS image contains unexpected image textures and the scales appear random, while original structure information can still be seen in the image polluted by Gaussian noise. In this paper, we mainly focus on the spatially clustered noise from ICMOS images.


**Figure 2.** An example of an ICMOS camera.

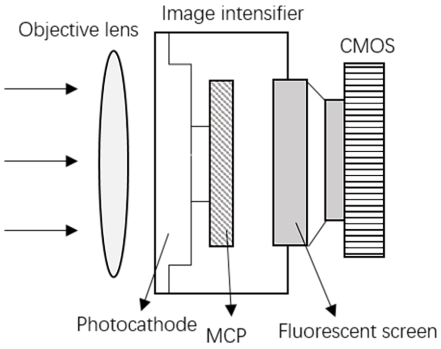
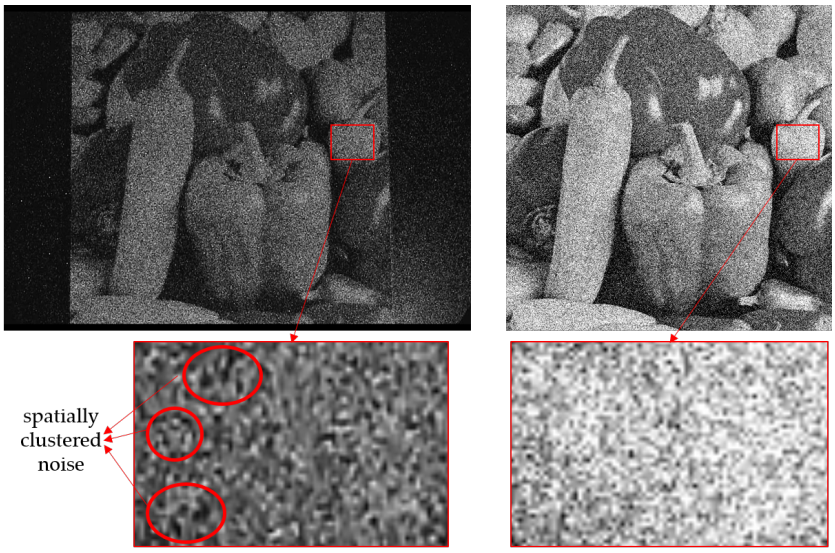



**Figure 3.** ICMOS imaging pipeline.




(a) ICMOS image noise under $10^{-3}lx$ (b) Gaussian noise with $\sigma = 55$

**Figure 4.** Comparison of ICMOS image noise and Gaussian noise at similar amplification.

## 3. Proposed Method

In this section, we provide the details of our proposed low-light cross-scale transformer (LL-CSFormer) for ICMOS image denoising.

### *3.1. Low Light Cross-Scale Transformer*

Due to the introduction of the microchannel plate, the noise in ICMOS images tends to be spatially clustered and unevenly scaled. To remove the complex noise and recover the latent texture in ICMOS sensing images, our network combines two core components: (1) multi-scale learning, which uses a cross-scale structure to extract multi-scale image features and ensure information interaction between different scales; (2) multi-range learning: a combination of short-range dependency convolution and long-range dependency transformer to extract local noise features and global signal features. Vision science has found that the local neuronal receptive fields of the human eye are of different sizes, demonstrating the importance of multi-scale information in neural networks. The cross-scale structure not only employs parallel multi-scale streams for fine-to-coarse and coarse-to-fine feature representations, but also allows for efficient extraction of multi-scale image features while ensuring the interaction of high- and low-resolution information at each step. In each scale, we designed a novel hybrid transformer module (HTM) as our sub-module, which combines the core mechanisms of convolution and transformer to capture both local and global image features. In ICMOS images, local image features mainly represent noise, while global image features mainly represent the signal. Hence, the proposed model effectively separates the signal and noise in ICMOS sensing images.

The entire network structure is illustrated in Figure 5. We denote a noisy ICMOS input image as $y$. Firstly, the input $y$ is encoded by a head module, which converts the input image from 3 channels to a 32-dimensional feature map through 2 convolutional layers and a PReLU activation function.

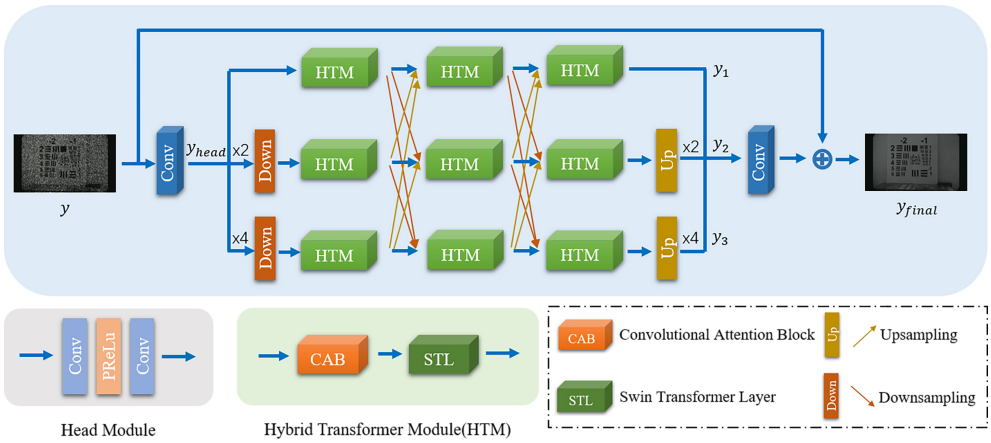



**Figure 5.** Cross-scale transformer structure.

In order to introduce multi-scale features, we employ the up- and downsampling operators to change the resolution of image features $y_{head}$, where we use a convolutional layer with a stride size of two for downsampling and a bilinear interpolation operator for upsampling. As shown in Figure 5, the three streams receive $y_{head}$ of different scales directly as input. We consider the three adjacent HTMs at different scales as one cross-scale step. After each step, three new results of different scales will be outputted, which will be transformed to the other two streams separately by up- or downsampling operators and added to the other two transformed results as the new input to the next step. Each HTM consists of a convolutional attention block (CAB) for local feature extraction and a Swin transformer layer (STL) for global feature extraction.

After obtaining the results of the multi-scale streams $y_1, y_2, y_3$, we concatenate them and pass through a convolution module, which is the same as the head module to achieve the final denoised image.

Hybrid Transformer Module (HTM)

The hybrid transformer module's core is multi-range learning, which considers both local image details and global contextual information for ICMOS image denoising. Each

HTM contains a CAB and an STL. The CAB is a CNN-based module in our network, and its schematic is shown in Figure 6. It extracts local image features through short-range dependency convolutional streams. Motivated by modern low-level vision tasks [25–27], we add spatial and channel attention modules based on the Res-block [28]. The attention modules share information within a feature tensor in terms of both spatial and channel dimensions, allowing the CAB to extract informative local features while suppressing redundant ones. The spatial/channel attention module generates a spatial/channel attention map by averaging and globally pooling to rescale the input feature map.

