Unsupervised Hybrid Transformer–Mamba Diffusion for Depth-Dependent Image Denoising in Scanning Acoustic Microscopy

Trong Nhan Nguyen, Vu Hoang Minh Doan, Tan Hung Vo, Dang Khoa Pham, Dinh Dat Vu, Thi Thu Ha Vu, Truong Tien Vo, Quoc Dung Nguyen, Jaeyeop Choi, and Junghwan Oh, *Member, IEEE*

***Abstract*—Scanning acoustic microscopy (SAM) plays a critical role in nondestructive inspection of semiconductor and industrial materials; however, deep-layer SAM images often suffer from severe attenuation-induced noise and the absence of clean reference data, significantly limiting reliable defect analysis. To address these challenges, this article proposes a hierarchical diffusion-based denoising framework that transfers structural and statistical knowledge from high-quality top-layer SAM images to restore degraded deep-layer images. A physics-guided synthetic data generation pipeline, combining Gaussian degradation modeling and unpaired domain transfer, is introduced to enable effective training. A hybrid Transformer–Mamba diffusion architecture is also constructed to jointly capture local spatial structures and long-range depth-dependent dependences, while a depthwise latent extraction strategy enhances the separation of noise and structural features. Extensive experiments on benchmark datasets and industrial SAM images demonstrate that the proposed approach achieves better performance on both reference-based metrics (peak signal-to-noise ratio, structural similarity index measure, learned perceptual image patch similarity, and Fréchet inception distance) and no-reference evaluation, while significantly improving downstream defect segmentation accuracy using the segment anything model. Although the proposed model has a higher parameter count than competing methods, the results indicate its strong potential for practical SAM-based industrial inspection.**

***Index Terms*—Hybrid transformer-Mamba, image denoising, scanning acoustic microscopy (SAM), unsupervised diffusion model.**

I. Introduction

SCANNING acoustic microscopy (SAM) is a specialized nondestructive testing (NDT) technique that utilizes high-frequency ultrasound to inspect internal structures of samples [1]. However, when inspecting thick, multilayered specimens, image quality is strongly affected by the attenuation of ultrasound as it penetrates the material [2]. The deeper the acoustic wave travels, the weaker the returning echoes become. In contrast, the background noise of the SAM system is largely independent of the ultrasound travel length [3]. For certain thick samples, the echoed A-scan signals from deep layers are often overwhelmed by background noise, resulting in severe degradation of the reconstructed C-scan images. In such cases, image quality may deteriorate to the point where structural features or defects cannot be distinguished from noise, severely limiting the reliability of SAM for nondestructive evaluation.

Conventional approaches to addressing this issue have mostly relied on adjusting hardware or scanning strategies. One common approach is to reduce the transducer frequency to enhance penetration capability, as lower frequency ultrasound travels further [4]. Although this increases the amplitude of the echo signals, it reduces spatial resolution, thereby degrading image quality. Another method is to increase the transducer signal amplitude by amplifying the excitation pulse; however, this is constrained by the electrical load capacity of the transducer [5]. Finally, reversing the sample to shorten the propagation distance can improve detection sensitivity, yet this strategy is often impractical in industrial environments where speed and accessibility are critical [6]. These limitations underscore the need for alternative solutions that can improve SAM image quality without requiring complex, costly, and often ineffective hardware modifications.

A deep learning (DL)-based denoising approach offers a promising solution to the depth-dependent noise problem in SAM. The key challenge, however, lies in the absence of

Received 7 November 2025; revised 13 December 2025; accepted 22 December 2025. Date of publication 15 January 2026; date of current version 6 April 2026. This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) under Grant 2022R1A5A8023404. Paper no. TII-25-7780. *(Trong Nhan Nguyen, Vu Hoang Minh Doan, and Tan Hung Vo contributed equally to this work.) (Corresponding author: Junghwan Oh.)*

Trong Nhan Nguyen, Thi Thu Ha Vu, and Quoc Dung Nguyen are with the Industry 4.0 Convergence Bionics Engineering, Department of Biomedical Engineering, Pukyong National University, Busan 48513, Republic of Korea (e-mail: 202355504@pukyong.ac.kr; havuthithu96@pukyong.ac.kr; quocdungxc@pukyong.ac.kr).

Vu Hoang Minh Doan, Tan Hung Vo, Dinh Dat Vu, and Jaeyeop Choi are with the Smart Gym-Based Translational Research Center for Active Senior's Healthcare, Pukyong National University, Busan 48513, Republic of Korea (e-mail: doanvuhminh@pukyong.ac.kr; Votanhung@pukyong.ac.kr; dinhdatvn96@pukyong.ac.kr; jaeyeopchoi@pknu.ac.kr).

Dang Khoa Pham is with Industrial and Systems Engineering Department, North Carolina A&T State University, Greensboro, NC 27411-0002 USA (e-mail: kdpham@aggies.ncat.edu).

Truong Tien Vo is with the Department of Engineering and Technology, Pham Van Dong University, Quang Ngai 570000, Vietnam (e-mail: vttien@pdu.edu.vn).

Junghwan Oh is with the Ohlabs Corp and Industry 4.0 Convergence Bionics Engineering and Department of Biomedical Engineering, Pukyong National University, Busan 48513, Republic of Korea (e-mail: jungoh@pknu.ac.kr).

Digital Object Identifier 10.1109/TII.2025.3648838

1941-0050 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

high-quality deep-layer scanning images that could serve as ground truth for supervised training. This limitation naturally frames the problem as an unsupervised learning task, requiring models that can operate without explicit clean references. Nevertheless, a key observation is that while image degradation increases with penetration depth, the surface layers generally produce high-quality images containing the same types of features and defects as those found in deeper layers. This creates an opportunity to train a model for learning the structural characteristics of top-layer images and transfer this knowledge to denoise degraded images from deeper layers. Such a model should capture not only spatial structures but also temporal transitions across depth, since attenuation and noise accumulation follow depth-dependent patterns.

Previous attempts to denoise SAM images—where high-frequency acoustic imaging suffers from low signal levels and high noise—have included both classical and DL-based approaches. A comparative study applied traditional filters (Gaussian filter, median filter, Wiener filter, and total variation) alongside a block-matching 4-D (BM4D) filter to time-domain SAM signals, showing that BM4D better suppresses noise while preserving structural details [7]. On the DL frontier, one study utilized a convolutional neural network (CNN) to achieve superresolution in SAM, enhancing 180-MHz images by approximating higher resolution 316-MHz ground-truth images with strong fidelity despite a minimal training set [8]. However, most recent DL-based denoising studies for SAM have focused on coin samples or surface-layer images, which contain little noise, highlighting their preliminary and impractical nature. In practice, effective deep-layer SAM image denoising remains largely unexplored.

Recent advances in DL have produced a new generation of powerful image denoising architectures, including vision transformer (ViT)-based and hybrid CNN–Transformer networks, which have achieved advanced performance in natural and medical imaging tasks. For instance, Zhang et al. [9] proposed Hformer, a highly efficient ViT–CNN hybrid for low-dose Computed tomography (CT) denoising, which effectively captures global dependences while preserving fine structural details. Similarly, Su et al. [10] introduced a dual-path Transformer–CNN architecture for CT denoising [11] that outperformed conventional U-shaped Network (U-Net) baselines on the Mayo Clinic Low-Dose CT Challenge benchmark. Zhao et al. [12] further conducted a comprehensive evaluation of various deep models (U-Net, Wasserstein generative adversarial network (Wasserstein GAN), and residual networks) and training strategies for low-dose CT images [13], [14], highlighting how network design choices impact noise suppression and structure recovery [15], [16] .

