Efficient and Robust SEM Image Denoising for Wafer Defect Inspection

HyunWoong Bae[1], JaeSeok Byun[2], YongWoo Lee[3], and Taesup Moon[1,2,4]

[1] *Interdisciplinary Program in Artificial Intelligence, Seoul National University, Gwanak-ro, 08826, South Seoul, Korea*
[2] *Department of Electrical and Computer Engineering, Seoul National University, Gwanak-ro, 08826, Seoul, South Korea*
[3] *Samsung Electronics, Samsung-ro, 144, Pyeongteak, Gyeonggi-do, South Korea*
[4] *Artificial Intelligence Institute of Seoul National University (AIIS)/Automation and Systems Research Institute (ASRI)/Institute of New Media and Communications (INMC), Seoul National University, Gwanak-ro, 08826, Seoul, South Korea*
Corresponding Author: TaeSup Moon <tsmoon@snu.ac.kr>

Abstract
Noise in scanning electron microscopy (SEM) often obscures details critical for accurate wafer defect inspection. Deep learning-based denoising methods have been widely used to address this problem, but they have two major limitations in SEM image denoising: lack of both efficient and powerful denoising methods, and poor generalization to image structures that are unseen during training. To this end, we propose Relaxed Noise2Noise with Input dropout (ReNIn), which includes components that address the above two issues. Firstly, our Relaxed Noise2Noise (RN2N) framework provides a much better trade-off between denoising performance and training data collection costs; namely, it shows nearly on-par denoising performance with much lower data collection cost than ordinary supervised learning-based methods. Secondly, we propose to apply the *input dropout* to boost generalization ability. It improves the performance of the images that are structurally different from the training images without compromising the performance of the normally structured images, thereby increasing the overall denoising performance. Consequently, our method supports downstream inspection tasks by reducing the failure rate in circle detection, which is a critical preprocessing step for circle-shaped product analysis. Overall, our ReNIn attains efficient training data collection cost with competitive denoising performance and enhances generalization capability across various structures.
Key Words: SEM, Denoising, Deep Learning, Denoising for inspection, (Received XX Y 20ZZ; revised XX Y 20ZZ; accepted XX Y 20ZZ)

Introduction

The Scanning Electron Microscope (SEM) has served as a standard tool for identifying manufacturing defects in semiconductor wafers (Dey et al., 2021). The raw images captured from the SEM are often degraded by noise, making denoising an essential step for fine-grained defect inspection. Traditionally, SEM image denoising is achieved by capturing multiple frames of the same image and averaging them; a method that is highly expensive, time-consuming, and potentially damaging the image samples (Giannatou et al., 2019). To address these limitations, several deep learning-based SEM image denoising techniques have been developed, which only require forward passes of neural networks during denoising.

Following the promising results of convolutional neural network (CNN)-based denoisers (Burger et al., 2012; Mao et al., 2016; Zhang et al., 2017; Tai et al., 2017; Lefkimmiatis, 2018) on natural noisy images from standard cameras, initial studies on SEM image denoising (Chaudhary et al., 2019; Ede and Beanland, 2019; Giannatou et al., 2019; Nagano et al., 2021) adopt the supervised learning framework, which utilizes clean-noisy image pairs for training. Despite promising results, obtaining sufficient clean-noisy SEM image pairs for training is highly expensive. To lift the dependency on obtaining clean SEM images, recent work (Yu et al., 2020; Sato et al., 2022) have adopted the Noise2Noise (N2N) (Lehtinen et al., 2018) training scheme in which two independent realizations of noisy images for the same source are utilized in training. Despite the cheap data acquisition cost, N2N shows inferior denoising performance compared to its supervised counterpart in SEM image denoising, as illustrated in Figure 1. Note that the supervised denoiser (Sup) and N2N are positioned at extremes where denoising performance and training data acquisition costs are either high or low. Specifically, when the raw noisy images (denoted by F01), are used as the input of the model, the supervised denoiser uses a highly clean target image, which can be typically obtained by averaging 64 frames of raw noisy images (denoted by F64), while N2N uses a single raw noisy images (F01) as a target.

As they represent two extremes, we believe there is a middle ground between them. To find a better trade-off between data collection cost and denoising performance, we

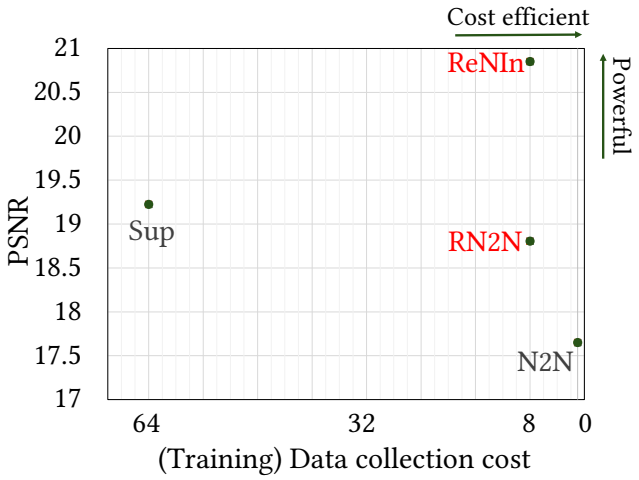


Fig. 1. Training data collection cost vs. Denoising performance (PSNR). We compare the models on our semiconductor wafer dataset (described in SEM images dataset section) using PSNR as the evaluation metric. Each number on the x-axis (training data collection cost) represents the number of raw noisy images (F#) required to obtain a single target image. For example, for a supervised denoiser, 64 frames of the raw noisy images are obtained and then averaged to get a single image (F64), which is regarded as a clean image. In the plot, a model is considered efficient and powerful as the point moves toward the upper-right direction. Here, RN2N and ReNIn denote our proposed methods.

devise the Relaxed Noise2Noise (RN2N), which relaxes the noise distribution constraint on the target image. For example, instead of using a highly clean image (F64) or highly noisy image (F01) as a target, we use a moderately clean image, such as an 8-frame averaged image (denoted by F08), as the target. The preview of the impact of our RN2N is shown in Figure 1 — it is evident that RN2N significantly reduces data collection cost (8× cheaper) while achieving competitive performance with its supervised counterpart (with a gap smaller than 0.5 dB).

Another challenge for SEM image denoising is poor generalization ability to image structures that are unseen during training. Namely, N2N shows decent denoising performance when test images are structurally similar to the training images. However, when the image structures of test images are significantly different from those of training images, N2N fails to successfully denoise them. For example, as depicted in Figure 2, the N2N model trained on regular circular images shows poor generalization ability when applied to test images with irregular and bumpy circles — note it particularly fails to preserve the edges of circles. Given the nature of the semiconductor manufacturing process where SEM images with irregular structures are frequently captured, handling *structurally different* test images is crucial for semiconductor SEM image denoising. To tackle this, we apply the input dropout mechanism which randomly drops pixels on

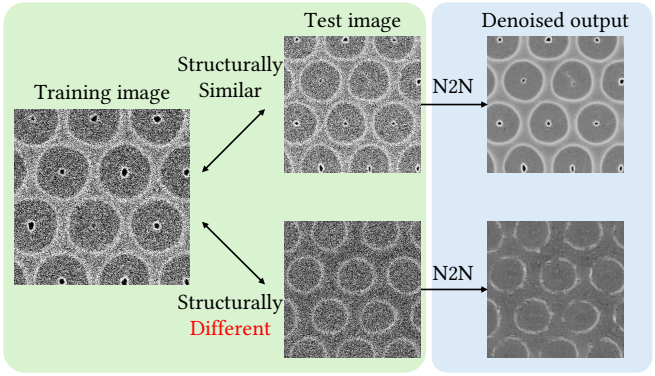


Fig. 2. Generalization issue of N2N. The green box shows the examples of the training SEM image along with two types of test images. In the green box, the upper right image represents a test image that is structurally similar to the training image (regular circles), while the lower right image denotes a structurally different test image (irregular and bumpy circles). The blue box represents the denoised images obtained from N2N. N2N fails to denoise the structurally different test image effectively, particularly in preserving edges.

the input image during training. Our experimental results demonstrate that it considerably enhances the conventional denoising metrics (PSNR/SSIM) but also in detecting circles, which is a crucial step for subsequent semiconductor defect inspection.