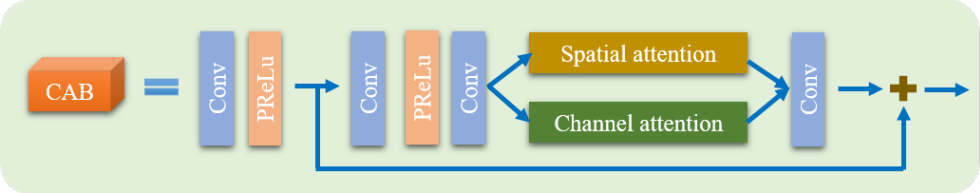



**Figure 6.** The structure of the convolutional attention block (CAB).

As mentioned earlier, capturing long-range dependencies and global image priors are important for image denoising. To achieve this, we adopt the Swin transformer layer (STL) [23] to extract global image features. The structure of the STL is illustrated in Figure 7. The STL is the improved version of classic multi-head self-attention [18]. The core of STL is the shifted window attention, which shows great promise for vision tasks. First, the input feature map $x_{in} \in R^{H \times W \times C}$ will be reshaped to a new tensor of size $\frac{HW}{M^2} \times M^2 \times C$ by partitioning it into non-overlapping local windows of size $M \times M$. We set $M = 8$ in our model. Then, the self-attention mechanism is computed in each window. For each window feature $x_i \in R^{M^2 \times C}, i = 1, \ldots, N$, the feature query $Q_i$, key $K_i$, and value $V_i$ can be formulated as

$$Q_i = x_i F_Q, K_i = x_i F_K, V_i = x_i F_V \tag{1}$$

where $F_Q, F_K, F_V \in R^{C \times d}$ are projection matrices for all of the local windows. Then, the classic self-attention matrix is calculated by

$$SA(Q_i, K_i, V_i) = softmax(\frac{Q_i \cdot K_i^T}{\sqrt{d}} + B) \cdot V_i \tag{2}$$

Here, $B$ is a learnable parameter for position coding and d is the dimension of the key feature. Abiding by the multi-head self-attention mechanism, we apply self-attention $h$ times in parallel for various attention distributions. Here, we set $h = 3$ in our work. Finally, a feedforward network (FFN) consisting of two fully connected layers and Gaussian error linear units (GELU) is employed for feature extraction. Layer normalization (LN) and residual skipping connections are performed before both MSA and FFN.

$$\begin{aligned} Z_i &= WMSA(LN(Q_i, K_i, V_i)) + x_i \\ Z_i &= FFN(LN(Z_i)) + Z_i \end{aligned} \tag{3}$$

The window multi-head self-attention mechanism only applies self-attention within each window, which neglects cross-window information. To address this limitation, we introduce a shifted window multi-head self-attention module (SWMSA) by shifting the window location by $(\left\lfloor \frac{M}{2} \right\rfloor, \left\lfloor \frac{M}{2} \right\rfloor)$ pixels during the partitioning process. This mechanism promotes information interaction between different windows by shifting their positions. Similar to WMSA, SWMSA can be formulated as

$$\begin{aligned} Z_i &= SWMSA(LN(Q_i^Z, K_i^Z, V_i^Z)) + Z_i \\ Z_i &= FFN(LN(Z_i)) + Z_i \end{aligned} \tag{4}$$

Furthermore, the receptive field of the self-attention mechanism in STL is limited in the two fixed windows by window shifting. Through multi-scale learning, the scale of the receptive field will be extended.

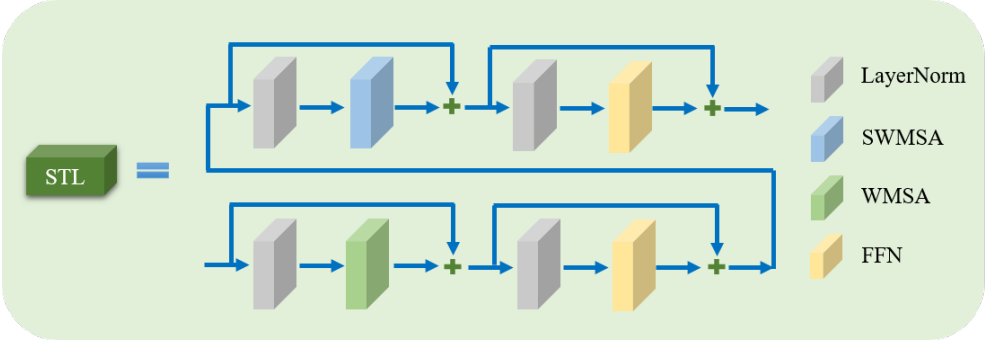



**Figure 7.** The structure of the Swin transformer layer.

### *3.2. Two-Stream Noise-to-Noise Training Strategy*

In general, acquiring pairs of noisy and clean images as data samples is necessary when training image denoising neural networks. However, obtaining high-quality noise-free ground truth images can be difficult for special imaging devices. In this study, we propose a novel two-stream noise-to-noise training strategy to address this problem, which is a variant of the noise-to-noise pipeline. The noise-to-noise pipeline attempts to relax the requirement of supervised mechanisms from noisy/clean pairs to noisy/noisy pairs, and performs almost as well as a noise-to-clean pipeline [29]. The theoretical premise behind this approach is that the noise is zero-mean and the noise from ICMOS images meets this requirement based on frame integral experiments. In this method, pairs of samples in the training dataset are all noisy images of the same scene, the two images can be formulated as

$$\begin{cases} y = x + n \\ y' = x + n' \end{cases} \tag{5}$$

where $n$ and $n'$ are different ICMOS image noises, and $x$ is the latent clean image. Then, the noise-to-clean pipeline trains the network by minimizing the empirical risk:

$$argmin \sum_i \left\{ \|F(y_i) - x_i\|_2^2 \right\} \tag{6}$$

where $F()$ denotes the denoising network. The noise-to-noise pipeline proves that the clean data $x$ used as training targets can be replaced with noisy images $y'$ without changing what the network learns, which can be formulated as

$$argmin \sum_i \left\{ \|F(y_i) - y_i'\|_2^2 \right\} \tag{7}$$

Then, we have

$$\begin{aligned} & argmin \sum_i \left\{ \|F(y_i) - y_i'\|_2^2 \right\} \\ =& argmin \sum_i \left\{ \|F(y_i) - x_i - n_i'\|_2^2 \right\} \\ =& argmin \sum_i \left\{ \|F(y_i) - x_i\|_2^2 - 2(n_i')^\top (F(y_i) - x_i) + (n_i')^\top n_i \right\} \\ =& argmin \sum_i \left\{ \|F(y_i) - x_i\|_2^2 \right\} - 2argmin \sum_i \left\{ (n_i')^\top F(y_i) \right\} + c \end{aligned} \tag{8}$$

Since the noise $n$ and $n'$ are uncorrelated and zero-mean, the second term can be simplified as $argmin \sum_i \{(n_i')^\top F(y_i)\} = 0$. Then, we can see that the empirical risk of the noise-to-noise pipeline is equivalent to that of the noise-to-clean pipeline, except for a constant $c$.

As shown in Figure 8a, the noise-to-noise training strategy is a special case where the ground truths for the regression task are noisy images instead of the desired clean images. In our ICMOS image denoising task, we have a limited dataset of still noisy bursts. The core of our proposed two-stream training strategy is to use paired noisy samples as both input and target, leading to interactive learning. In this case, the number of noisy samples for the denoiser is doubled, allowing for data augmentation. However, overfitting is a common issue during the noise-to-noise training strategy [30], mainly due to the use of noisy samples, such as the labels. Given the two images, i.e., $y$ and $y'$, the overfitting result is such that $F(y)$ estimates $y'$ and $F(y')$ estimates $y$. Based on the two-stream training strategy, this problem can be solved by measuring the difference between $F(y)$ and $F(y')$. In other words, if the denoiser reaches the best performance, the difference between $F(y)$ and $F(y')$ should be minimal, because $y$ and $y'$ contain the same latent image $x$.