In the ultrasound domain, Zhang and Zhang [17] proposed a GAN-based despeckling method built on a U-Net residual backbone (GAN-RW) that effectively suppresses speckle while preserving details in B-mode ultrasound, but it is trained on typical ultrasound datasets and assumes stationary noise statistics. Cammarasana et al. [18] introduced a deep model optimized for real-time ultrasound denoising, demonstrating practical inference speeds, yet this work focuses on common clinical imaging conditions rather than the extreme signal-to-noise ratio (SNR) drop seen in deep SAM layers. Recent content-aware and attention-driven networks have improved despeckling by learning image priors and attention maps, but they still rely on supervised or weakly supervised training and do not explicitly model depth-dependent acoustic attenuation [19]. These existing denoising approaches—ranging from classical filtering techniques to modern CNN- and Transformer-based models—are limited in their applicability to deep-layer SAM imaging. Classical methods fail to handle depth-dependent, nonstationary noise, while most DL approaches assume access to clean ground-truth images and are validated primarily on shallow or low-noise datasets. None of the existing methods explicitly leverages the hierarchical relationship between high-quality surface-layer images and severely degraded deep-layer images, nor do they model long-range depth-dependent acoustic attenuation. These limitations highlight a critical gap in current SAM restoration research.

To address this gap, our study introduces a diffusion model-based denoising approach for SAM using a novel hybrid architecture named scanning acoustic microscopy image denoising (SAMID), which integrates Transformer blocks and Mamba state-space modules within a multiscale encoder–decoder framework. We first constructed a synthetic SAM dataset through a three-stage pipeline combining Gaussian degradation modeling and CycleGAN-based domain transfer (DT), overcoming the data scarcity problem. We then introduced a depthwise latent extraction model (DLEM) to disentangle noise-specific and structure-specific features, enhancing noise suppression without sacrificing structural integrity. Finally, we embed these modules within a probabilistic diffusion framework to progressively recover clean representations from noisy inputs. The contributions of this article can be summarized as follows:

1) A hierarchical denoising framework that learns structural and statistical features from high-quality top-layer SAM images and transfers this knowledge to restore noisy deep-layer images.
2) A physics-guided synthetic data generation pipeline that combines Gaussian degradation modeling with CycleGAN-based DT to accurately emulate depth-dependent SAM noise, enabling effective training without requiring paired clean–noisy datasets.
3) A DLEM that disentangles structure-specific and noise-specific features, improving fidelity in fine-detail preservation while enabling more selective suppression of heterogeneous deep-layer noise.
4) A hybrid Transformer–Mamba diffusion architecture that jointly captures local spatial structures and long-range depth-related dependences, providing a robust generative framework for reconstructing degraded SAM images.
5) Extensive evaluations on both benchmark datasets and real industrial SAM scans, demonstrating state-of-the-art (SOTA) denoising performance across supervised and unsupervised metrics, as well as significant improvements in downstream defect-segmentation accuracy.

The rest of this article is organized as follows. Section II presents the methodology of the proposed SAMID model. Section III describes the experimental setup and implementation details. Section IV discusses the results and provides an in-depth analysis. Section V reports the ablation study to further validate the effectiveness of our method. Finally, Section VI concludes this article.

II. Methodology

A. Characteristics Analysis of Noise

In SAM imaging, noise behavior is governed by both acoustic propagation physics and hardware-induced disturbances. As an

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

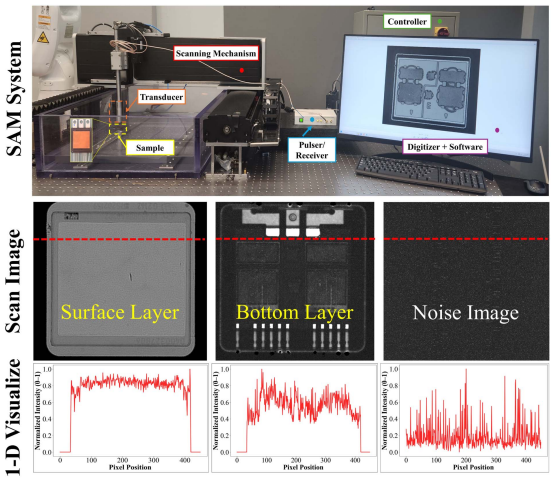


Fig. 1. SAM system and representative examples of noise characteristics. From left to right: scan images and 1-D normalized amplitude plots for the surface layer, bottom layer, and noise-only case. The results show stronger fluctuations and distortions at deeper layers and random background noise in the absence of a sample.

ultrasonic wave travels through a solid medium, its amplitude decays exponentially with depth according to

$$A(d) = A_0 e^{-2\alpha d} \tag{1}$$

where $A(d)$ is the received amplitude at depth d, A_0 is the initial amplitude at the surface, and α is the material-dependent acoustic attenuation coefficient [20]. Impedance mismatches at material interfaces where acoustic impedance is defined as $Z = \rho c$, with ρ denoting the material density and c the longitudinal sound speed generate reflection and scattering losses characterized by

$$R = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{2}$$

where Z_1 and Z_2 are the impedances of two adjacent media [21]. Beyond these physical mechanisms, the SAM system inherently introduces random electronic noise, primarily originating from the pulser and receiver circuitry. The pulser's SNR specification imposes a baseline noise floor that cannot be eliminated, and this electronic noise becomes increasingly dominant as the echo amplitude weakens with depth. These combined effects lead to the depth-dependent degradation observed in Fig. 1.

Fig. 1 illustrates the SAM system setup (top), representative scan images (middle), and their corresponding 1-D normalized amplitude plots (bottom). At the surface layer, the scan image appears relatively uniform, and the 1-D plot confirms that the amplitude fluctuates only within a narrow range (0.75–0.96). In contrast, the bottom layer scan shows stronger artifacts, and the 1-D profile reveals larger fluctuations (0.4–0.9), indicating that deeper regions suffer from more severe noise and signal distortion. The right column ("Noise Image") corresponds to a scan without any physical sample. Here, the scan image mainly contains random background patterns, while the 1-D normalized shows irregular oscillations with extended lobes and large amplitude variations. Unlike Gaussian or Poisson distributions, this noise is highly random and does not follow classical statistical assumptions. Overall, the figure highlights that noise in SAM imaging is depth-dependent and increasingly severe at deeper layers. These findings motivate the need for robust denoising strategies. In this work, we propose a DL-based method to suppress noise and recover reliable structural information from SAM images.

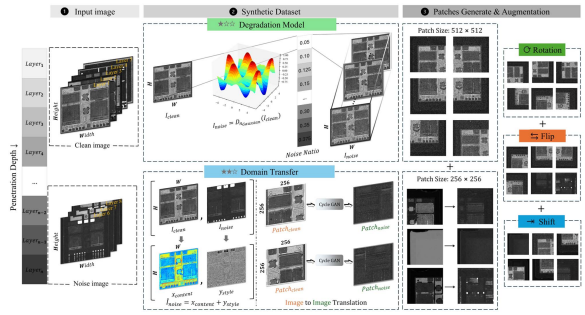


Fig. 2. Three-stage synthetic data generation pipeline for SAM denoising: 1) real acoustic images are captured from a commercial SAM system; 2) synthetic noisy images are created using a Gaussian DM and a CycleGAN-based DT to mimic realistic depth-dependent industrial noise; and 3) clean and noisy images are segmented into fixed-size patches and augmented for training. This pipeline alleviates data scarcity and improves robustness for industrial SAM denoising.

B. Proposed Synthetic Data Generation and Augmentation Framework

To overcome the shortage of labeled training data and improve denoising performance in SAM, we propose a three-stage framework for synthetic image generation and augmentation (see Fig. 2). The process includes: real data capture, synthetic data generation, and patches generator and augmentation. This approach leverages synthetic data to address data scarcity while enhancing robustness.