In summary, our final model, Relaxed Noise2Noise with Input Dropout (ReNIn), successfully addresses two major issues in semiconductor SEM image denoising: lack of efficient training consideration and poor generalization ability. In Figure 1, we indeed observe that ReNIn largely enhances the denoising performance over RN2N, even surpassing that of supervised denoiser, with a much lower data collection cost.

Related Works

[Deep learning-based denoisers without clean image] As mentioned above, to remove the requirements of clean images for supervised training, several attempts have been proposed denoisers which can solely be trained with noisy images in the natural image domain. The first approach adopts the Noise2Noise (N2N) training scheme (Lehtinen et al., 2018) in which two noisy realizations of the same source are utilized in training. This approach demonstrates competitive denoising performance with supervised counterparts in the natural image domain. However, the necessity of two noisy realizations of the exactly same source can still be impractical in real-world settings. Therefore, another approach adopts a self-supervised learning-based denoising framework (Batson and Royer, 2019; Laine et al., 2019; Krull et al., 2019; Moran et al., 2020; Xu et al., 2020; Wu et al., 2020;

Byun and Moon, 2020; Quan et al., 2020; Huang et al., 2021; Byun et al., 2021), which utilizes only *single* noisy images for training. To enable self-supervision, most works have adopted the blind spot network (BSN), which prevents the model from simply learning identity mapping by masking the central pixel in the receptive field of the noisy image. Recently, FBI-denoiser (Byun et al., 2021) achieved superior performance on real-world natural noisy image benchmark datasets and Fluorescence Microscopy Denoising dataset (Zhang et al., 2019). It proposed an efficient BSN architecture (named FBI-Net) for pixel-wise affine denoiser and a separate network that estimates Poisson-Gaussian noise parameters. A similar work to our Relaxed Noise2Noise is Moran et al. (2020). It uses two different noisy images with different noise levels for training. Their approach is to generate additional noisy images by adding known noise to single noisy images. It is similar to ours in that it uses two different noisy images, each with noise from different distributions. It assumes that the noise comes from a known noise model. However, the noise modeling for our SEM image is not straightforward since our images are obtained directly from the SEM. In addition, it is further complicated by the possible presence of spatially correlated noise, which will be discussed further in Discussion section. It makes it impractical to apply their approach to our SEM images. Our method, on the other hand, requires additional noisy images without any constraints rather than hard-to-define precise noise information.

The presence of correlated noise also demonstrates why the denoising results of the above second approach (the self-supervised denoisers) are significantly degraded in our experiments. Thus, we exclude them from the main baseline in our experiments and instead select another N2N-based training approach (Lehtinen et al., 2018). In this work, between the N2N approach and its supervised counterpart, we aim to find a better trade-off in terms of performance and data collection cost.

[Overfitting issues in low-level vision] To mitigate overfitting issues in low-level vision tasks, several works (Quan et al., 2020; Kong et al., 2022; Liu et al., 2023; Chen et al., 2023; Ma et al., 2024; Gu et al., 2024) have been proposed. For example, the dropout technique (Srivastava et al., 2014), which randomly deactivates neurons during training, has been applied to handle overfitting in *single* image denoising (Quan et al., 2020) and super-resolution (Kong et al., 2022). Similarly, recent studies (Chen et al., 2023; Ma et al., 2024) have adopted masked autoencoder (MAE), a network architecture designed to reconstruct missing or corrupted input data, to mitigate overfitting of the model to the noise distribution during training. Our proposed input dropout mechanism is closely related to both the dropout technique and MAE, but, to the best of our knowledge, it is the first application of an input masking scheme specifically focused on enhancing the generalization ability for image structures in SEM image denoising.

Preliminaries and Notations

Supervised Learning

In supervised learning, the D pairs of noisy-clean images $\{\boldsymbol{y}^i, \boldsymbol{x}^i\}_{i=0}^{D}$ are used, where the i-th noisy image is referred to as $\boldsymbol{y}^i$ and the corresponding clean image as $\boldsymbol{x}^i$. Its training objective with a learnable parameter θ for the model is given as

$$\arg\min_{\theta} \sum_{i} L(f_\theta(\boldsymbol{y}^i), \boldsymbol{x}^i) \tag{1}$$

in which L is the loss function, which typically adopts the L_2 loss, also known as the MSE loss.

Noise2Noise

Instead of using the noisy-clean image pairs for training, two noisy image pairs $\{\boldsymbol{y}^i, (\boldsymbol{y}')^i\}_{i=0}^{D}$, are utilized, where $(\boldsymbol{y}')^i$ is another noisy realization sampled from the same underlying clean image $\mathbf{x}^i$ of $\boldsymbol{y}^i$. The key difference between $\boldsymbol{y}^i$ and $(\boldsymbol{y}')^i$ lies in their noise realizations, which are assumed to be independent and identically distributed (i.i.d.) with zero mean. Under this assumption, training the neural network with these two noisy images enables the model to learn the statistical mean of the noisy images. The training objective for training the denoiser f_θ with Noise2Noise is as follows:

$$\arg\min_{\theta} \sum_{i} L(f_\theta(\boldsymbol{y}^i), (\boldsymbol{y}')^i) \tag{2}$$

While its approach relies on the assumptions of zero mean and i.i.d. noise, experimental results in (Lehtinen et al., 2018, Section 3.2 Other Synthetic Noises) demonstrate that Noise2Noise can effectively denoise even when the noise does not strictly adhere to these assumptions, such as in the case of text overlay. Additionally, it shows competitive denoising performance compared to supervised learning-based approaches, while significantly reducing the cost of data acquisition by eliminating the need for clean images during training.

Materials and Methods

In this section, we will describe our main contributions: we first present the details of the semiconductor SEM image dataset we used, then delineate the intuition and details of Relaxed Noise2Noise with Input dropout (ReNIn).

SEM images dataset

Our SEM images consist of noisy images from eight distinct scenes, each with a resolution of 3072x2048 pixels. The term

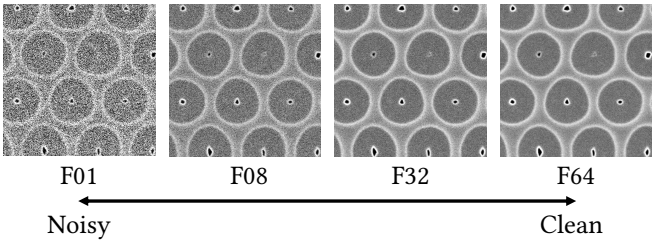


Fig. 3. Examples of our SEM images with different noise levels. The ground truth clean images (F64) are obtained by averaging 64 raw images (F01). As more frames are used for averaging (moving to the right in the figure), the image becomes less noisy.