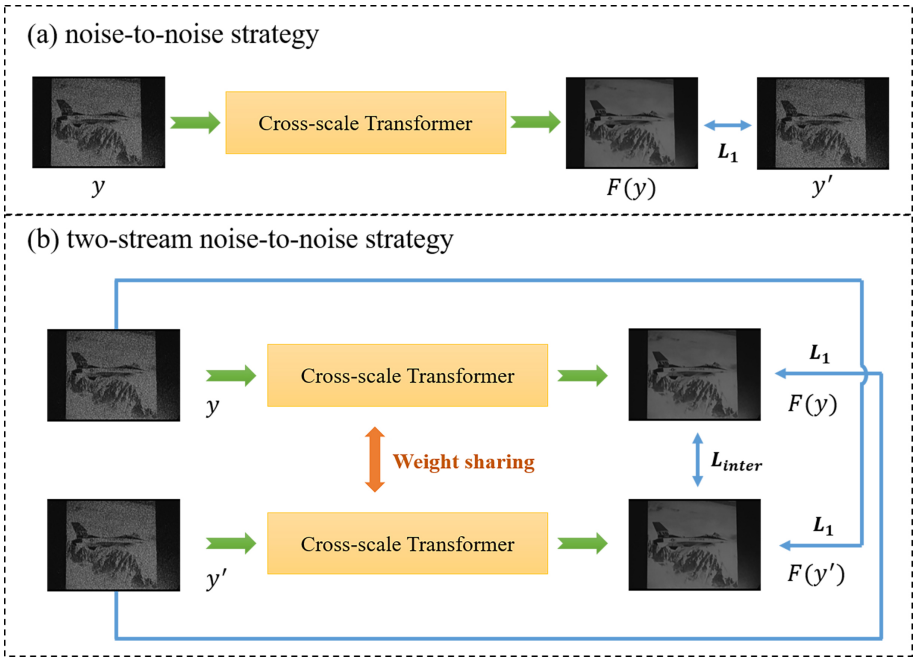



**Figure 8.** The pipeline of the noise-to-noise training strategy and two-stream noise-to-noise training strategy.

Our proposed two-stream noise-to-noise training strategy is illustrated in Figure 8b. After receiving $y$ and $y'$ as input, the two-stream multi-scale transformer produces $F(y)$ and $F(y')$ as the output, respectively. The two-stream networks share their weights and parameters during training. Therefore, only one stream network is utilized during testing. In addition to the losses between $F(y)$ and $y'$ and between $F(y')$ and $y$, we also consider the loss between $F(y)$ and $F(y')$ to avoid the overfitting problem, which is called interactive loss in our model. All of the losses employed in our model are Charbonnier losses [31]. The total loss is formulated as

$$L_1 = \sqrt{\|y - F(y')\|^2 + \varepsilon} + \sqrt{\|y' - F(y)\|^2 + \varepsilon} (\varepsilon = 10^{-6}) \tag{9}$$

$$L_{inter} = \sqrt{\|F(y) - F(y')\|^2 + \varepsilon} \tag{10}$$

$$L_{total} = \lambda_1 L_1 + \lambda_{inter} L_{inter} \tag{11}$$

where $\lambda_1$ and $\lambda_{inter}$ are the corresponding coefficients. After the experiments, we empirically set $\lambda_1 = 1$ and $\lambda_{inter} = 0.1$ in our model.

## 4. Experimental Results and Analysis

### *4.1. ICMOS Noisy Image Dataset*

Existing learning-based image-denoising methods have always focused on synthetic or real data captured by normal CMOS sensors. In this study, we first collect a real ICMOS image dataset for training. We adopted a directly-coupled ICMOS camera that couples a 25 mm diameter image intensifier to a Canon EOS M3 sensor [32]. This camera is able to capture videos under $10^{-3}$ lx with $1920 \times 1080$ pixels in spatial resolution. We captured still image bursts directly on the camera display at 30 fps. Our dataset is divided into two parts, i.e., indoor scenes and outdoor scenes. The indoor data collection experiment was carried out in a dark room. We used an integrating sphere to control scene illumination, and an illuminometer was used to measure the illuminance of the low-light scene accurately. Figure 9 shows the indoor experiment scene. The outdoor data were collected in the urban night environment, and we also used the illuminometer to measure the illuminance of the target scene. The setting of our dataset is shown in Table 1.

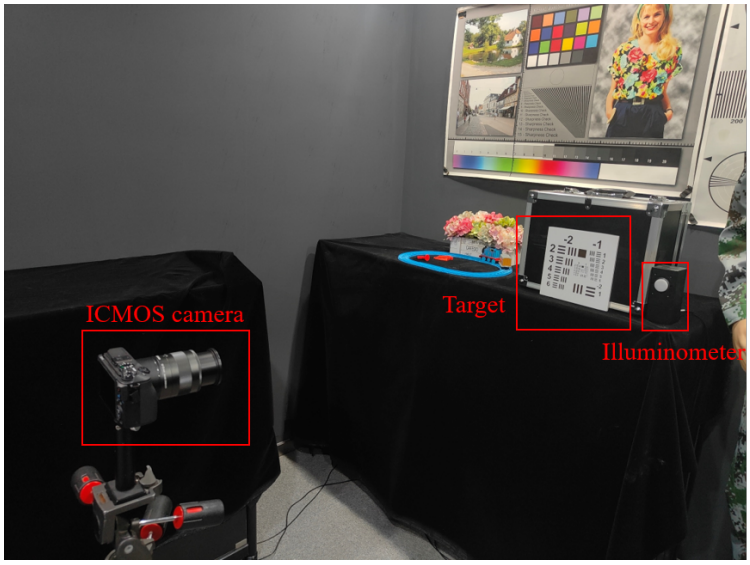



**Figure 9.** The indoor data collection experiment scene.

To enable the two-stream noise-to-noise training strategy, we need to capture a large number of image bursts, and all the scenes in our dataset must be static. To better compare the performance of our method, we applied the frame integral algorithm to these images to obtain relatively clean results for each scene sequence. Examples of noisy images and corresponding clean images from the indoor part are shown in Figure 10. We captured images under two illumination levels, namely $1 \times 10^{-3}$ lx and $1 \times 10^{-2}$ lx. For each illumination level, we captured video sequences of 20 static scenes, with each sequence consisting of 1000 frames. These images have the same underlying scene but different noise distributions. Each video was processed using a frame's integral algorithm to obtain a corresponding clean image. Therefore, this indoor dataset contains a total of 20 clean images and 20,000 noisy images. We randomly selected 5 video sequences as the training set, where 5000 images were equally divided into 2 parts for input and ground truth to satisfy our noise-to-noise training strategy. Next, we randomly extracted 5 noisy images from each of the remaining 15 video sequences as the validation set, which consisted of 75 test images in total. Examples of noisy images and corresponding clean images from the outdoor part are shown in Figure 11.

We captured video sequences of 15 different static scenes in the urban night environment. Similar to the previous indoor scenes, we have a total of 15,000 images and 15 corresponding clean images. Due to the unevenness of urban night lighting, the illumination range of the target scene is between $2 \times 10^{-2}$ lx and $1 \times 10^{-1}$ lx. As shown in these figures, the frame integral images basically exclude noise components, and the detailed information of the image itself is well preserved, so we will later employ frame integral images as ground truth images for evaluation.

$$y_{clean} = \frac{y_1 + y_2 + \cdots + y_N}{N} \tag{12}$$

where $N = 1000$ in our experiment.

**Table 1.** Setting of our ICMOS image dataset.

| Dataset | Location | Illumination Level | Total Scenes | Total Images |
|---|---|---|---|---|
| part1 | indoor | $1 \times 10^{-2}$ lx | 20 | 20,000 |
| | | $1 \times 10^{-3}$ lx | 20 | 20,000 |
| part2 | outdoor | $2 \times 10^{-2}$ lx–$1 \times 10^{-1}$ lx | 15 | 15,000 |