1) Real Data Capture: In this stage, real acoustic images were collected using the SAM system across multiple penetration depths. The dataset includes two primary types of images: clean images, typically captured from the uppermost layers near the surface, and noisy images obtained from deeper layers, where acoustic signal attenuation and scattering introduce significant noise artifacts. These depth-dependent variations provide a natural distribution of both high- and low-quality data.

Clean images, which preserve the structural content with minimal noise interference, are selected as the base inputs for synthetic data generation. In contrast, noisy images from the bottom layers are characterized by severe signal degradation and are utilized as visual style references in the DT stage. This separation of content (clean) and style (noise) forms the foundation for simulating realistic degradation scenarios in subsequent stages.

2) Synthetic Data Generation:

a) Degradation model (DM): To simulate noisy data, we adopted a classical Gaussian DM, where synthetic noisy images I_{noise} are generated from clean images I_{clean} using the relation

$$I_{\text{noise}} = D_{\text{gaussian}}(I_{\text{clean}}, \sigma) \tag{3}$$

where σ denotes the noise ratio, and multiple noise levels are applied to create a diverse synthetic dataset. Specifically, each clean image is degraded using 15 predefined noise ratios: [0.05, 0.1, 0.125, 0.135, 0.15, 0.165, 0.175, 0.185, 0.195, 0.2, 0.225, 0.25, 0.3, 0.35, 0.375]. This process yields 15 noisy variants per clean image, ensuring a wide range of noise intensities for model robustness.

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

b) Domain transfer: To simulate realistic and structurally consistent noisy images, we adopted a DT approach based on CycleGAN. Unlike classical noise injection, this method enables data-driven learning of degradation patterns from real-world acoustic images without requiring paired clean-noisy examples. Let $x_{\text{clean}} \in \mathcal{D}_{\text{clean}}$ denote a clean image sampled from the near-surface region, and $x_{\text{noisy}} \in \mathcal{D}_{\text{noisy}}$ denote a noisy image from deeper layers. The goal of the generator $G : \mathcal{D}_{\text{clean}} \to \mathcal{D}_{\text{noisy}}$ is to transfer the style (noise characteristics) of x_{noisy} to x_{clean}, while preserving the content of the clean image

$$x_{\text{translated}} = G(x_{\text{clean}}) \approx x_{\text{content}} + x_{\text{style}} \tag{4}$$

where x_{content} represents the preserved microstructural detail from x_{clean}, and x_{style} captures the noise patterns learned from $\mathcal{D}_{\text{noisy}}$.

Our contribution lies in leveraging this DT paradigm to construct a large-scale synthetic noisy dataset tailored for SAM, where labeled data are inherently limited due to imaging cost, acquisition time and restrictions related to data confidentiality. By emulating the stochastic yet structured degradations found in real SAM images, the synthetic samples not only augment the training space but also preserve semantic and material-specific textures. This enables subsequent denoising models to generalize better across varying defect types, depths, and material properties-a critical advancement for industrial defect inspection where clean-noisy image pairs are impractical to obtain.

3) Patches Generator and Augmentation: To prepare data for training, both clean and noisy images are divided into fixed-size patches. Patches of 512×512 are extracted from degradation-based images, while 256×256 patches are used for DT samples to accommodate architectural input constraints. To further enhance data diversity and improve model generalization, standard augmentation techniques, including random rotation, horizontal/vertical flipping, and pixel shifting, are applied to all patches.

C. Depthwise Latent Extraction Model

In this study, the depthwise latent extraction model (DLEM) is introduced as a proposed component for enhancing image denosing ability. DLEM leverages depthwise feature separation to disentangle and isolate noise-specific and structure-specific representations across different hierarchical levels of the network. Unlike traditional convolutional blocks that aggregate spatial and channel information uniformly, DLEM emphasizes selective information propagation across depth, enabling more precise retention of structural content and more aggressive suppression of noise artifacts.

1) Mathematical Formulation: For each feature map $\mathbf{F}^{(l)} \in \mathbb{R}^{H_l \times W_l \times C_l}$ at depth l, the *DLEM* is formulated as

$$\mathbf{F}^{(l)}_{\text{dlem}} = \sigma\left(\sum_{c=1}^{C_l} \mathbf{W}^{(l)}_c * \mathbf{F}^{(l)}_c\right) \tag{5}$$

where

1) $\mathbf{W}^{(l)}_c \in \mathbb{R}^{k \times k}$ is a *depthwise convolution kernel* applied only to channel c;
2) $*$ denotes convolution;
3) $\sigma(\cdot)$ is a nonlinear activation (GELU);
4) C_l is the number of channels at depth l.

After the depthwise convolution stage, the intermediate latent features $\mathbf{F}^{(l)}_{\text{dlem}}$ are fused using a *pointwise projection*

$$\mathbf{F}^{(l)}_{\text{out}} = \phi\left(\mathbf{P}^{(l)} * \mathbf{F}^{(l)}_{\text{dlem}}\right) \tag{6}$$

where $\mathbf{P}^{(l)} \in \mathbb{R}^{1 \times 1 \times C_l}$ represents a 1×1 convolution kernel for channel mixing and $\phi(\cdot)$ is an optional normalization or gating function.

2) Computational Efficiency: The DLEM decomposes a standard convolution into a depthwise convolution followed by a pointwise projection, significantly reducing the computational cost. The complexity is reduced from

$$\mathcal{O}\left(k^2 \cdot C^2\right) \quad \text{to} \quad \mathcal{O}\left(k^2 \cdot C + C^2\right) \tag{7}$$

where k is the kernel size and C is the number of channels. For small k ($k = 3$), this leads to approximately k^2 times faster execution with minimal performance degradation, making DLEM highly suitable for high-resolution image denoising. In SAM images, noise and structural components exhibit markedly different spatial distributions across depth. Therefore, separating channelwise responses becomes crucial for isolating noise-dominant frequencies at deeper layers. The proposed DLEM exploits this property by performing depthwise convolutions before channel fusion, enabling more selective suppression of depth-induced noise while maintaining the fine acoustic reflections that correspond to material interfaces or inner defects.

D. Diffusion Model

As shown in Fig. 3(b), the diffusion model is a probabilistic framework that estimates the data distribution $q(z)$ by inverting a fixed Markov chain of length T. It comprises a forward process, where Gaussian noise is gradually added to clean samples, and a reverse process, where the noise is iteratively removed to reconstruct the original image. Unlike existing diffusion-based denoisers that assume spatially uniform noise patterns, the proposed hybrid generator explicitly models the depth-dependent and spatially heterogeneous noise characteristics inherent in SAM imaging. The Transformer blocks provide fine-grained structural encoding, while the Mamba-based state-space units efficiently capture the long-range attenuation patterns produced by acoustic propagation.

1) MultiScale Encoder–Decoder Diffusion Model: We propose a four-level encoder–decoder diffusion architecture that denoises latent variables $\mathbf{z}_t \in \mathbb{R}^{B \times C \times H \times W}$ by progressively recovering clean representations across multiple resolutions.

The encoder begins with a shallow convolutional head (Level 0, $\mathbb{R}^{B \times 32 \times H \times W}$) and applies progressive downsampling through Levels 1–3 (64, 128, 256 channels). Each stage integrates MammerBlock units, combining Mamba-based state-space modeling and Transformer self-attention, while latent priors $\mathbf{z}_{\text{prior}}$ are injected via learned 1x1 convolutions to embed diffusion dynamics.

The decoder mirrors this structure with upsampling, convolution, and skip connections to reconstruct features from Level 3 back to Level 0. At each stage, MammerBlock modules refine features, and a final 1x1 convolution projects outputs into the latent space.

This hierarchical design balances fine-grained detail with large-scale structure, making it effective for challenging denoising tasks in high-dimensional acoustic and visual domains.