"scene" refers to the different view of NAND flash semiconductor products, which are core components of Solid State Disk (SSD). These images are sourced from the manufacturing systems of Samsung Electronics, Inc. Within each scene, noisy images at four different noise levels (F01, F08, F32, and F64, as shown in Figure 3) are obtained by adjusting the number of frames averaged during acquisition. Namely, the number following "F" indicates the number of frames used. For example, F01 represents a single-frame (raw) image captured directly from the SEM, while F08 corresponds to images with reduced noise, achieved by averaging 8 frames. F32 images are primarily used to represent the upper bound performance of the median filter, as applying the median filter to F01 images often yields undesirable results. Despite the higher data collection cost, the median filter on F32 is preferred in practical manufacturing due to its stability and high performance. F64 images are used as ground-truth clean images in supervised learning and evaluation. Depending on the training scheme and the image to be denoised, the different F# image pairs are selected for training. For instance, in Noise2Noise (N2N) training with F01 images as input, pairs of F01 images from the same scene are used. In supervised learning with F01 images as input, a pair consisting of one F01 image and one F64 image is utilized for training.

The scenes can be categorized into two structural types: six scenes with normal structures and two *structurally different* structures. As shown in the bottom row of Figure 2, the *structurally different* F# images, abbreviated as SDF#, have oval or dented circles that differ from the precisely circular F01 images observed in the normally structured training image. Additionally, the small circle pattern within the larger circle is often barely visible, in contrast to the normal structure. A significant challenge with these *structurally different* images is that the N2N denoiser, trained exclusively on normal F01 images, struggles to process them effectively, as highlighted in the blue box of Figure 2. Among the eight scenes, five normally structured scenes are selected to form the training dataset, and three remaining scenes are selected as the test dataset to assess the performance and generalization ability of the denoising model: one from the normal structure and two from the *structurally different* scenes. For simplicity, these test scenes are referred to as 1st F01, 1st SDF01, and 2nd SDF01 in the experimental results. The key difference between the two SDF01 images is that the 2nd SDF01 image has a much smaller dented circle than the 1st SDF01. Since the model is trained only on normal structure images, successful denoising of the three evaluation scene images indicates that the model effectively handles both types of images, demonstrating good generalization ability.

Moreover, to resolve a spatial misalignment problem exists due to the inevitable motion shaking of the SEM and specimen during image acquisition, we devise a straightforward PSNR metric-based image registration algorithm, which is described in the Appendix (PSNR matching section).

Table 1. PSNR (dB)/SSIM results of ReNIn and application of regular dropout. PSNR and SSIM values are averaged over three test F01 images. "ID" refers to our input dropout strategy, while "CD" refers to the conventional dropout strategy, which drops neurons in hidden layers. D1 to D5 indicate the positions within the FBI-Net architecture where dropout is applied. The F01-F08 model (RN2N) is used as the base model. More details of the dropout layer position and experiment can be found in Dropout experiment details section. The PSNR and SSIM are average results based on the three test F01 images.

[t!] model	PSNR	SSIM
RN2N	18.8050	0.2400
RN2N + ID (ReNIn)	**20.8514**	**0.2712**
RN2N + CD (D1)	18.6695	0.2329
RN2N + CD (D2)	18.7859	0.2384
RN2N + CD (D3)	18.8553	0.2393
RN2N + CD (D4)	18.7790	0.2362
RN2N + CD (D5)	18.7836	0.2351

Relaxed Noise2Noise with Input dropout (ReNIn)

Now, we explain our approach, Relaxed Noise2Noise with Input dropout (ReNIn), specifically designed to enhance generalization while maintaining efficient training data collection costs for SEM image denoising.

[Relaxed Noise2Noise (RN2N)] As described in Preliminaries and Notations section, Noise2Noise requires training image pairs to have noise realizations that are independent and identically distributed, meaning they must have the exact same noise level. To enhance denoising performance, we propose to relax this constraint, allowing training with noisy images from the same noise distribution but with different noise levels. For example, when the input noisy image is F01, a relatively cleaner target image like F02 or other F# image can be used instead of using a target image with the same noise level ($\mathbf{y}'$ in Equation 2).

[Input dropout (ID)] To address overfitting, which can de-

Fig. 4. Overfitting issue of N2N. The green and red graphs represent the test and training loss, respectively. The intersection of these two graphs clearly indicates an overfitting issue of the N2N model.

grade denoising performance on *structurally different* images, we propose an input dropout (ID) strategy, which applies dropout directly to the input pixels, rather than to neurons in hidden layers as in conventional dropout strategies. Namely, the pixels of the image are dropped by multiplying a binary random mask $\boldsymbol{m}$, generated by i.i.d. Bernoulli distribution with the dropout rate p. With the masked image, the model is trained to perform dual tasks: removing the noise present in unmasked pixels while predicting the values of the masked pixels. As shown in Table 1, conventional dropout (rows 3 to 7) shows no improvement, while our input dropout significantly enhances generalization.

Overall, the training objective of ReNIn is:

$$\arg\min_{\theta} \sum_{i} L\left(f_\theta(\boldsymbol{y}^i \odot \boldsymbol{m}^i), (\boldsymbol{y}')^i\right) \tag{3}$$

in which $\boldsymbol{m}^i$ is the pixel-wise mask for $\boldsymbol{y}^i$, $\odot$ is the element-wise multiplication, and $(\boldsymbol{y}')^i$ is the (relatively clean) target image. In our experiments, we mainly use F01 images as input ($\boldsymbol{y}$) and F08 images as target ($\boldsymbol{y}'$).

After training, the denoiser takes test images without applying input dropout and returns its expected clean image. The estimated clean image $\hat{x}^t$ is denoted as :

$$\hat{x}^t = f_\theta(y^t) \tag{4}$$

where y^t is the test image. In our experiments, the F01 images are mainly used as test input images (y^t) since they lead to high efficiency without capturing multiple frames in the manufacturing systems.

Experimental Results

Experimental settings

Training images are randomly cropped to 256x256 pixels, resulting in a total of 21,600 cropped patches. Random horizontal and vertical flips are applied as data augmentation. For evaluation, the original size images (3072×2048) are used.

Baseline We have conducted experiments and comparisons with several denoising methods, including the Median filter, BM3D (Dabov et al., 2007), Noise2Void (N2V) (Krull et al., 2019), FBI-denoiser (Byun et al., 2021), Noise2Noise (N2N) (Lehtinen et al., 2018), and the supervised learning method. These methods are categorized into conventional denoising methods (Conventional) and deep learning-based methods.

For conventional denoising, we evaluate three methods: the median filter (F01), the median filter (F32), and BM3D. The median filter is a training-free approach widely used in real-world manufacturing for its stable, edge-preserving results, especially for the already *relatively clean* images like F32. The input image used for each median filter method is indicated in the parentheses. As expected, the median filter (F01) produces suboptimal results, while the median filter (F32) serves as an upper bound for denoising performance, assuming *expensive* F32 images are available during denoising. As another baseline, we use BM3D, recognized as the state-of-the-art among classical denoising methods. Since BM3D requires specifying the noise standard deviation (σ), we report results with $\sigma = 140$, which achieves the best PSNR.