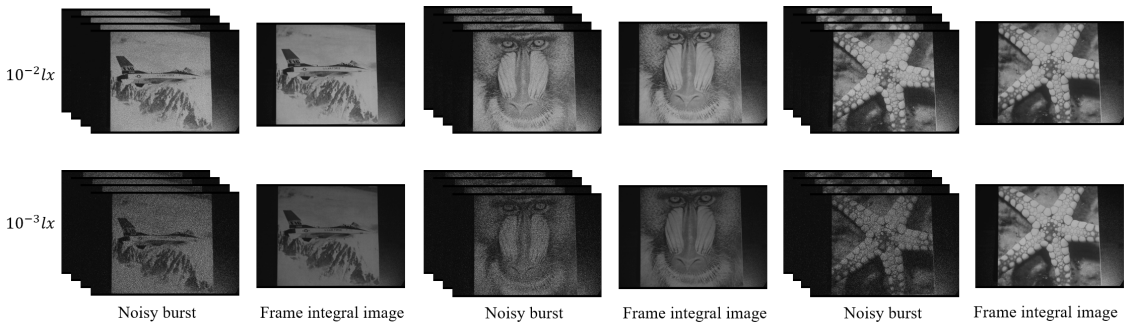



**Figure 10.** Examples of noisy images and corresponding clean images from indoor scenes.

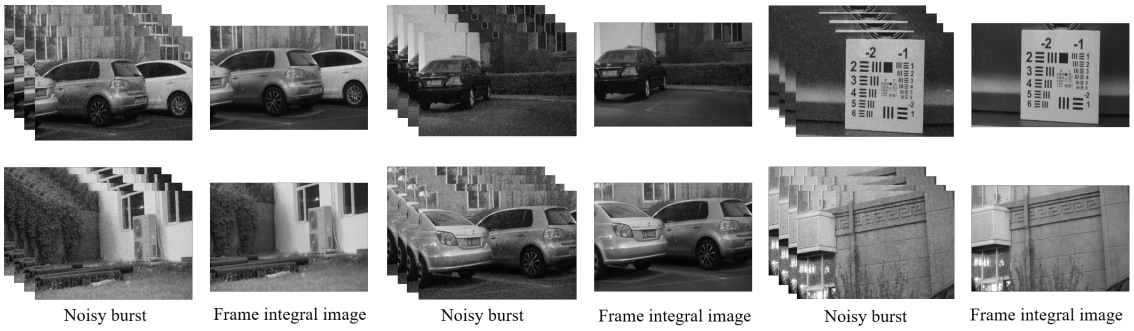



**Figure 11.** Examples of noisy images and corresponding clean images from outdoor scenes.

### *4.2. Implementation Details*

We adopted the Adam optimization method to optimize the parameters, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial value of the learning rate is $2 \times 10^{-4}$, with decay by cosine annealing as training progresses. The images in the dataset have a uniform size of $1920 \times 1080$, but the images are randomly cropped to $128 \times 128$ when training. We also apply random rotation and flipping to the image patches for data augmentation. All experiments were performed using two NVIDIA GeForce RTX 3090 GPUs. To compare the performance of our method with others, we use the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) for an objective evaluation.

### *4.3. Ablation Study*

In this section, we describe several ablation experiments conducted to verify the effectiveness of each component of our model.

#### 4.3.1. The Analysis of Cross-Scale Structure

To verify the effectiveness of the cross-scale structure, we remove the cross-scale pipeline from the LL-CSFormer. This means that the network is performed in a fixed resolution, and the three-hybrid transformer module is conducted in the only branch. The results in Figure 12c show that the denoising performance is very poor without the cross-scale structure. It also achieves the lowest scores of PSNR and SSIM in Table 2, indicating that the cross-scale structure is crucial for ICMOS image denoising.

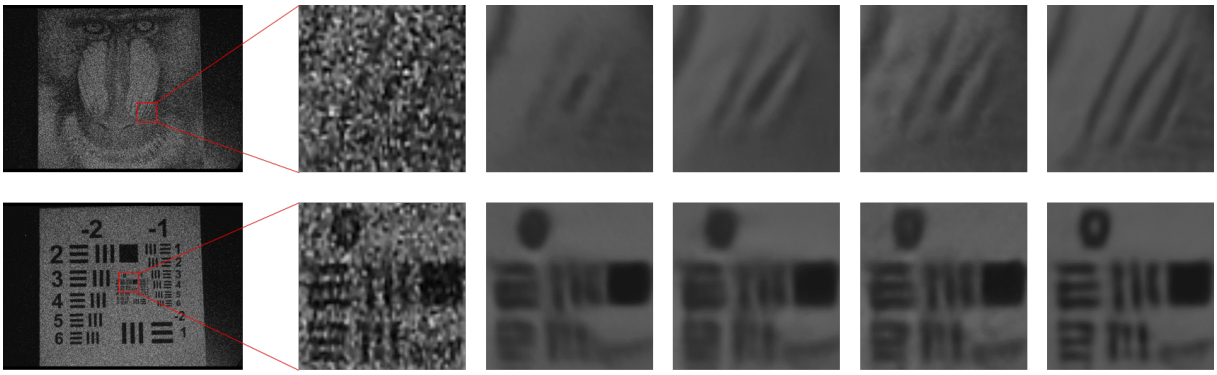



**Figure 12.** Results of the ablation study.

#### 4.3.2. The Analysis of STL

To test the performance of the STL, we replace all of the STL in the network with CAB. As can be seen in Figure 12d, the noise is removed relatively well with the CAB, but the images are oversmoothed and the latent texture details are severely lost. The results in Figure 12f indicate that STL can help the model recover image texture details by extracting global features. Table 2 demonstrates that the PSNR increases by an average of 0.46 dB with the help of STL. So, STL is an essential part in our model.

**Table 2.** Quantitative results under different configurations ('CS' means cross-scale structure, 'w/o' means without).

| | CS | STL | CAB | PSNR/dB (↑) | | SSIM (↑) | |
|---|---|---|---|---|---|---|---|
| | | | | $10^{-2}$ lx | $10^{-3}$ lx | $10^{-2}$ lx | $10^{-3}$ lx |
| w/o CS | | ✓ | ✓ | 33.63 | 33.45 | 0.8595 | 0.8880 |
| w/o STL | ✓ | | ✓ | 34.13 | 33.74 | 0.8857 | 0.9004 |
| w/o CAB | ✓ | ✓ | | 33.95 | 33.42 | 0.8810 | 0.8833 |
| LL-CSFormer | ✓ | ✓ | ✓ | **34.50** | **34.29** | **0.8904** | **0.9077** |

#### 4.3.3. The Analysis of CAB

CAB plays the role of extracting local noise features from images. To verify its effectiveness, we replaced all CAB with STL (similar to before). As can be seen from Figure 12e, it retains more image details compared to Figure 12d, but it still suffers from noise contamination. Whereas the results in Figure 12f achieve the best denoising effects among all experiments. After the introduction of CAB, the PSNR rose by an average of 0.71 dB. It is clear that CAB also significantly improves the performance of the network.

### *4.4. Results and Analysis*

To verify the performance of our LL-CSFormer, we compare it to several state-of-the-art image-denoising methods, including NLM [33], BM3D [6], CBDnet [12], VDN [13], DANet [14], DeamNet [17], and Uformer [22]. For the deep learning-based methods, we used the publicly available source code provided by the authors and trained them using our ICMOS noisy dataset. Due to the specificity of the dataset in this paper, we used the noise-to-noise training strategy for the other methods here.

#### 4.4.1. Visual Comparison

Figures 13–16 show a visual comparison of indoor images of these methods, it is clear that the ICMOS image noise becomes worse as illumination decreases. We find that NLM is the least effective among all of the methods, as the results show that NLM can barely remove the noise from the ICMOS image. BM3D, the best conventional image denoising method, is able to perform well at $10^{-2}$ lx illumination, but cannot work at $10^{-3}$ lx illumination. As shown in Figures 15c and 16c, many unreasonable bright spots and artifacts still exist in the pictures. CBDnet, VDN, and DeamNet exhibit poor generalization performance for the spatially clustered noise of ICMOS images, and tend to destroy image edges and textures during denoising across different illumination levels. These models are unable to distinguish noise and texture clearly. DANet is able to restore some image textures, but also introduces artifacts such as striped textures as shown in Figures 13f and 15f. Uformer is able to remove the noise component, but still damages the image details as shown in Figure 15h, and produces ripple artifacts as shown in Figures 13h and 14h.

Figures 17 and 18 show the visual comparisons from outdoor images. It can be seen that NLM, BM3D, and VDN have difficulty completely removing noise from ICMOS images. The processed images still contain strong noise interference. CBDnet, DANet, and DeamNet over-smooth the signals in the images, even producing artifacts. As shown in Figure 17d,g, the striped texture on the building was directly blurred. Although Uformer can restore image details, it still cannot effectively restore texture details. There is a noticeable ghosting effect at the wall joint in Figure 18h.

In comparison, our LL-CSFormer exhibits a strong denoising performance for all images, removing image noise to the maximum extent and simultaneously retaining image details.