2) Mammer Generator Model: The proposed Mammer generator introduces a novel hybrid architecture that integrates

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

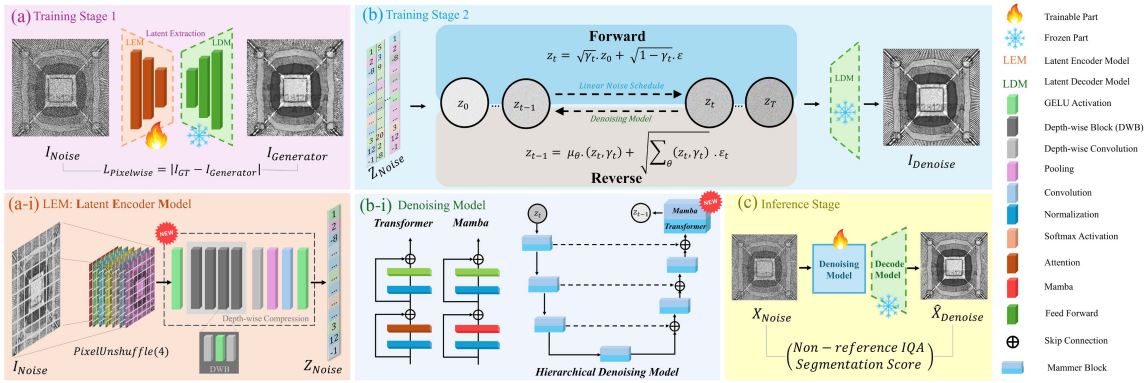


Fig. 3. Framework for denoising tasks using the proposed Mammer Block. (a) Training Stage 1 involves latent extraction with the latent encoder model (LEM) and generation of noisy latent codes using a generator. (b) Training Stage 2 employs a forward process to add noise and a reverse process with the hierarchical denoising model, combining transformer and Mamba architectures. (c) Inference stage utilizes the trained denoise model to produce denoised outputs, evaluated with a nonreference IQA segmentation score.

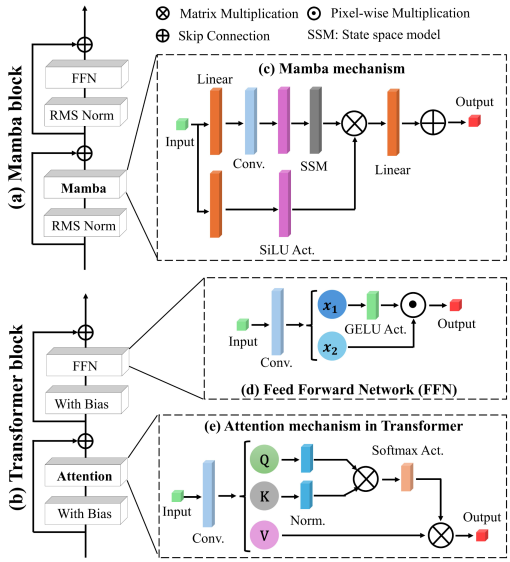


Fig. 4. Overview of the proposed Mammer block structure. The hybrid architecture integrating (a) Mamba mechanism for efficient long-range modeling via selective SSMs and (b) transformer attention for spatial correlations. Sequential coupling enables both local precision and global context modeling.

both transformer-based attention [22] and sequence modeling via Mamba mechanisms [23], [24], arranged in a multiscale encoder–decoder framework, as illustrated in the Fig. 4. Each processing block in the encoder and decoder consists of two sequential components: a Transformer block and a Mamba block, enabling the model to effectively capture both spatial correlations and long-range dependences across diffusion steps.

a) Transformer block: Within the Mammer generator, each encoder–decoder level begins with a Transformer block [25], designed to capture spatial correlations across the input feature map $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$. The block first applies bias-based layer normalization, followed by multihead self-attention. This attention mechanism operates through the standard scaled dot-product formulation

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{8}$$

where $\mathbf{Q} = \mathbf{x}W_Q$, $\mathbf{K} = \mathbf{x}W_K$, and $\mathbf{V} = \mathbf{x}W_V$ are the query, key, and value projections with learned matrices $W_Q, W_K, W_V \in \mathbb{R}^{C\times d_k}$. The result is projected via W_O and combined with the input via residual connection

$$\mathbf{y}_{\text{attn}} = \mathbf{x} + (\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})\cdot W_O). \tag{9}$$

This output is passed into a feed-forward network consisting of a gated linear unit (GLU) and Gaussian error linear unit (GELU) activation, where branched processing with elementwise multiplication allows for nonlinear interaction

$$\text{GLU}(\mathbf{a},\mathbf{b}) = \mathbf{a}\odot\text{GELU}(\mathbf{b}). \tag{10}$$

This block excels in modeling spatial dependences but is limited in sequential representation and long-range temporal modeling, particularly relevant in the reverse denoising process of diffusion models.

b) Mamba block: Inspired by the recent vMamba architecture [26], our design retains the original formulation of the Mamba block to preserve its core strengths in modeling long-range sequential dependences. To complement the Transformer, the Mamba block is introduced to capture long-range temporal dependences and model sequential transitions in latent variables $\mathbf{z}_t \to \mathbf{z}_{t-1}$ during reverse diffusion. Each Mamba block begins with root mean square layer normalization (RMSNorm), defined as

$$\text{RMSNorm}(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}}\cdot\gamma \tag{11}$$

where γ is a learnable parameter. The block then applies a selective state-space model (SSM), which can be represented as

$$\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \mathbf{B}\mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}\mathbf{h}_t + \mathbf{D}\mathbf{x}_t \tag{12}$$

where matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and $\mathbf{D}$ are learned from input via lightweight convolutional projections. The Mamba architecture leverages parallelizable scan operations and modern hardware acceleration to achieve $\mathcal{O}(L)$ runtime for sequence length L, significantly more efficient than the $\mathcal{O}(L^2)$ complexity of attention-based methods.

Within the Mammer block, the Transformer first extracts spatial content features, while the Mamba block sequentially processes the same features to model their evolution across time or diffusion steps. This hybrid design ensures that both local defect boundaries and global depth-related degradations are faithfully reconstructed, which is essential for reliable SAM-based

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

defect inspection. Moreover, the hybrid Mamba–Transformer configuration enables the network to simultaneously handle fine-scale structural cues and depth-dependent signal degradation. This complementary behavior aligns closely with the physical formation of SAM images, allowing the model to reconstruct deep-layer information that conventional denoisers cannot reliably recover.

III. Experimental

A. Dataset

Our experiments used both public benchmarks and a private industrial dataset. For pretraining, we adopted the smartphone image denoising dataset (SIDD) [27], selecting 2 240 images of resolution 512×512. Evaluation is conducted on standard benchmarks, including CBSD68 [28], Kodak24 [29], McMaster [30], and Urban100 [31].

For domain-specific adaptation, we constructed the SAM dataset with 221 acoustic images from 15 power semiconductor samples acquired by SAM system. It consists of 209 clean surface images and 12 noisy deep-layer images affected by severe signal attenuation. To ensure rigorous evaluation and prevent data leakage, we performed dataset splitting at the sample level rather than the image level. Specifically, the 15 physical samples were first divided into independent training and testing groups before any image acquisition. This strategy eliminates overlap between training and testing domains, mitigates overfitting, and provides a more faithful measure of real-world generalization, where the model must operate on unseen components without prior exposure. Subsequently, the resulting raw images were further expanded using a synthetic pipeline with two pathways: a degradation-based method (DM) and a DT module, as previously discussed in Section II-B2. The final dataset is generated as a weighted combination

$$\text{Dataset} = \alpha \cdot \text{DM} + \beta \cdot \text{DT} \tag{13}$$

with optimal results obtained at $\alpha : \beta = 7 : 3$. This process yields 2 881 synthetic images at original resolution, forming the complete training and validation set.