We evaluate several deep learning-based methods: Noise2Void (Krull et al., 2019) and FBI-denoiser (Byun et al., 2021), Noise2Noise (Lehtinen et al., 2018), and a supervised learning method. To ensure fair comparisons among them, we used the same BSN architecture, FBI-Net, proposed in (Byun et al., 2021) for all methods. Noise2Void and FBI-Denoiser are the self-supervised learning baselines, which are trained on single noisy input images using BSN — the difference is that the FBI-Denoiser estimates affine mapping coefficients for the input images to generate the denoised output, while Noise2Void directly estimate the denoised output. On the other hand, Noise2Noise and supervised learning use target images that are separate from the input for training, and the only difference between them in our setting is the noise level of the target images; *i.e.,* Noise2Noise uses a separate noisy F01 image as a target and the supervised learning uses F64 as a "clean" target. Note both of these baselines do not necessarily have to use the BSN, but we fixed to use the BSN architecture so as to directly compare the performance with the self-supervised learning baselines. Moreover, as we show in our Ablation studies on modeling choices, the performance of Noise2Noise turns out to not differ significantly even when we use UNet (Ronneberger et al., 2015), which has a non-BSN architecture.

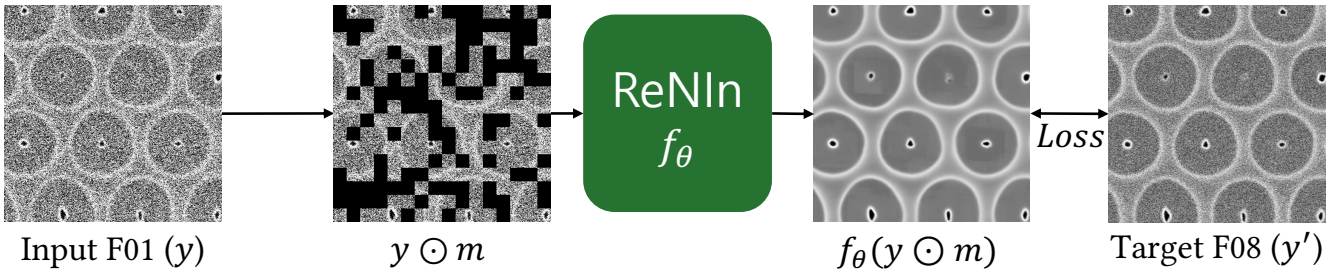


Fig. 5. Overall training procedure of ReNIn. Here, $\odot$ denotes element-wise multiplication. Note that the masking strategy is only applied in the training phase, not during the inference phase.

Evaluation metric

The evaluation is based on two standard image quality metrics and a failure rate metric in the circle detection task, which identifies the edges of circular patterns in each SEM image—a crucial pre-processing step in the manufacturing analysis and inspection.

[PSNR and SSIM] Image quality assessment is conducted using two standard metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), utilizing the F64 images that correspond to the input F01 images as clean reference images. As mentioned in SEM images dataset, three scenes are selected for sampling test images: a normally structured single F01 image and two structurally different F01 (SDF01) images.

[Failure rate in circle detection] We utilize the circle detection failure rate—defined as the proportion of undetected circles relative to the total number of circles—as a key evaluation metric. This metric is crucial for ensuring that denoised images retain clear structural information around edges, which is vital for distinguishing between cut and uncut parts in semiconductor manufacturing. Therefore, regardless of the above image quality, we believe a low circle detection failure rate clearly shows the effectiveness of the denoising process which has a crucial role in identifying manufacturing defects in semiconductor wafers.Therefore, regardless of the above image quality, we believe a low circle detection failure rate clearly shows the effectiveness of the denoising process which has a crucial role in identifying manufacturing defects in semiconductor wafers. It would contribute to subsequent product inspection and analysis, such as critical dimension or line edge roughness measurement. The failure rate (FR) for the test image y^t is represented as :

$$fr(y^t) = \frac{\boldsymbol{C}_{i,fail}}{\boldsymbol{C}_i} \times 100\ (\%) \tag{5}$$

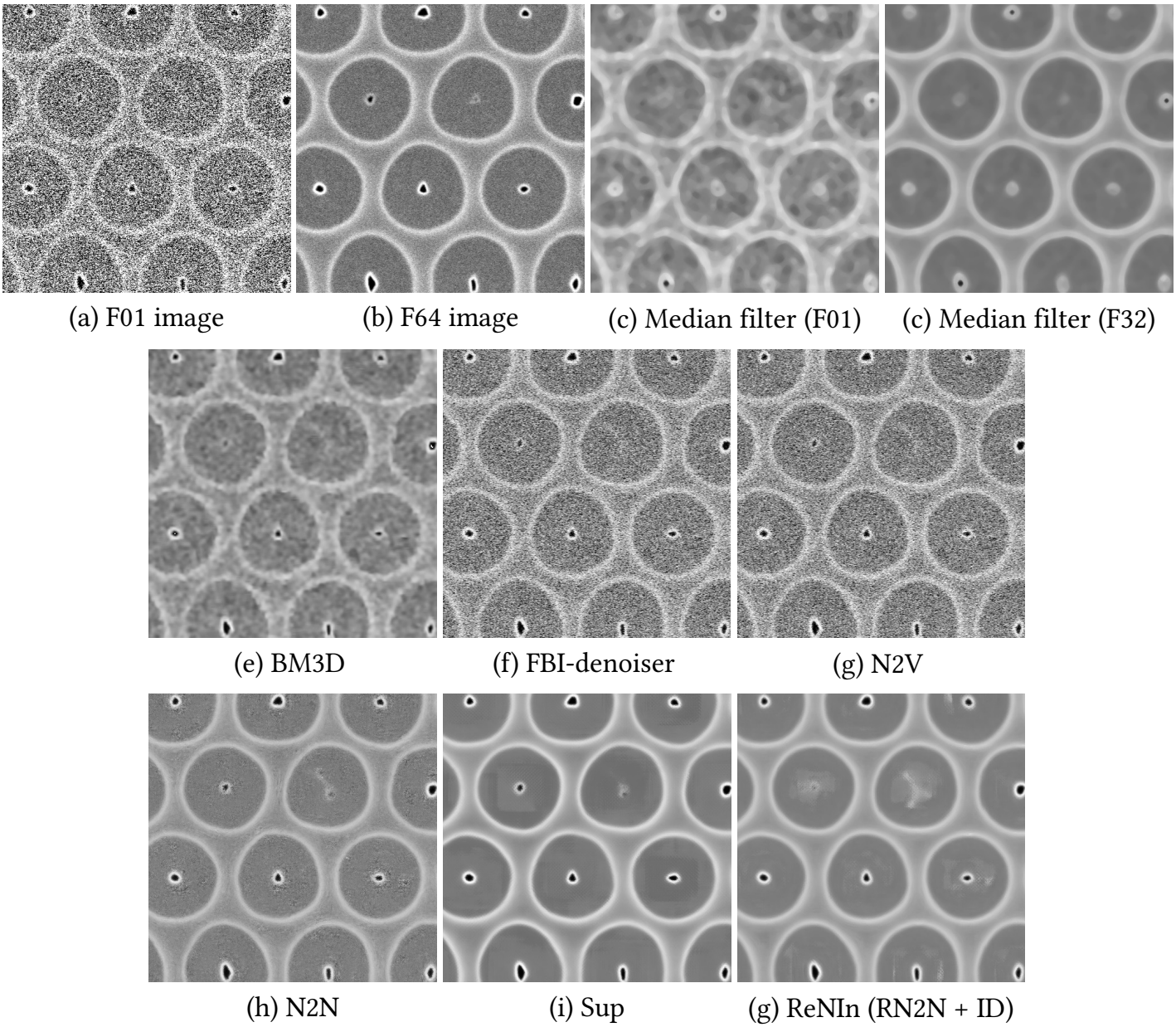
where $\boldsymbol{C}_i$ is the number of total circles to capture in i-th test images $\boldsymbol{y}^t$, and $\boldsymbol{C}_{i,fail}$ is the number of circles which fail to detect. In our experiments, three scenes are used for evaluation, and 108 circles ($\boldsymbol{C}_i$ = 108) are captured for evaluation.