#### 4.4.2. Quantitative Comparison

We also compare our method to other state-of-the-art methods in terms of objective evaluation metrics. Table 3 shows the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) results of each method on the test set. It is worth noting that the overall indicators of indoor scene images will be higher than outdoor scenes due to the relatively simple background. Similar to the subjective evaluation, with the traditional methods, NLM achieved the lowest scores among all the methods in terms of PSNR and SSIM. Due to the limitations of the model, CBDnet, VDN, and DeamNet also scored poorly in terms of PSNR and SSIM. Since BM3D performs well at $10^{-2}$ lx illumination, the scores are relatively high. In general, the results demonstrate that our proposed LL-CSFormer outperforms other methods in all cases. At the same time, we compare the parameters and running times of different methods in Table 4. The running time experiments were performed by averaging 75 images of size $1920 \times 1080$. For NLM and BM3D, the codes are operated by CPU, while the deep learning-based methods are operated by GPU. As shown in Table 3, our method has a minimal number of parameters with only 0.97M. In terms of running time, traditional methods have huge time costs. Due to the introduction of transformer mechanism, our method operates slightly slower than CBDnet and VDN.

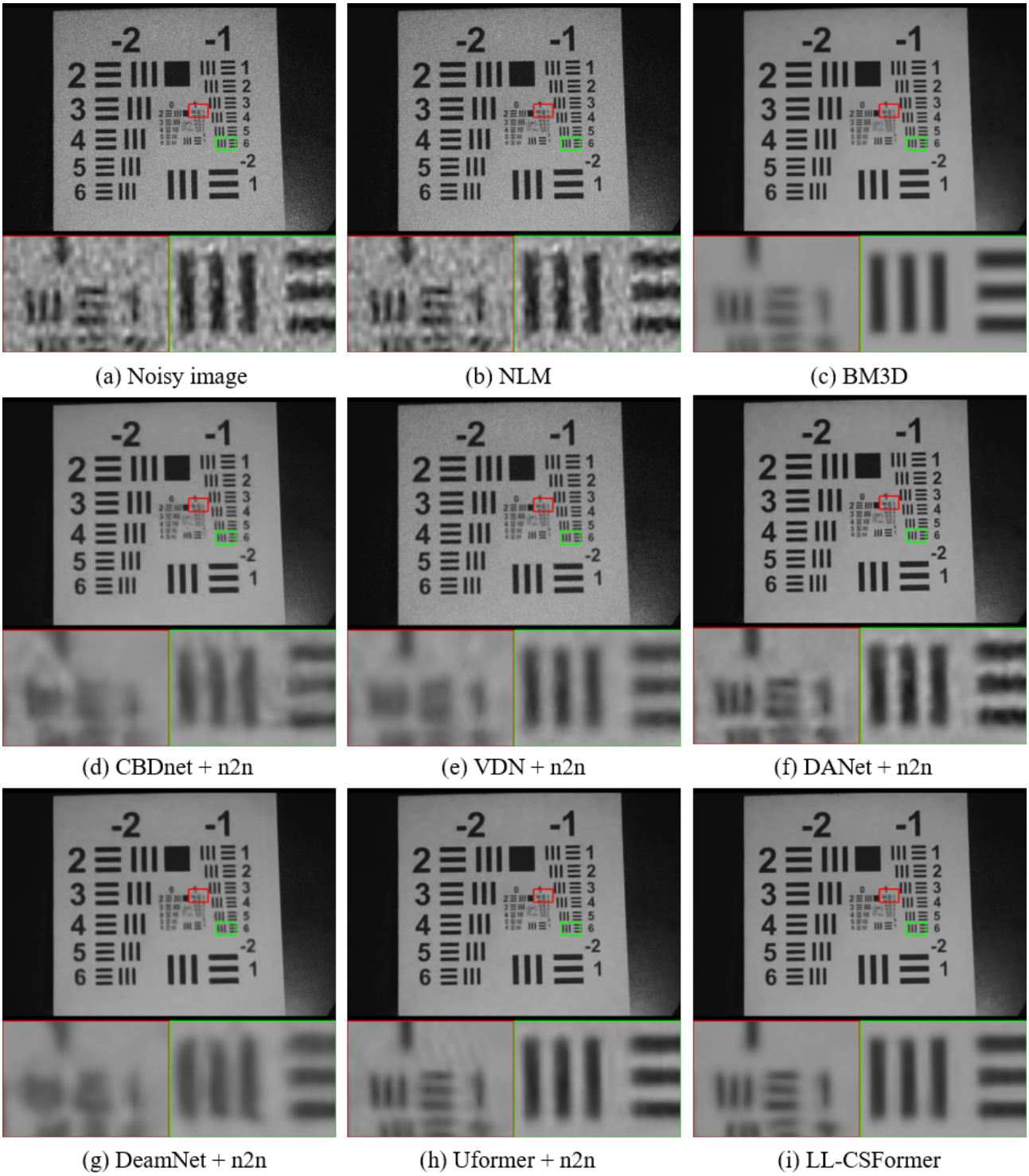



**Figure 13.** Visual comparison results under $10^{-2}$ lx; 'n2n' means the noise-to-noise training strategy.

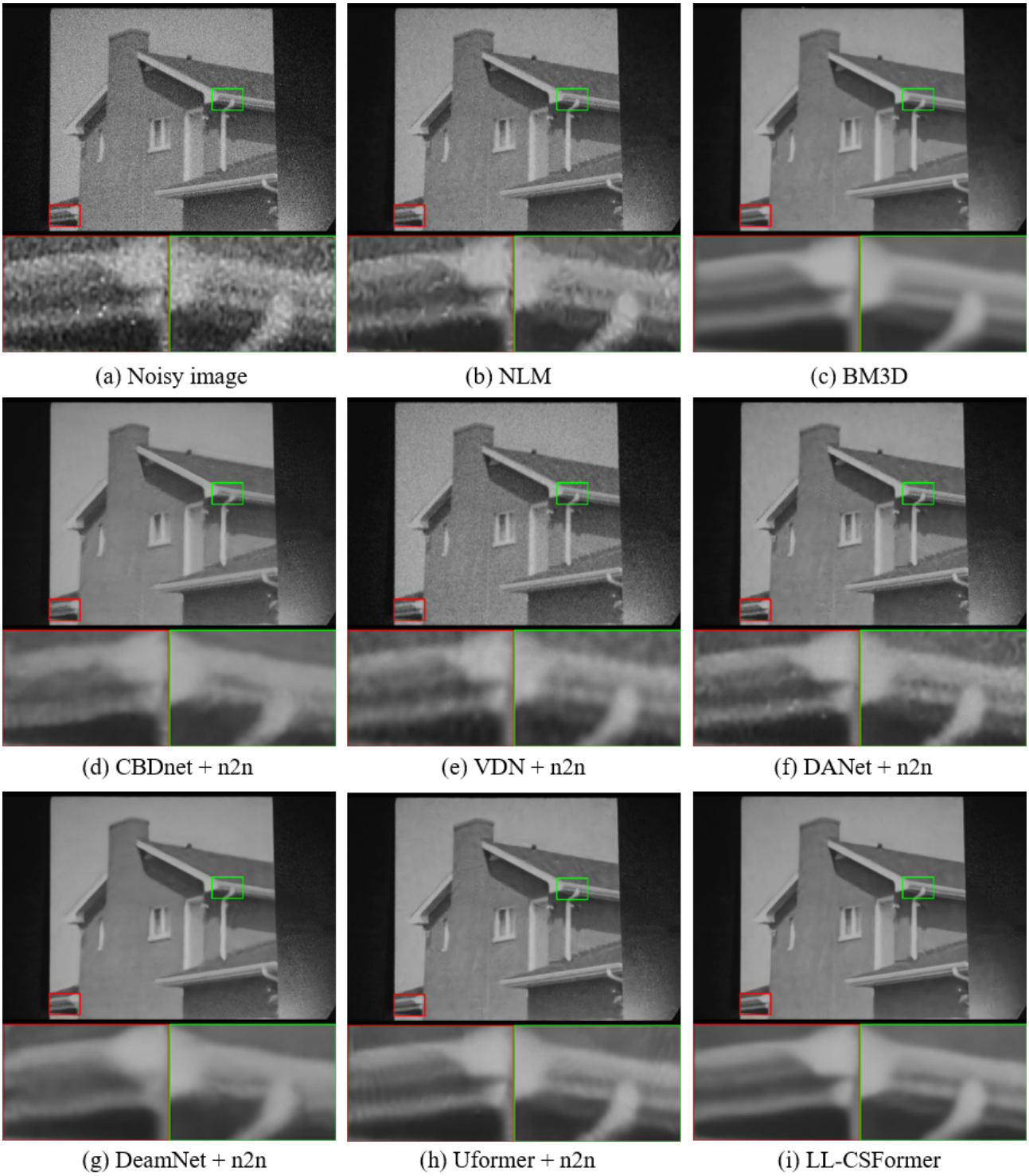



**Figure 14.** Visual comparison results under $10^{-2}$ lx; 'n2n' means noise-to-noise training strategy.

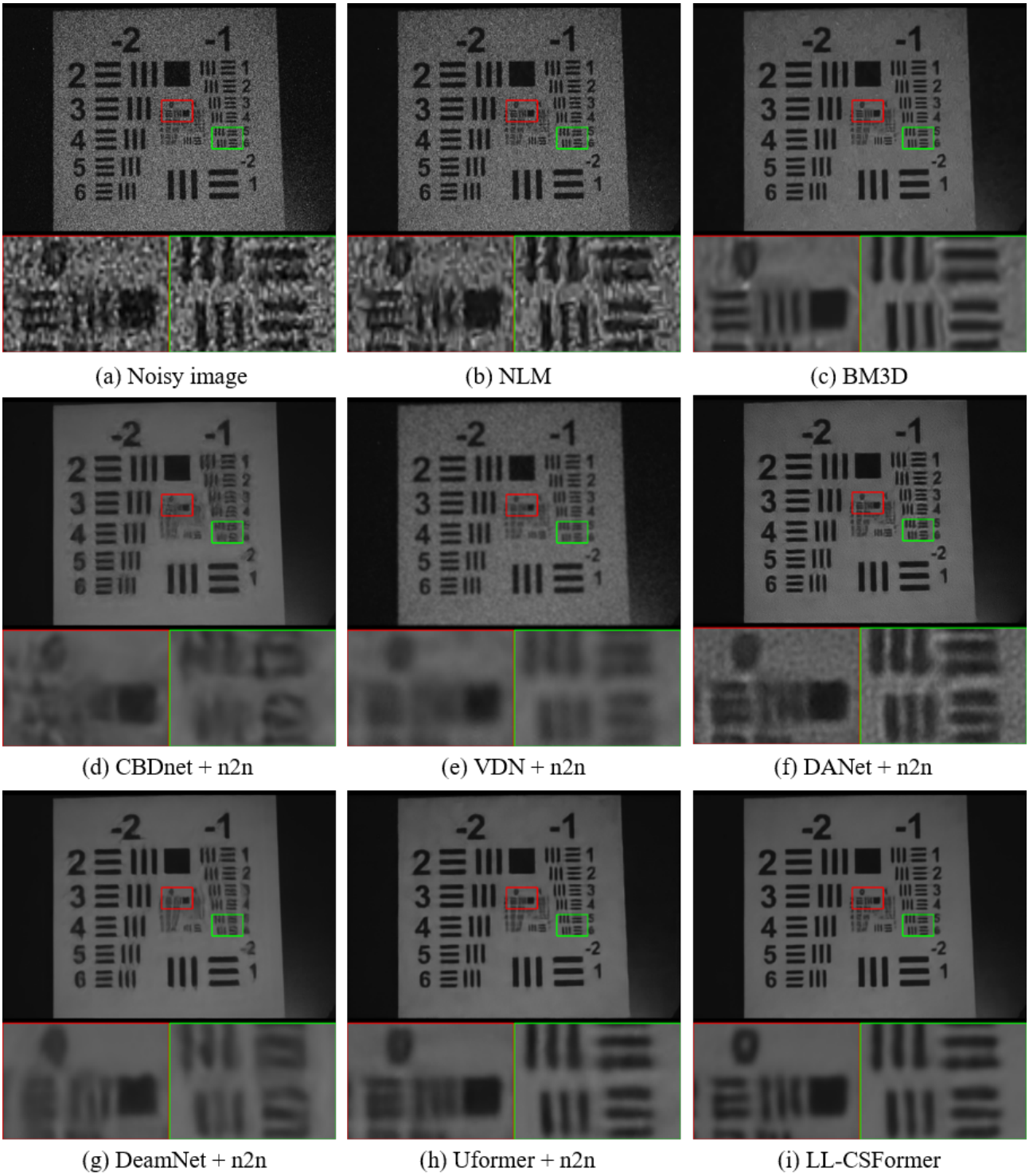



**Figure 15.** Visual comparison results under $10^{-3}$ lx; 'n2n' means noise-to-noise training strategy.

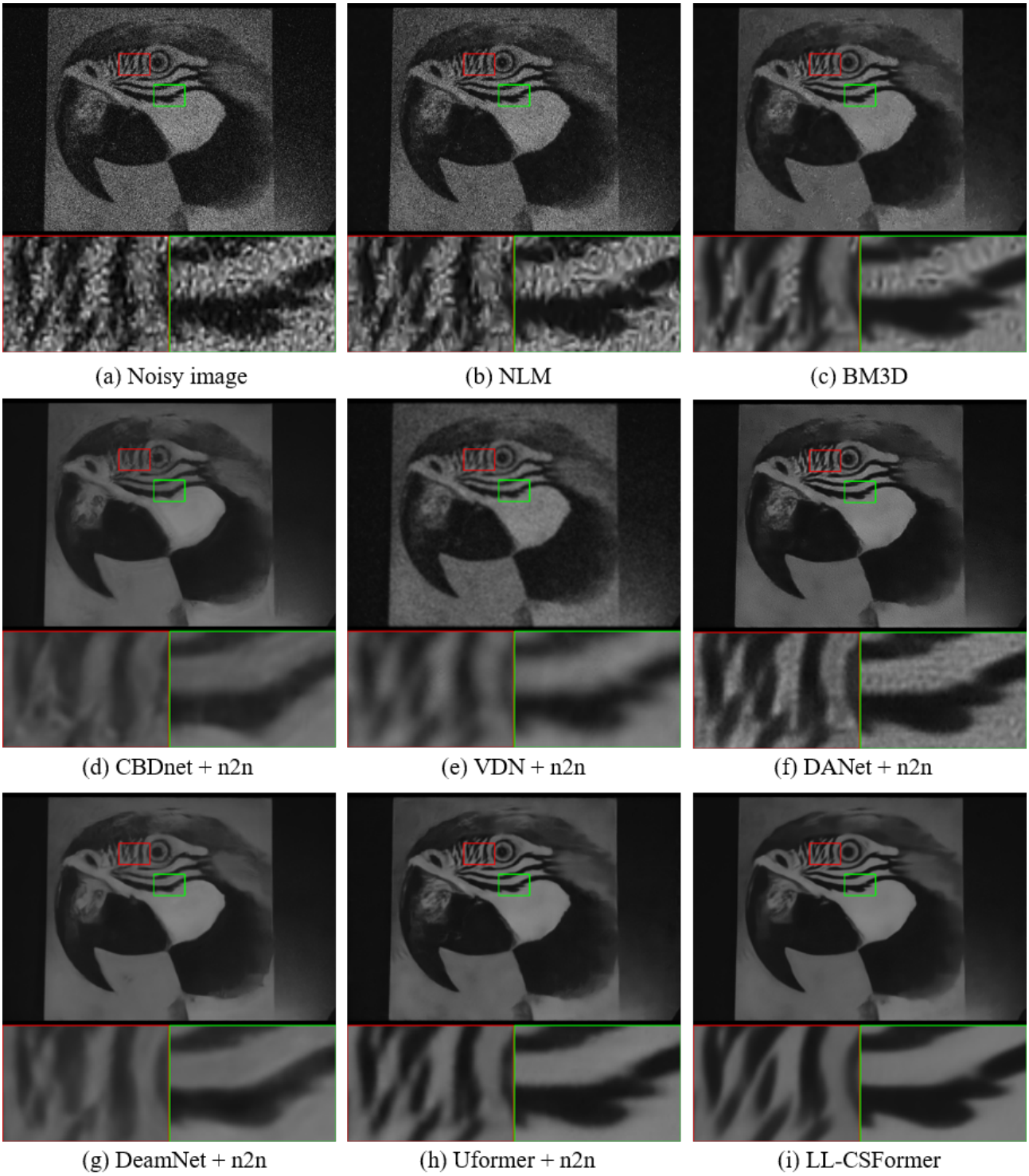



**Figure 16.** Visual comparison results under $10^{-3}$ lx; 'n2n' means the noise-to-noise training strategy.

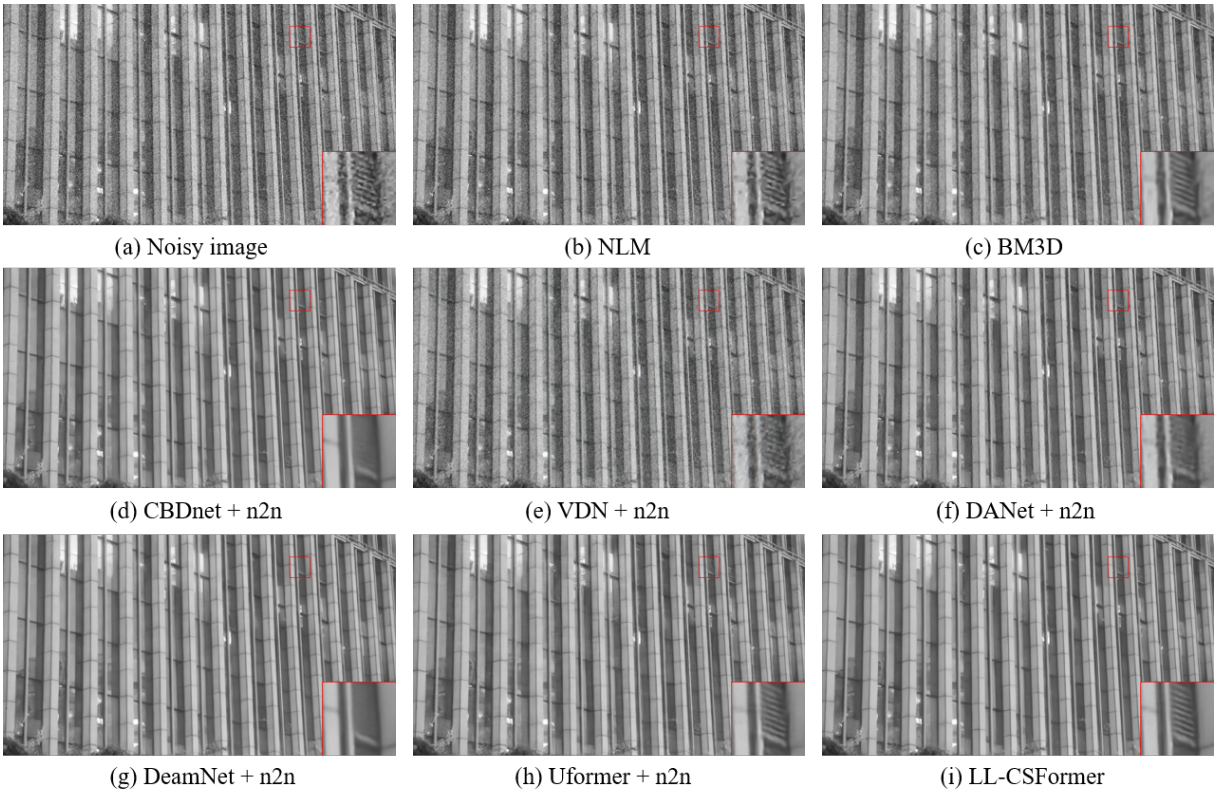