B. Evaluation Metric

To comprehensively evaluate the performance of our denoising model across both supervised and unsupervised datasets, we adopted a combination of full-reference and no-reference image quality metrics, as well as task-specific assessments, ensuring the robustness and practicality of our approach.

For supervised datasets, we employed widely used full-reference metrics. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are calculated on the Y channel of the YCbCr color space to reflect luminance-based distortions. In addition, the perceptual quality is assessed using the learned perceptual image patch similarity (LPIPS), with deep features extracted by a pretrained AlexNet. To further evaluate the statistical similarity between the distribution of generated and reference images, the Fréchet inception distance (FID) is computed using InceptionNet features.

For unsupervised datasets, where ground-truth images are unavailable, a diverse set of evaluation criteria is employed to ensure the reliability of denoising performance. First, we used a nonreference image quality metric (NIQA) [32], evaluated based on the pseudocode proposed in recent literature on autodenoising models. This approach enables quantitative assessment of perceptual quality without relying on clean ground truth. Second, we analyzed the statistical distribution of noise components, following methodologies introduced in recent unsupervised learning approaches for real-world denoising. By comparing noise histograms and variance distributions before and after denoising, we validate the model's capability to attenuate complex noise patterns while preserving structural information. Finally, to assess the applicability of the denoised images for downstream tasks, we integrated the segment anything model (SA-Model) [33] to perform semantic segmentation. The effectiveness of the denoising process was evaluated using two segmentation metrics: Dice similarity coefficient (DSC) and intersection over union (IoU).

In addition, we employed 1-D intensity profile analysis to directly compare amplitude transitions across noisy, ground-truth (supervised), and denoised images. This visualization-based metric clearly reveals the model's ability to suppress noise while preserving edge sharpness and defect-relevant signal variations that may not be fully captured by global image-level metrics.

This combination of metrics provides a rigorous and comprehensive evaluation of the proposed denoising framework across various data regimes and application scenarios.

C. Training Setting

1) Training Strategy: Our model was trained in two stages: latent extraction and compression, and the diffusion process. To improve generalization and prevent overfitting or underfitting on limited industrial data, we employed a transfer learning approach. This transfer learning strategy enables the model to benefit from generic noise patterns learned from SIDD while adapting effectively to the specific characteristics of acoustic microscopy images. Specifically, both stages are first trained on the SIDD dataset—widely recognized for real-world image denoising—serving as a large-scale pretraining source. In Stage 1, we trained the latent encoder and generator components for 2000 iterations. The resulting weights are then reused as pretrained parameters for Stage 2, where the diffusion model is trained for 10 000 iterations.

2) Implementation Details: The model is trained using an augmented dataset generated through rotation, flipping, and shifting. We adopted the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and applied a weight decay of 1×10^{-4}. The initial learning rate is set to 2×10^{-4}.

For the diffusion process, we employed a linear noise schedule, which demonstrates superior performance compared to the cosine variant. The total number of denoising time steps is set to 10. In the Transformer backbone, refinement modules are introduced to stabilize feature representations at each level, with two refinement blocks placed after decoding. The network is structured with a total of four hierarchical levels, where the number of Transformer blocks is configured as $[3, 5, 6, 6]$, and the number of attention heads as $[1, 2, 4, 8]$, respectively. This design helps maintain gradient flow and avoids degradation in deeper layers.

For the Mamba modules, we follow the original implementation with a state size of 8 to model long-range dependences efficiently. The entire system is implemented using BasicSR (version 1.4.2), PyTorch (version 2.2.0), and Zetascale (version 2.8.6), running on CUDA version 11.7. All training experiments are conducted on a distributed GPU environment, including NVIDIA H200 SXM (on RunPod platform), RTX 3060 Ti, and RTX 3090 Ti.

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

TABLE I
QUANTITATIVE COMPARISON ON GAUSSIAN COLOR IMAGE DENOISING WITH SOTA METHODS

Method	CBSD68			Kodak24			McMaster			Urban100		
	$\sigma = 0.15$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 0.15$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 0.15$	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 0.15$	$\sigma = 0.25$	$\sigma = 0.5$
IRCNet	33.86	31.16	27.86	34.69	32.18	28.93	34.58	32.18	28.91	33.78	31.20	27.70
ADNet	31.74	29.25	26.29	-	-	-	-	-	-	-	-	-
BM3D	33.52	30.71	27.38	34.28	32.15	28.95	34.06	31.66	28.51	33.93	31.36	27.93
DNCNN	33.89	31.23	27.92	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
SwinIR	34.42	31.78	28.56	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82
Restormer	34.40	31.79	28.60	35.47	33.04	30.01	35.61	33.34	30.30	35.13	32.96	30.02
Xformer	34.43	31.82	28.63	35.39	32.99	29.94	35.68	33.44	30.38	35.29	33.21	30.36
MambaIR	34.48	32.24	28.66	35.42	32.99	29.92	35.70	33.43	30.35	35.37	33.21	30.30
SAMID	34.96	32.18	29.03	35.64	33.18	30.21	36.10	33.71	31.02	35.55	33.05	30.52

The highest value in each Column is highlighted in red.

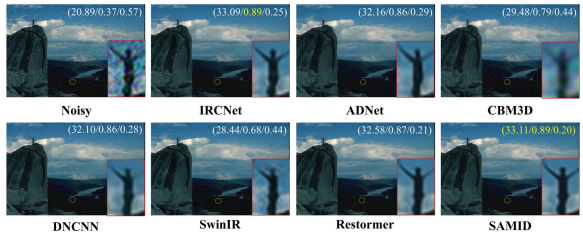


Fig. 5. Visual comparison of Gaussian color image denoising on the *CBSD68* dataset with noise level $\sigma = 0.25$. Our proposed *SAMID* model preserves structural details and suppresses noise more effectively than other SOTA methods (PSNR↑ /SSIM↑ /LPIPS↓).

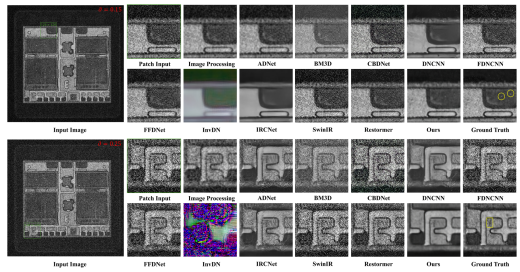


Fig. 6. Visual comparison of denoising results on synthetic SAM data from power semiconductor layers at noise ratios $\sigma = 0.15$ and $\sigma = 0.25$. Each row shows a cropped patch from the noisy input image, followed by results from image processing techniques and SOTA DL models, including FFDNet, BM3D, CBDNet, DnCNN, FDN-CNN, InvDN, IRCNet, SwinIR, and Restormer. Our model clearly preserves structural details and suppresses noise more effectively compared to other methods. Ground-truth (clean) images are included for reference.

D. Comparison Methods

To demonstrate the effectiveness of our proposed approach, we compared it against both classical and SOTA denoising methods. As a traditional baseline, we applied a blur kernel commonly used in conventional image processing. For SOTA benchmarks, we evaluated across multiple categories: Color block-matching and 3D filtering (CBM3D) [34] (classical method), CNN-based models (fast and flexible denoising network (FFDNet) [35], fast denoising convolutional neural network (FDnCNN) [36], denoising convolutional neural network (DnCNN) [28], convolutional blind denoising network (CBDNet) [37], attention-based denoising network (ADNet) [38], and image restoration convolutional neural network (IRC) [39]), downsampling-based architectures (invertible denoising network (InvDN) [40]), transformer-based methods (SwinIR [41], Restormer [42], and Xformer [43]), and a representative SSM, MambaIR [44]. Among these, Restormer and Xformer serves as a hierarchical transformer baseline, offering a strong reference point for assessing the advantage of our hybrid Transformer–Mamba generator within a diffusion framework. For a fair evaluation, each baseline model was retrained or fine-tuned on our dataset under the same training schedule and data conditions. Hyperparameters were adopted the default settings provided by their authors.