Main results

As shown in Table 2, Figure 6, and Figure 7, ReNIn shows superior denoising results in terms of PSNR, SSIM, and visual quality. In Table 2, ReNIn achieves an averaged PSNR of 20.85 dB, showing a significant improvement in denoising performance, particularly with the *structurally different* images. It even surpasses the PSNR of supervised learning which is about 19.22 dB, mainly due to its enhanced performance on SDF01 images. Specifically, ReNIn almost reaches the performance of supervised learning for 1st F01 images within 0.61 dB difference in PSNR at 8× cheaper cost (F64 → F08) for collecting target trainig data. For SDF01 images, ReNIn achieves 20.39 dB and 19.12 dB, respectively. These values are significantly higher compared to the PSNR results from supervised learning, which are approximately 16.24 dB and 17.77 dB. We attribute this performance enhancement on SDF01 images to the superior generalization ability of ReNIn, supported by input dropout. We observe that the self-supervised methods (N2V and FBI-Denoiser), which assume pixel-wise independent noise, show inferior denoising performance compared to other baselines. We believe this is primarily due to the complexity of SEM noise, which includes pixel-wise correlation that violates the core assumption of these methods. This will be further discussed in the Noise correlation in SEM images section.

The effectiveness of ReNIn is further validated by qualitative results for both normal and *structurally different* images, as shown in Figure 6 and Figure 7. On normal structure images (Figure 6), our method (Figure 6j and deep learning-based methods except for self-supervised learning methods (Figure 6h and Figure 6i show good visual results while self-supervised learning methods and other conventional denoising approaches fall short. Especially, the denoised result of the median filter (F32) looks blurry compared to ReNIn although it shows superior performance in PSNR and SSIM. On *structurally different* images (Figure 7), other methods struggle with noise removal, as seen in Figure 7g, or fail to preserve edge information, as observed in Figure 7i. In contrast, our method effectively removes noise while preserving outer edge details, as depicted in Figure 7j, achieving results

:



(a) F01 image (b) F64 image (c) Median filter (F01) (c) Median filter (F32)

(e) BM3D (f) FBI-denoiser (g) N2V

(h) N2N (i) Sup (g) ReNIn (RN2N + ID)

Fig. 6. Qualitative denoising results on a structurally similar test image. Here, the input test image is structurally similar to the training image.

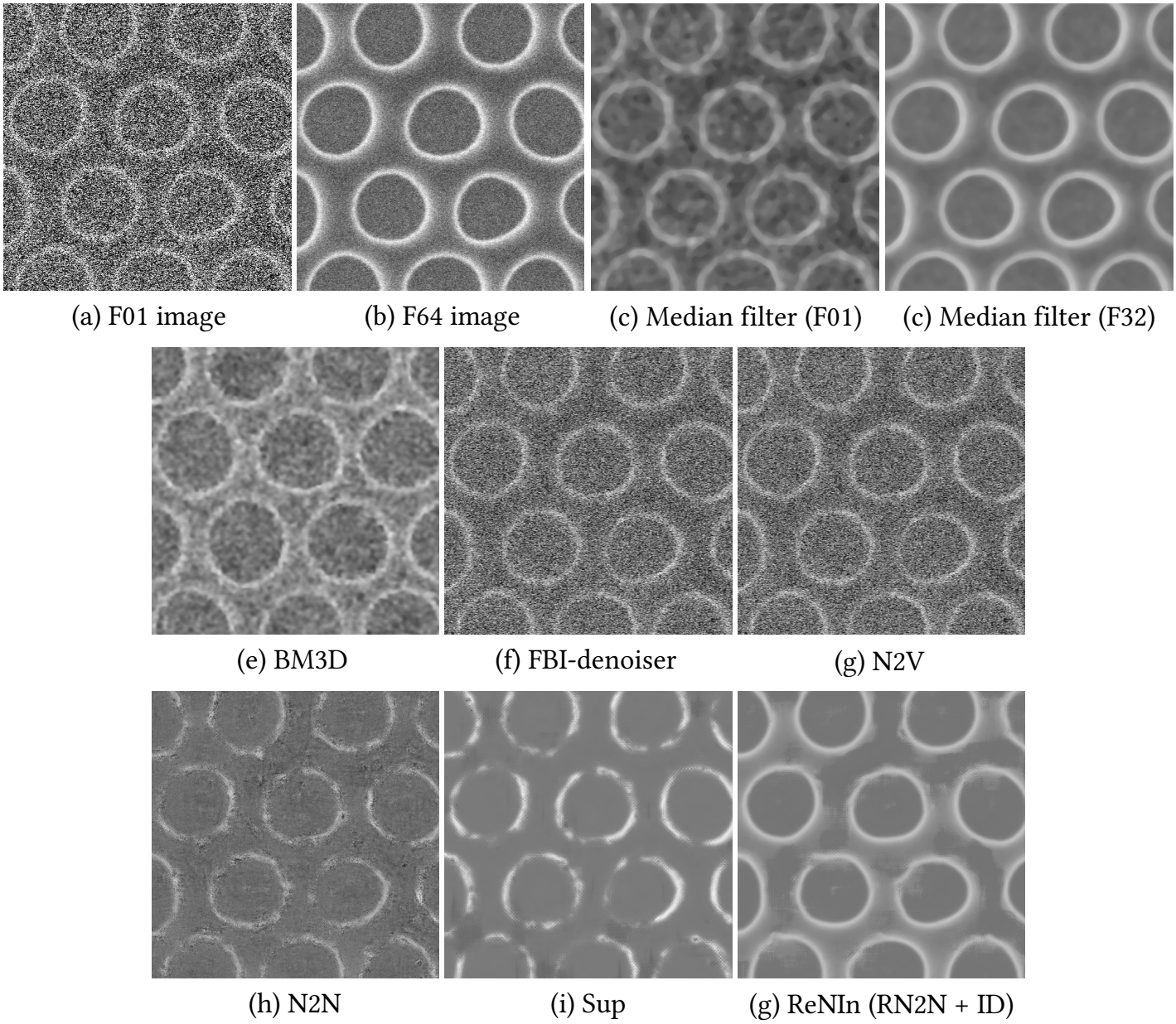


Fig. 7. Qualitative denoising results on a *structurally different* test image (SDF01). Here, the input test image is structurally different from the training image. The results of another SDF01 image are omitted for simplicity since it shows a similar tendency.

:

Table 2. PSNR(dB)/SSIM results of various methods. Since the evaluation images consist of one F01 and two SDF01 images, they are referred to as 1st F01, 1st SDF01, and 2nd SDF01 for convenience. The "average" column in PSNR/SSIM shows the averaged PSNR/SSIM values of three evaluation images. The best results for each image are marked in **bold** while the second ones are underlined except for the median filter (F32) since it utilized high-cost F32 images as input.