**Figure 17.** Visual comparison results from the outdoor scene; 'n2n' means noise-to-noise training strategy.

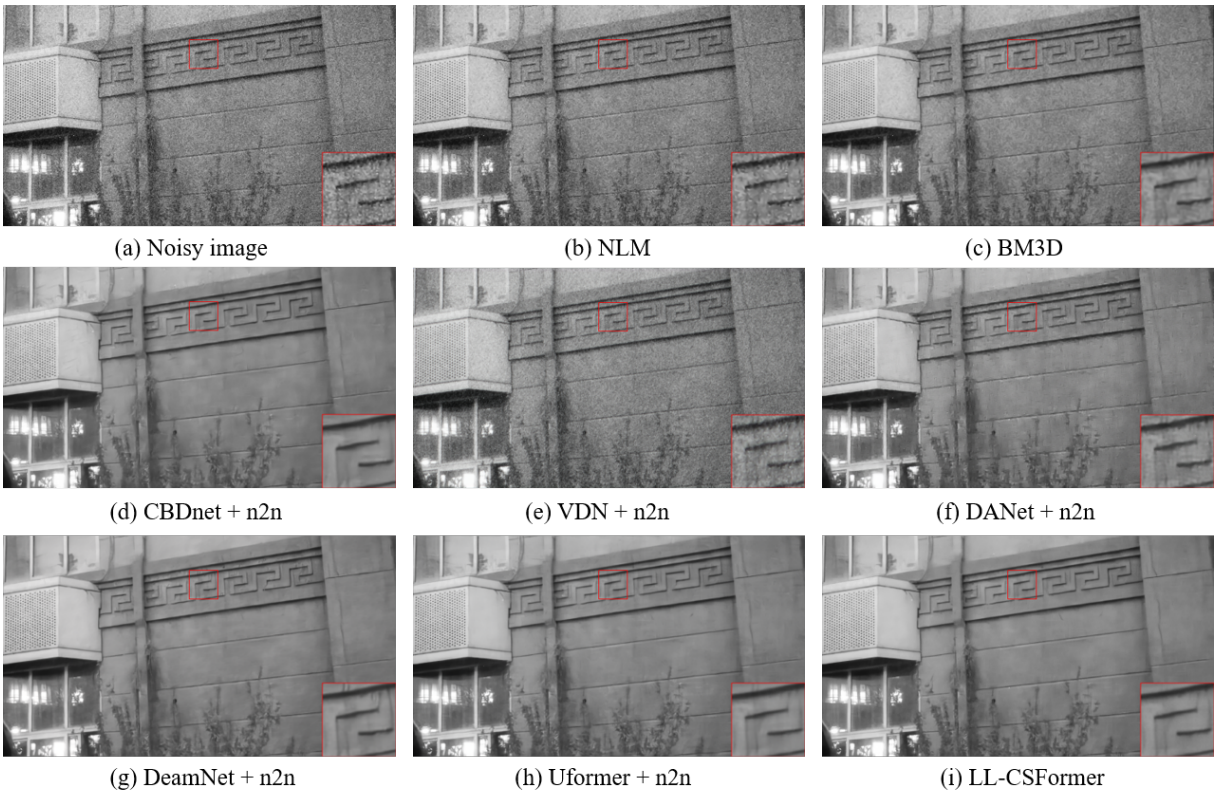



**Figure 18.** Visual comparison results from the outdoor scene; 'n2n' means the noise-to-noise training strategy.

**Table 3.** Objective evaluation of different methods.

| Methods | PSNR/dB (↑) | | | SSIM (↑) | | |
|---|---|---|---|---|---|---|
| | $10^{-2}$ lx | $10^{-3}$ lx | Outdoor | $10^{-2}$ lx | $10^{-3}$ lx | Outdoor |
| NLM | 30.39 | 22.62 | 18.92 | 0.7753 | 0.5255 | 0.5559 |
| BM3D | 33.28 | 31.66 | 18.00 | 0.8725 | 0.8389 | 0.6947 |
| CBDnet | 31.82 | 31.49 | 22.03 | 0.8667 | 0.8794 | 0.7564 |
| DANet | 32.81 | 30.88 | 21.08 | 0.8407 | 0.7580 | 0.6870 |
| VDN | 31.23 | 31.38 | 19.12 | 0.8241 | 0.8479 | 0.5341 |
| DeamNet | 32.34 | 32.85 | 22.04 | 0.8745 | 0.8937 | 0.7595 |
| Uformer | 33.98 | 33.89 | 22.07 | 0.8791 | 0.8975 | 0.7639 |
| ours | **34.50** | **34.29** | **22.24** | **0.8904** | **0.9077** | **0.7666** |