IV. RESULT AND DISCUSSION

A. Comparison With SOTA Models

1) Supervised Metric: a) Benchmark dataset: The results of Gaussian color image denoising are presented in Table I, where evaluations are conducted under different noise levels $\sigma = 15, 25$, and 50. Compared to SOTA baselines, our proposed SAMID model consistently achieves the highest PSNR values across on most benchmark datasets, including CBSD68, Kodak24, McMaster, and Urban100. Our proposed SAMID framework consistently achieving the highest PSNR values across the majority of benchmark datasets, including leading results on Kodak24 and McMaster. Notably, SAMID surpasses recent architectures such as Xformer and MambaIR by demonstrating superior noise suppression and feature preservation, solidifying its effectiveness across varying noise levels ($\sigma = 0.15$ to $\sigma = 0.5$). In addition to quantitative benchmark evaluations, we present a representative image from the CBSD68 dataset to qualitatively validate our results. As shown in Fig. 5, the qualitative results demonstrate that SAMID exhibits superior capability in both fine-detail preservation and robust noise suppression compared with recent SOTA denoisers, including Restormer. In particular, the integration of the DLEM enables SAMID to effectively capture and restore critical local features that are often blurred or missed by models relying primarily on global representations. This enhanced fidelity is essential for preserving structural integrity and ensuring reliable inspection in practical industrial quality-control scenarios.

b) Synthetic SAM dataset: We selected two random images from the testing synthetic dataset to compare with image processing methods and some SOTA models. In Fig. 6, results on 128×128 patch images with two noise levels ($\sigma = 0.15$ and $\sigma = 0.25$) demonstrate that our model provides the best visual quality compared to competing methods. In the black region of the top patch image, DnCNN, IRCNet, and our model exhibit strong denoising performance; however, DNCNN and IRCNet tend to oversmooth, inadvertently losing many characteristics of the original patch image. Regarding feature preservation, image processing methods, ADNet, Restormer, and our model are effective approaches, but their denoising ability diminishes when noise levels are excessively high. Our model's robust denoising capability stems from the application of its multiscale hierarchical architecture combined with the global attention mechanism in the Transformer (yellow region). Furthermore, the effective handling of dense, randomly distributed noise while preserving the characteristic color fidelity of SAM images is achieved by leveraging the long-range dependence capability of the SSM, trained robustly via the proposed synthetic framework. Specifically, defects highlighted in the yellow regions are well-preserved after denoising by SAMID. This highlights the significance and efficacy of our initial DLEM, which ensures feature preservation in the shallow extraction stage. Quantitatively, Table II shows that our model outperforms all competing methods, surpassing the second-best approach by 5.3397 dB on the PSNR metric, 0.2341 on the SSIM metric, 0.1004, and 47.4317 on the latent space metrics: LPIPS and FID. Although this comparison is conducted solely on the synthetic dataset, it

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

TABLE II
QUANTITATIVE COMPARISON ON SAM DATASET ACROSS DIFFERENT NOISE RATIOS

Noise Ratio	Metric	Blur	ADNet	BM3D	CBDNet	DNCNN	FDNCNN	FFDNet	IRCNet	InvDN	Restormer	SwinIR	SAMID
0.05	FID↓	17.9903	71.1666	64.1040	21.0526	16.5246	87.7225	78.9056	132.9892	125.1128	44.1925	15.1400	**9.6352**
	LPIPS↓	0.3535	0.4578	0.4013	0.3306	0.3591	0.4555	0.4462	0.5951	0.5039	0.3379	0.2769	**0.2685**
	PSNR↑	30.5917	30.6538	23.1295	28.1848	32.9212	29.8988	29.7839	26.5376	28.6354	27.6027	31.5246	**33.0948**
	SSIM↑	0.7856	0.7517	0.5675	0.6623	0.8172	0.7357	0.7333	0.6291	0.6631	0.5953	0.7851	**0.8366**
0.15	FID↓	119.2814	135.8377	117.4810	128.5532	178.2706	141.0337	138.4606	221.8437	137.5959	135.6992	91.0029	**43.6888**
	LPIPS↓	0.4698	0.5045	0.4990	0.5189	0.4597	0.5371	0.5436	0.7893	0.5343	0.5190	0.4548	**0.3949**
	PSNR↑	25.4203	26.7190	17.0320	18.8545	26.1564	20.4330	20.7106	19.5295	25.9970	18.3588	21.6775	**29.0984**
	SSIM↑	0.5274	0.6130	0.2956	0.2096	0.5655	0.2581	0.2656	0.4173	0.5753	0.1941	0.3182	**0.7454**
...													
0.35	FID↓	186.6758	195.7014	198.5421	205.8318	183.3142	192.9327	194.0800	178.0028	502.7681	176.0443	198.2191	**122.3746**
	LPIPS↓	0.5723	0.6187	0.6329	0.6379	0.5744	0.6402	0.6439	0.6018	0.8008	0.6246	0.6474	**0.4530**
	PSNR↑	18.3599	13.5963	11.6783	12.3458	17.1015	12.4535	12.7311	16.4568	6.6478	13.0430	11.7417	**27.4856**
	SSIM↑	0.2560	0.0859	0.0864	0.0680	0.1914	0.0669	0.0700	0.1645	0.0317	0.0749	0.0618	**0.6687**
0.375	FID↓	186.3069	200.9655	199.5411	209.4516	186.6894	199.6212	199.9339	183.3551	485.6160	185.9180	202.9458	**132.4229**
	LPIPS↓	0.5805	0.6286	0.6418	0.6454	0.5890	0.6479	0.6505	0.6153	0.8001	0.6319	0.6552	**0.4581**
	PSNR↑	17.7532	12.9695	11.3411	11.8669	16.1645	11.9301	12.1950	15.3624	6.6587	12.6417	11.2440	**27.2040**
	SSIM↑	0.2376	0.0768	0.0787	0.0620	0.1545	0.0608	0.0644	0.1197	0.0317	0.0689	0.0562	**0.6606**
Average	FID↓	175.6768	151.1524	132.2298	138.3259	130.3630	149.1277	147.4366	302.2636	138.7127	142.5497	109.2889	**61.8572**
	LPIPS↓	0.4386	0.5186	0.6308	0.7346	0.4484	0.6681	0.6591	0.8776	0.5224	0.7136	0.5839	**0.3382**
	PSNR↑	23.7078	23.5204	15.7145	17.5759	24.0801	18.7050	18.9496	15.8982	23.7100	17.1169	19.6894	**29.4198**
	SSIM↑	0.4653	0.4836	0.2463	0.1982	0.4817	0.2355	0.2392	0.3118	0.4958	0.1806	0.2735	**0.7299**

The best and second-best results are highlighted in bold and underlined, respectively.

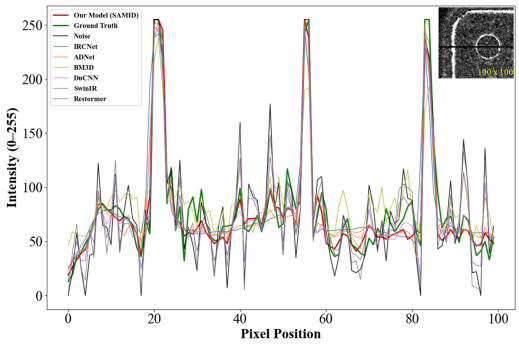


Fig. 7. Intensity-profile analysis on a 100 × 100 cropped region. The proposed SAMID model (red) closely aligns with the ground truth (green), accurately restoring major signal peaks while effectively suppressing noise. Traditional methods retain residual fluctuations, whereas transformer-based models exhibit oscillatory artifacts. SAMID achieves the best balance between noise reduction and structural fidelity across the entire profile.

still reflects the learning capability and denoising performance across different noise levels, suggesting strong generalization to real-world denoising tasks.