		PSNR / SSIM (↑)			
category	model	1st F01	1st SDF01	2nd SDF01	average
Conventional	Median filter (F32)	22.64 / 0.3382	21.70 / 0.2616	21.19 / 0.2409	21.84 / 0.2802
	Median filter (F01)	20.44 / 0.3143	14.09 / 0.2206	17.56 / 0.2130	17.36 / 0.2493
	BM3D (Dabov et al., 2007)	21.34 / 0.2974	15.12 / 0.2008	17.70 / 0.1905	18.05 / 0.2296
Deep learning-based	Noise2Void (Krull et al., 2019)	14.56 / 0.0792	12.20 / 0.0500	13.35 / 0.0540	13.37 / 0.0611
	FBI-denoiser (Byun et al., 2021)	14.71 / 0.0827	12.30 / 0.0505	13.42 / 0.0537	13.48 / 0.0623
	Noise2Noise (Lehtinen et al., 2018)	21.92 / 0.2990	14.40 / 0.1150	16.63 / 0.1403	17.65 / 0.1848
	Supervised learning	**23.66 / 0.3668**	16.24 / 0.1727	17.77 / 0.1830	19.22 / 0.2408
	ReNIn (Ours)	23.05 / 0.3446	**20.39 / 0.2508**	**19.12 / 0.2181**	**20.85 / 0.2712**

Table 3. Failure rate (FR) results of various methods. Details are the same as Table 2.

		FR(↓)			
category	model	1st F01	1st SDF01	2nd SDF01	average
Conventional	Median filter (F32)	0.00%	0.00%	0.00%	0.00%
	Median filter (F01)	12.50%	36.57%	52.31%	33.80%
	BM3D (Dabov et al., 2007)	8.33%	34.26%	67.59%	36.73%
Deep learning-based	Noise2Void (Krull et al., 2019)	0.93%	96.30%	87.04%	61.42%
	FBI-denoiser (Byun et al., 2021)	1.85%	96.30%	83.33%	60.49%
	Noise2Noise Lehtinen et al. (2018)	0.00%	100.00%	100.00%	66.67%
	Supervised learning	0.00%	100.00%	100.00%	66.67%
	ReNIn (Ours)	**0.00%**	**0.93%**	**2.31%**	**1.08%**

comparable to the median filter (F32), which uses expensive F32 images. We believe this significant improvement in handling challenging images underscores the strong generalization ability of ReNIn.

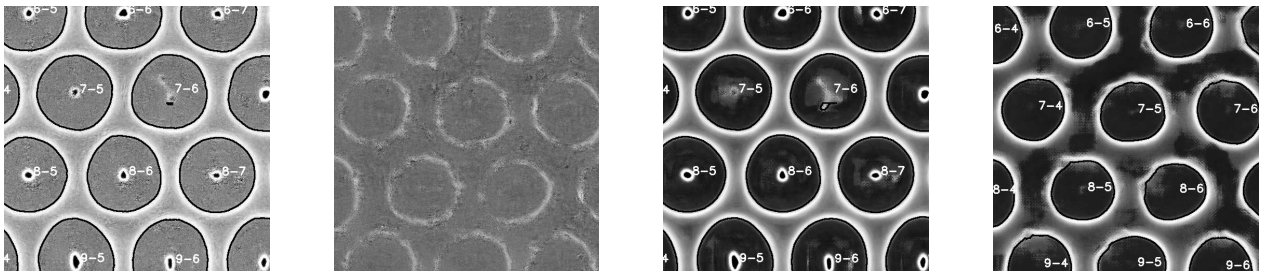
The effectiveness of the ReNIn is also evaluated using the failure rate (FR) of the circle detection metric, as presented in Table 3 and Figure 8. As shown in Table 3, ReNIn achieves superior circle detection performance compared to other baselines, demonstrating its strong generalization ability across various image structures. Specifically, ReNIn achieves a significantly lower average failure rate (FR) of 1.08%, whereas Noise2Noise struggles with effective circle detection, resulting in an average FR of 66.67%. In the visualization results presented in Figure 8, we again observe that Noise2Noise fails to perform well with *structurally different* images (Figure 8b), in contrast to its performance with normal images (Figure 8a). Whereas, our ReNIn performs consistently well in circle detection, regardless of whether the images are normal images (Figure 8c) or *structurally different* images (Figure 8d).

Ablation studies on modeling choices

Table 4 presents the effectiveness of the proposed components, Relaxed Noise2Noise (RN2N) and Input Dropout (ID), across different model architectures: FBI-Net (BSN) and

Table 4. Ablation results of ReNIn. Table shows each component of ReNIn: Relaxed Noise2Noise (RN2N) and Input dropout (ID). We evaluate the models in two different architectures: FBI-Net and UNet. FBI-Net and UNet are representative model architectures for BSN and non-BSN, respectively.

Model	+RN2N	+ID	PSNR	SSIM
FBI-Net	X	X	17.65	0.1848
	O	X	18.81	0.2400
	X	O	19.78	0.2545
	O	O	**20.85**	**0.2712**
UNet	X	X	18.09	0.2139
	O	X	18.76	0.2328
	X	O	19.80	0.2466
	O	O	**20.35**	**0.2636**



(a) F01 with N2N **(b)** SDF01 with N2N **(c)** F01 with ReNIn **(d)** SDF01 with ReNIn

Fig. 8. Qualitative results on the circle detection. The circle detection is performed over the image Figure 6 and Figure 7. Their results are indicated by the black circle line above the white edge (For better understanding, please check Figure 6 and Figure 7). Due to the circle line, the color of the images may appear shifted, but this is purely a visualization difference and does not affect the material properties. The numbers within the images indicate the circle numbering, ranging from 1-1 to 9-12. It is used to identify each NAND cell for further inspection.

Table 5. Target images with different noise levels. Here, all variants apply the input dropout (ID), with only the noise level of target images differing.

Model variants	PSNR	SSIM	FR
F01-F01 (N2N + ID)	19.78	0.2545	7.56%
F01-F02	19.96	0.2570	5.25%
F01-F08	20.85	0.2712	1.08%
F01-F16	20.87	0.2806	2.62%
F01-F64 (Sup + ID)	20.80	0.2939	1.54%

UNet (non-BSN). Following the Lehtinen et al. (2018), we use their implementation of UNet architecture. Regarding batch size in training, we use batch size of 1 for FBI-Net and 4 for UNet following their implementations. The base model (Noise2Noise, F01-F01) is shown in rows 1 and 5. Both RN2N (rows 2 and 6) and ID (rows 3 and 7) individually improve denoising performance, showing their effectiveness and compatibility with different architectures. Lastly, when both components are combined (rows 4 and 8, our final method, ReNIn), there is a further improvement in performance. Furthermore, we observe there is no significant performance difference between the BSN and non-BSN architecture[1]..

More analyses on ReNIn

In this subsection, we show more detailed analyses on ReNIn.

[Target images with different noise levels] Table 5 compares variants of ReNIn which utilize the different noise levels of target images, such as F02 and F16 (F08 images are used in our main experiment). All variants incorporate the input dropout (ID), with the only change being the target images. For simplicity, we use the notation F#1-F#2 to denote the training scheme, where F#1 represents input images and F#2 represents target images. Specifically, F01-F01 refers to "Noise2Noise with input dropout", while F01-F64 denotes the "supervised learning method with input dropout". We observe incremental performance improvements, particularly in SSIM, as target images become less noisy. It highlights the practical usability of ReNIn, which provides broader training options regarding both data collection cost and denoising performance.