**Table 4.** Parameters and running times of different methods.

| Methods | Parameters (M) | Running Time (s) |
|---|---|---|
| NLM | / | 163.2 |
| BM3D | / | 2236 |
| CBDnet | 4.3 | 0.124 |
| DANet | 9.2 | 0.318 |
| VDN | 7.8 | **0.109** |
| DeamNet | 2.2 | 0.369 |
| Uformer | 50.9 | 0.886 |
| ours | **0.97** | 0.186 |

## 5. Discussion

In this section, we will further analyze the experimental results and findings of this paper.

1. ICMOS image noise is spatially clustered with a strong spatial correlation, which differs from the independent and identically distributed noise in natural images. As shown in Figure 4, the noise in ICMOS images significantly degrades the image details and introduces undesirable textures. Currently, denoising methods designed for real-world natural noises struggle to remove ICMOS image noise. In contrast, our method introduces a cross-scale and multi-range learning approach that can effectively extract the characteristics of ICMOS image noise and achieve optimal denoising results.

2. As the illumination decreases, the noise intensity of ICMOS images increases and the denoising effects of different methods become worse. As shown in Section 4, the illumination of the scene seriously affects the noise intensity of ICMOS. Some algorithms can effectively remove noise interference under the $10^{-2}$ lx environment, but their performances significantly deteriorate under $10^{-3}$ lx and in outdoor environments. However, our method achieves the best denoising effects under different illumination conditions. Our method can also handle the situation of uneven illumination in outdoor urban scenes.

## 6. Conclusions

In this study, we propose a learning-based low-light cross-scale transformer (LLCSFormer) for denoising ICMOS sensing images. To remove the spatially clustered and unevenly scaled ICMOS noise, we introduce multi-scale and multi-range learning. The proposed cross-scale transformer structure is able to extract multi-scale features and ensures information exchange across different scales. In each scale, we employ both convolutions and transformer blocks to extract noise-wise local features and signal-wise global contextual information. We also establish a novel ICMOS image dataset of still noisy bursts under different illumination levels to enable network training and evaluation. The two-stream noise-to-noise training strategy proposed in this paper offers a new trick in the study of denoising of specific devices. The experimental results show that our proposed method can remove the ICMOS image noise under different illumination levels, and also greatly restore

the corrupted detail features of ICMOS sensing images. We performed several ablation studies to verify the effectiveness of our core component in the model. Comparisons with other state-of-the-art image-denoising methods demonstrate the superiority of our method in both visual effects and objective quality metrics.

**Author Contributions:** Conceptualization, X.Z. and X.W.; methodology, X.Z.; software, X.Z.; validation, X.Z.; formal analysis, X.Z. and C.Y.; investigation, X.Z.; resources, X.Z.; data curation, X.Z. and C.Y.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z.; visualization, X.Z.; supervision, X.Z. and X.W.; project administration, X.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: No. 62031018.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors wish to thank the editors and the reviewers for their valuable suggestions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Wang, F.; Wang, Y.; Yang, M.; Zhang, X.; Zheng, N. A denoising scheme for randomly clustered noise removal in ICCD sensing image. *Sensors* **2017**, *17*, 233. [CrossRef] [PubMed]
2. Yang, M.; Wang, F.; Wang, Y.; Zheng, N. A denoising method for randomly clustered noise in ICCD sensing images based on hypergraph cut and down sampling. *Sensors* **2017**, *17*, 2778. [CrossRef] [PubMed]
3. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 7 January 1998; pp. 839–846.
4. Zhang, W.; Li, J.; Yang, Y. A fractional diffusion-wave equation with non-local regularization for image denoising. *Signal Process.* **2014**, *103*, 6–15. [CrossRef]
5. Su, Y.; Xu, Z. Parallel implementation of wavelet-based image denoising on programmable PC-grade graphics hardware. *Signal Process.* **2010**, *90*, 2396–2411. [CrossRef]
6. Dabov, K.; Foi, A.; Katkovnik, V.; Egiazarian, K. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **2007**, *16*, 2080–2095. [CrossRef] [PubMed]
7. Bui, A.T.; Im, J.K.; Apley, D.W.; Runger, G.C. Projection-free kernel principal component analysis for denoising. *Neurocomputing* **2019**, *357*, 163–176. [CrossRef]
8. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. Non-local sparse models for image restoration. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 2272–2279.
9. Ou, Y.; Swamy, M.; Luo, J.; Li, B. Single image denoising via multi-scale weighted group sparse coding. *Signal Process.* **2022**, *200*, 108650. [CrossRef]
10. Nie, T.; Wang, X.; Liu, H.; Li, M.; Nong, S.; Yuan, H.; Zhao, Y.; Huang, L. Enhancement and Noise Suppression of Single Low-Light Grayscale Images. *Remote Sens.* **2022**, *14*, 3398. [CrossRef]
11. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef] [PubMed]
12. Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; Zhang, L. Toward convolutional blind denoising of real photographs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1712–1722.
13. Yue, Z.; Yong, H.; Zhao, Q.; Meng, D.; Zhang, L. Variational denoising network: Toward blind noise modeling and removal. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1690–1701.
14. Yue, Z.; Zhao, Q.; Zhang, L.; Meng, D. Dual adversarial network: Toward real-world noise removal and noise generation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 41–58.
15. Cheng, S.; Wang, Y.; Huang, H.; Liu, D.; Fan, H.; Liu, S. Nbnet: Noise basis learning for image denoising with subspace projection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4896–4906.
16. Liu, Y.; Wan, B.; Shi, D.; Cheng, X. Generative Recorrupted-to-Recorrupted: An Unsupervised Image Denoising Network for Arbitrary Noise Distribution. *Remote Sens.* **2023**, *15*, 364. [CrossRef]
17. Ren, C.; He, X.; Wang, C.; Zhao, Z. Adaptive consistency prior based deep network for image denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8596–8606.
18. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.

19. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Yang, S.; Quan, Z.; Nie, M.; Yang, W. Transpose: Keypoint localization via transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11802–11812.
22. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 17683–17693.
23. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
24. 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, New Orleans, LA, USA, 19–20 June 2022; pp. 5728–5739.
25. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Learning enriched features for real image restoration and enhancement. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 492–511.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
27. Zhang, X.; Wang, X. MARN: Multi-Scale Attention Retinex Network for Low-Light Image Enhancement. *IEEE Access* **2021**, *9*, 50939–50948. [CrossRef]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
29. Pang, T.; Zheng, H.; Quan, Y.; Ji, H. Recorrupted-to-Recorrupted: Unsupervised Deep Learning for Image Denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2043–2052.
30. Calvarons, A.F. Improved Noise2Noise denoising with limited data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 796–805.
31. Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; Barlaud, M. Two deterministic half-quadratic regularization algorithms for computed imaging. In Proceedings of the 1st International Conference on Image Processing, Austin, TX, USA, 13–16 November 1994; Volume 2, pp. 168–172.
32. Gao, T.; Cao, F.; Wang, X.; Cui, Z. Direct Coupling of Low Light Image Intensifier with Large Size CMOS. *Infrared Technol.* **2021**, *43*, 537–542.
33. Buades, A.; Coll, B.; Morel, J.M. A non-local algorithm for image denoising. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 60–65.