In addition, a random 100× 100 region was extracted from each selected sample to perform a 1-D intensity-profile analysis, providing a more detailed quantitative comparison of restoration quality. As illustrated in the line profile comparisons (see Fig. 7), our model demonstrates superior noise suppression while accurately preserving critical signal peaks at pixel positions 20, 60, and 85, achieving near-perfect alignment with ground truth. In contrast, transformer-based methods (SwinIR and Restormer) exhibit pronounced oscillations and fail to preserve the signal shape consistently in the region between pixels 30–55, particularly after abrupt intensity transitions (such as the white area). This behavior indicates a reduced ability to capture fine-grained local textures and high-frequency structures, partly due to their limited inductive bias. Meanwhile, traditional methods such as BM3D still retain noticeable residual noise fluctuations. Our model achieves an optimal balance between aggressive noise reduction and structural detail preservation across the entire signal profile, outperforming competing approaches in maintaining signal fidelity.

2) Unsupervised Metric: a) Qualitative analysis: To assess the effectiveness of the proposed approach in an industrial context, we conducted experiments on a real dataset of bottom-layer images acquired from the SAM system. As illustrated in Fig. 8, the evaluation was performed on a real testing image of size 900 × 900 pixels, which was cropped into nonoverlapping patches of 64 × 64 pixels for accuracy. The visualization demonstrates that our model achieves effective denoising while retaining intricate structural patterns.

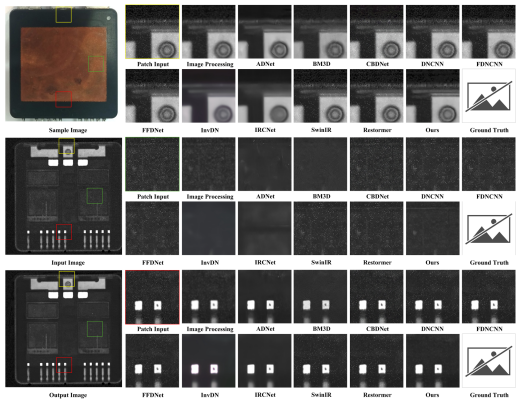


Fig. 8. Visual comparison of the bottom-layer image after denoising using different methods. The first column shows the sample image and selected regions of interest. Subsequent columns present results from various denoising algorithms, with our method (Ours) demonstrating clearer structural preservation. Some regions still contain visible defects, which are highlighted for qualitative assessment. Ground-truth images are shown for reference.

For comparison, the image processing baseline (*blur*) reduces noise but not completely, while simultaneously blurring the image and removing fine details. CNN-based methods such as CBDNet, DNCNN, FDNCNN, and FFDNet tend to preserve the original structures but struggle to sufficiently suppress noise, whereas ADNet, BM3D, InvDN, and IRCNet often oversmooth the input patches, distorting fine features. Transformer-based methods, including SwinIR and Restormer, maintain certain structural elements but fail to deliver convincing denoising performance.

A closer inspection of highlighted regions provides further insight.

1) *Yellow box:* Our method effectively denoises both the background and foreground while preserving circular features. In contrast, some methods (ADNet, InvDN, and IRCNet) generate skewed or overly blurred circles.
2) *Green box:* This region corresponds to a defect (void) on the chip. Our method produces a clear representation with strong contrast between the background and the defect, which is crucial for downstream inspection tasks.
3) *Red box:* The terminal component maintains characteristics consistent with the noisy input, indicating that our approach can suppress noise without altering essential structural properties.

These observations confirm that our method delivers high-quality denoising while retaining critical features, making it suitable for defect detection and other industrial downstream applications.

b) Quantitative analysis: Table III presents the quantitative evaluation on the unsupervised dataset using nonreference image quality assessment metrics. Across all 12 test images, our method consistently achieves the highest score in most cases, with an overall average of **80.720**, surpassing the second-best method (Restormer) by a considerable margin.

While certain methods such as SwinIR and Restormer demonstrate competitive results for specific images, they exhibit noticeable performance drops on others, indicating lower robustness to varying noise patterns. CNN-based methods, including CBDNet, DNCNN, and FFDNet, generally produce lower scores,

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

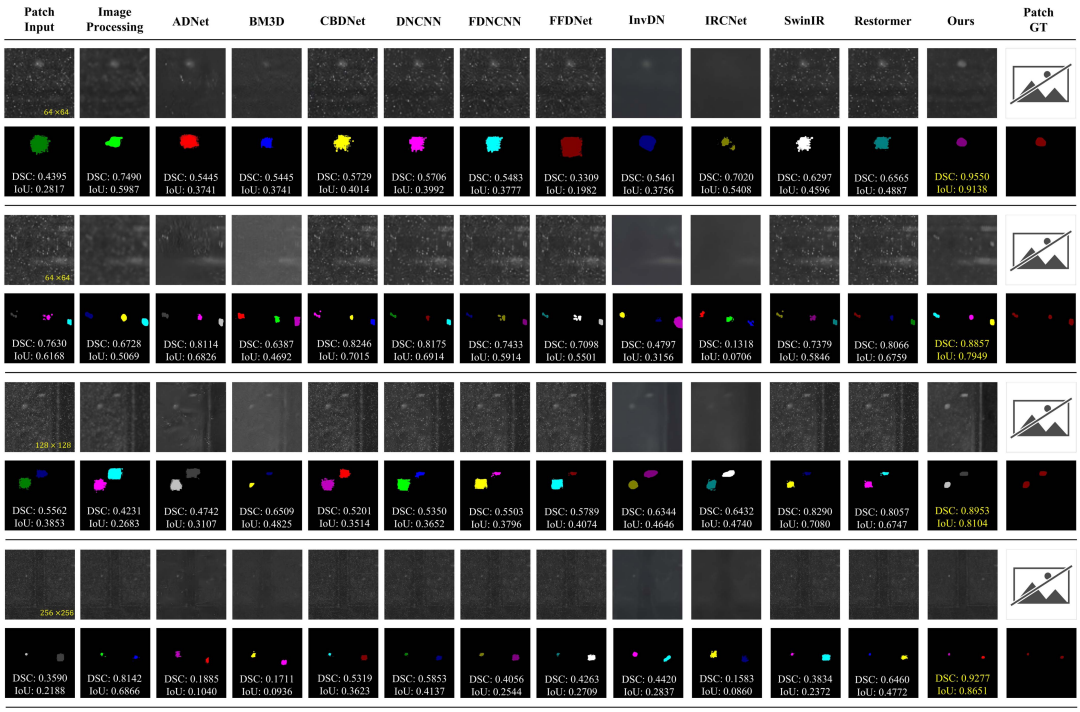


Fig. 9. Performance comparison of segmentation models on patch defects.