[1]Note that the supervised learning method also has no improvement with non-BSN architecture.

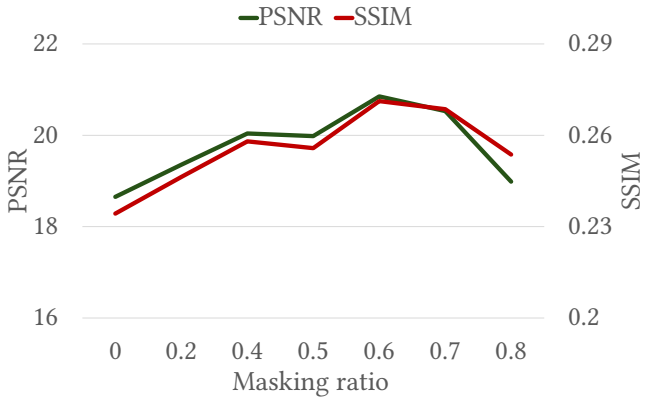


Fig. 9. Impact of the varying dropout ratio. F01-F08 model is used as a base model (*i.e.,* all models use the F08 image as the target). The dropout ratio indicates the proportion of the masked pixel among all the pixels. Note that the zero masking ratio indicates the F01-F08 model without input dropout.

[Impact of the varying dropout ratio] We assess the impact of different dropout ratios—the percentage of masked pixels in the input image—on denoising performance. Here, the F01-F08 model is used as a base model (*i.e.,* all models use the F08 image as the target). As shown in Figure 9, we test various ratios and identify the probability of 0.6 as the most effective, consistently yielding the highest performance in both PSNR and SSIM metrics. This ratio was therefore selected as the optimal setting for our final model.

[Impact of varying dropout patch size] We explore the

:

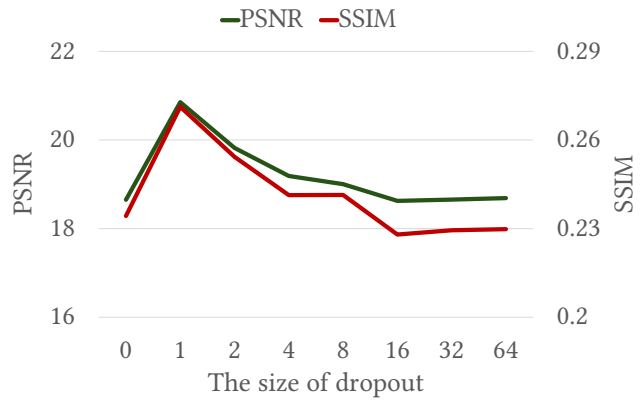


Fig. 10. Impact of the varying dropout patch size. This figure follows the same details as Figure 9, using the F01-F08 model as the base. A dropout patch size of 0 indicates training without input dropout, while a size of 1 corresponds to the pixel-wise dropout strategy.

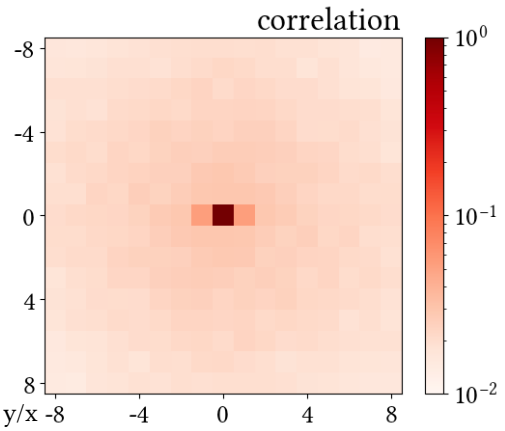


Fig. 11. The spatial correlation map between pixels based on relative positions. The x-axis denotes horizontal positions and the y-axis denotes vertical ones. Each coordinate represents the correlation between the pairs of pixels that differ by x and y. For example, the (1,2) of (x,y) of coordinate indicates the correlation between pairs of pixels that are 1 pixel apart horizontally and 2 pixels apart vertically. Note that the (0,0) coordinate indicates the correlation between the pixel and itself.

alternative strategies that drop patches, rather than individual pixels, within the image. To analyze this, we conduct experiments with various patch dropout sizes, ranging from 0 to 64. Here, a patch size of 2 refers to a 2×2 area of pixels being dropped. A dropout size of 0 indicates no dropout, while a size of 1 corresponds to the pixel-wise dropout used in the main experiments. As shown in Figure 10, we observe that pixel-wise dropout, where the dropout size is one pixel, consistently outperforms other dropout strategies.

Discussion

Noise correlation in SEM images

We observe that self-supervised learning methods (Noise2Void and FBI-denoiser) exhibit significantly inferior denoising performance in Table 2. We hypothesize that this is primarily due to the noise in our SEM images not being pixelwise independent. To verify this, following (Lee et al., 2022), we conduct noise correlation analysis. We calculate the Pearson correlation coefficient between the pairs of noise pixels with a difference of x and y (horizontal and vertical direction, respectively). Here, the noise was approximated by calculating the difference between F01 and F64 images. As shown in Figure 11, the correlation analysis reveals a relatively high correlation between adjacent neighboring pixels on the left and right, suggesting a spatial correlation. This highlights discrepancies between the noise assumptions in self-supervised learning methods, which assume noise is conditionally pixel-wise independent given the clean images, and the actual characteristics of our SEM image noise. We believe these discrepancies lead to the inferior denoising performance observed in self-supervised learning methods.

Conclusions

Our paper presents ReNIn which effectively addresses key challenges in SEM image denoising by offering an efficient balance between denoising performance and data collection costs through the Relaxed Noise2Noise (RN2N) framework. Additionally, the integration of input dropout strategy enhances generalization to unseen image structures without compromising performance on familiar ones, leading to a reduced failure rate in circle detection—a critical step in product analysis for semiconductor manufacturing. Overall, ReNIn offers an efficient and robust solution for SEM image denoising.

Acknowledgements

This work was supported by Samsung Electronics Co., Ltd [IO220623-09625-01]. It was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

Conflict of Interest

Competing interests: Author YongWoo Lee is employed at company Samsung.

References

Batson, J. and Royer, L. (2019). Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. Proceedings of Machine Learning Research.

Burger, H. C., Schuler, C. J., and Harmeling, S. (2012). Image denoising: Can plain neural networks compete with BM3D? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2392–2399. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Byun, J., Cha, S., and Moon, T. (2021). FBI-denoiser: Fast blind image denoiser for poisson-gaussian noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5768–5777. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Byun, J. and Moon, T. (2020). Learning blind pixelwise affine image denoiser with single noisy images. *IEEE Signal Processing Letters*, 27:1105–1109.

Chaudhary, N., Savari, S. A., and Yeddulapalli, S. S. (2019). Line roughness estimation and Poisson denoising in scanning electron microscope images using deep learning. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, pages 024001–024001.

Chen, H., Gu, J., Liu, Y., Magid, S. A., Dong, C., Wang, Q., Pfister, H., and Zhu, L. (2023). Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, pages 2080–2095.

Dey, B., Halder, S., Khalil, K., Lorusso, G., Severi, J., Leray, P., and Bayoumi, M. A. (2021). SEM image denoising with unsupervised machine learning for better defect inspection and metrology. In *Metrology, Inspection, and Process Control for Semiconductor Manufacturing XXXV*, volume 11611, pages 245–254. SPIE.

Ede, J. M. and Beanland, R. (2019). Improving electron micrograph signal-to-noise with an atrous convolutional encoder-decoder. *Ultramicroscopy*, pages 18–25.

Giannatou, E., Papavieros, G., Constantoudis, V., Papageorgiou, H., and Gogolides, E. (2019). Deep learning denoising of sem images towards noise-reduced ler measurements. *Microelectronic Engineering*, page 111051.

Gu, J., Ma, X., Kong, X., Qiao, Y., and Dong, C. (2024). Networks are slacking off: Understanding generalization problem in image deraining. *Advances in Neural Information Processing Systems*.

Huang, T., Li, S., Jia, X., Lu, H., and Liu, J. (2021). Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14781–14790. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Kong, X., Liu, X., Gu, J., Qiao, Y., and Dong, C. (2022). Reflash dropout in image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6002–6012. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Krull, A., Buchholz, T.-O., and Jug, F. (2019). Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2129–2137. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Laine, S., Karras, T., Lehtinen, J., and Aila, T. (2019). High-quality self-supervised deep image denoising. *Advances in Neural Information Processing Systems*.

Lee, W., Son, S., and Lee, K. M. (2022). Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Lefkimmiatis, S. (2018). Universal denoising networks: a novel CNN architecture for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3213. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. (2018). Noise2Noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pages 4620–4631. International Machine Learning Society.

Liu, Y., Zhao, H., Gu, J., Qiao, Y., and Dong, C. (2023). Evaluating the generalization ability of super-resolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ma, X., Wei, Z., Jin, Y., Ling, P., Liu, T., Wang, B., Dai, J., Chen, H., and Chen, E. (2024). Masked Pre-trained Model Enables Universal Zero-shot Denoiser. *arXiv preprint arXiv:2401.14966*.

Mao, X., Shen, C., and Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in Neural Information Processing Systems*.

Moran, N., Schmidt, D., Zhong, Y., and Coady, P. (2020). Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Nagano, K., Mukouyama, Y., Nishimura, T., Fujioka, H., Watanabe, K., Kurita, T., and Hidaka, A. (2021). Noise Reduction of SEM Images using U-net with SSIM Loss Function. *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*, pages 65–72.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison,

A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Quan, Y., Chen, M., Pang, T., and Ji, H. (2020). Self2self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1890–1898. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

Sato, Y., Kazui, M., and Kobayashi, S. (2022). Noise Reduction in SEM Images using Deep Learning. In *International Symposium on Semiconductor Manufacturing*, pages 1–4.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Tai, Y., Yang, J., Liu, X., and Xu, C. (2017). Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4539–4547. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Wu, X., Liu, M., Cao, Y., Ren, D., and Zuo, W. (2020). Unpaired learning of deep image denoising. In *European Conference on Computer Vision*.

Xu, J., Huang, Y., Cheng, M.-M., Liu, L., Zhu, F., Xu, Z., and Shao, L. (2020). Noisy-as-clean: Learning self-supervised denoising from corrupted image. *IEEE Transactions on Image Processing*, pages 9316–9329.

Yu, L., Zhou, W., Pu, L., and Fang, W. (2020). SEM image quality enhancement: an unsupervised deep learning approach. In Adan, O. and Robinson, J. C., editors, *Metrology, Inspection, and Process Control for Microlithography XXXIV*, volume 11325, page 1132527. International Society for Optics and Photonics, SPIE.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *Transactions on Image Processing*, pages 3142–3155.

Zhang, Y., Zhu, Y., Nichols, E., Wang, Q., Zhang, S., Smith, C., and Howard, S. (2019). A Poisson-Gaussian Denoising Dataset with Real Fluorescence Microscopy Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Institute of Electrical and Electronics Engineers/The Computer Vision Foundation.

Appendix

Dropout experiment details

To evaluate the effect of the dropout layer, we follow the same setting as our main experiment. The only difference is in the FBI-Net network architecture which adopts an additional dropout layer within the model. Similar to Kong et al. (2022), we choose five different locations to add the dropout layer. These are marked from D1 to D5, which are represented in Figure 12. We implement a dropout layer with PyTorch (Paszke et al., 2019). The details of FBI-Net architecture such as the masked convolutional layer and residual module are described in Byun et al. (2021).

PSNR matching

To address the problem of pixel mismatch between images depicting the same scene, we propose an intuitive and straightforward algorithm known as PSNR matching. This algorithm identifies pixel shift information by locating the best-matching patches using the PSNR metric. The algorithm consists of three steps:

1. Selecting appropriate patches with identical positions in the images.
2. Fixing the reference patch (typically the cleaner patch is chosen) and shifting the other patch horizontally and vertically.
3. Compute the PSNR for each shifted patch to identify the position with the highest PSNR, which indicates the best matching position.

In detail, we choose the reference patch from F64 images and randomly select the position to crop the patch, as the image displays similar patterns repeatedly. We opt for a crop size of 256, and the search range to shift vertically and horizontally is set to 40. The Figure 13 is illustrated a simple overview of this algorithm.

With this algorithm, we align the images to match appropriate pixel positions within the scene. However, there is a loss of information near the borders, where the search range multiplied by 2 represents the maximum number of lost pixels.

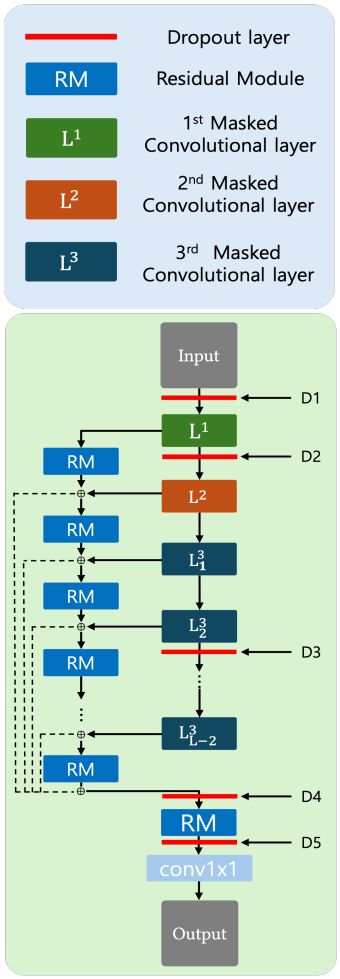


Fig. 12. The architecture of FBI-Net and dropout position from D1 to D6. The details of the masked convolutional layer and residual module can be found in Byun et al. (2021)

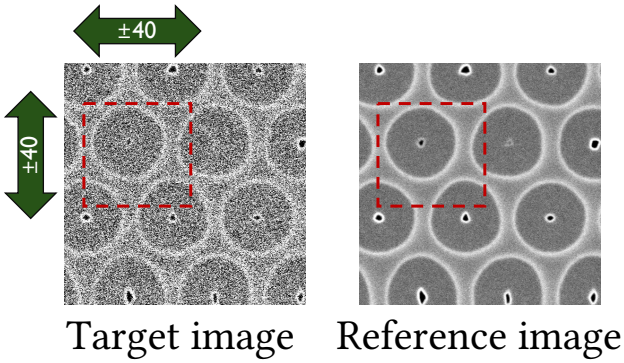


Fig. 13. A simplified overview of PSNR matching. For clarity, we demonstrate using image patches rather than the entire image.