TABLE III
QUANTITATIVE EVALUATION ON THE UNSUPERVISED DATASET USING NIQA METRICS

Image	Blur	ADNet	BM3D	CBDNet	DNCNN	FDNCNN	FFDNet	IRCNet	InvDN	Restormer	SwinIR	Xformer	MambaIR	SAMID
Image 1	5.649	0.968	52.660	66.100	8.555	0.207	0.649	7.625	25.101	9.123	9.069	16.292	20.592	**80.319**
Image 2	6.081	2.238	56.366	4.953	6.389	4.482	2.408	6.332	20.129	11.076	3.843	11.147	20.289	**80.658**
Image 3	3.463	1.819	46.054	20.453	9.438	4.204	1.374	2.291	23.430	8.377	10.516	16.578	14.607	**79.757**
Image 4	7.672	3.166	36.918	4.362	6.494	3.174	1.413	7.370	20.418	9.816	3.780	10.471	24.252	**79.713**
Image 5	1.102	2.418	46.007	35.746	11.278	1.031	1.683	2.210	22.413	10.676	19.615	26.049	19.011	**79.679**
Image 6	1.102	4.120	41.247	68.485	7.415	0.430	0.737	8.163	21.883	16.621	13.441	25.232	18.147	**83.430**
Image 7	1.202	2.726	25.721	76.681	7.857	1.564	0.500	7.654	25.399	13.005	12.593	25.309	20.551	**83.841**
Image 8	1.849	3.064	43.022	69.965	8.430	0.780	0.386	7.579	24.658	12.792	11.508	21.620	21.429	**81.818**
Image 9	2.579	1.841	36.493	62.061	7.774	0.826	0.523	8.179	19.322	11.764	10.866	20.791	21.662	**82.397**
Image 10	0.311	3.768	52.913	11.095	7.873	1.871	0.436	6.095	19.068	5.976	7.109	16.905	22.714	**81.633**
Image 11	18.218	5.088	51.564	47.022	6.691	2.736	0.657	6.087	21.849	11.258	0.966	4.634	8.619	**72.458**
Image 12	3.575	0.935	54.670	64.256	8.307	0.289	1.508	8.215	22.663	11.553	10.500	20.091	21.330	**82.935**
Average	4.400	2.679	45.303	44.265	8.042	1.800	1.023	6.483	22.194	11.003	9.484	17.108	19.818	**80.720**

Each image corresponds to a bottom-layer image affected by strong acoustic noise. The scores Represent the average value calculated over all 128 × 128 Cropped patches from each image.

TABLE IV
PSNR RESULTS ACROSS DIFFUSION STEPS AND DM/DT RATIOS

Diffusion Step	6	7	8	9	10	11	12
PSNR (dB)	31.595	31.584	31.610	31.530	**31.612**	31.602	31.528
DM/DT Ratio	Only DM	Only DT	9&1	8&2	7&3	6&4	5&5
PSNR (dB)	31.417	27.499	31.610	31.635	**31.661**	31.536	31.525

reflecting their limited generalization ability in handling strong acoustic noise. In contrast, our approach maintains high performance across all samples, it outperforms the nearest competitor by a large margin.

This superior performance can be attributed to our generator architecture, which integrates both *Mamba* and *Transformer* blocks. The Mamba component efficiently captures long-range dependences with high computational efficiency, while the Transformer enhances global feature representation [45]. This hybrid design enables our model to effectively suppress strong acoustic noise while ensuring robust and consistent results across diverse bottom-layer SAM images.

These findings confirm that our method not only delivers superior performance but also demonstrates strong consistency and robustness, making it well-suited for industrial inspection tasks where noise conditions are highly variable and sensitive.

c) Defect segmentation analysis: Voids are among the most prevalent defects in power semiconductor packages, and accurately quantifying their area is critical for industrial inspection, as it directly affects defect percentage calculations for quality control. In this study, we employed the SA-Model to segment defect regions automatically. The segmented regions were evaluated against ground-truth annotations using DSC and IoU metrics. As shown in Fig. 9, our denoising model produces outputs in which defect boundaries are preserved with high clarity, enabling the segmentation algorithm to delineate void regions more accurately. Compared to SOTA denoising models and traditional image processing methods, our approach achieves the highest DSC and IoU scores across all tested patch sizes. This demonstrates that the integration of our Mamba–Transformer-based denoising generator not only improves visual quality but also substantially enhances the performance of downstream inspection tasks. Such capability is essential for industrial scenarios where precise defect localization and measurement are critical to maintaining high production quality.

3) Comparison of Parameters, Running Time, and Computations: Table V provides a comparative analysis of denoising methods considering the number of parameters, computational complexity (GFLOPs), inference time, and average NIQA score. Lightweight CNN-based approaches such as FFDNet, FDnCNN, DnCNN, and ADNet use fewer parameters but achieve relatively low NIQA scores, reflecting limited generalization to industrial acoustic noise. Transformer-based models (SwinIR, Restormer, and Xformer) substantially increase computational demand—SwinIR requiring over 187 GFLOPs—yet deliver only moderate improvements in NIQA. In contrast, the proposed SAMID model attains the highest NIQA score, significantly surpassing all baselines. To further validate these gains, paired *t*-tests confirm the statistical robustness of the improvements. These results underline the effectiveness of combining Mamba

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

TABLE V
COMPARISON OF DIFFERENT DENOISING METHODS IN TERMS OF PARAMETERS, GFLOPS, AND NIQA EVALUATION RESULTS (MEAN ± STANDARD DEVIATION)

Methods	Param (M)	FLOPs (G)	NIQA Average ↑
FFDNet	0.480000	1.980000	1.023 ± 0.639***
FDnCNN	0.555137	9.080000	1.800 ± 1.510***
DnCNN	0.666497	10.900000	8.042 ± 1.361***
CBDNet	4.365382	10.070000	44.265 ± 27.595***
InvDN	2.641000	23.770000	22.194 ± 2.179***
SwinIR	11.500000	187.930000	9.484 ± 5.034***
Restormer	26.093952	35.210000	11.003 ± 2.645***
ADNet	0.519171	8.540000	2.679 ± 1.248***
IRC	0.185857	3.040000	6.483 ± 2.122***
Xformer	25.230000	42.200000	17.108 ± 6.634***
MambaIR	20.400000	28.490000	19.818 ± 4.177***
SAMID	47.855429	27.320000	80.720 ± 2.989

Note: FLOPs were calculated on an NVIDIA RTX 3090 Ti with the input vector size 128 × 128 × 3. Statistical significance was assessed using paired *t*-tests, with symbols *** indicating improvements over each baseline with *p*-values < 0.001.

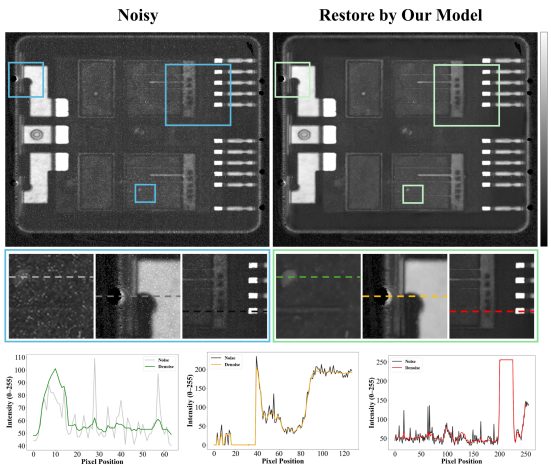


Fig. 10. Model effectiveness analysis.

and Transformer blocks in SAMID, offering a superior tradeoff between denoising quality and computational feasibility for industrial defect inspection.

B. Model Effectiveness Analysis

As shown in Fig. 10, for a clearer comparison, the chart (last line) illustrates representative 1-D profiles taken along the dotted lines in the zoomed regions, showing a direct comparison between pre- and postdenoising results. From the chart, it is evident that the edges after denoising remain similar to, or appear slightly sharper than, those in the noisy images. This suggests that the spatial resolution can be regarded as comparable or slightly enhanced following de-noising. The effectiveness of de-noising is more convincingly demonstrated through side-by-side examination of zoomed regions rather than full-field images, since the differences in performance are more distinct in the magnified areas. These experiments demonstrate that our model is capable of effectively removing strong acoustic noise while maintaining the intrinsic texture and structural characteristics of the samples.

V. ABLATION STUDY

1) Effects of Noise Schedule: The noise schedule strongly influences denoising dynamics. We compared cosine and linear schedules, finding that the linear schedule achieved slightly higher PSNR (31.61 versus 31.57 dB). Although the gain is modest, linear scheduling provides more stable reconstruction and is adopted as the default configuration.

2) Effects of Latent Encoder Design: We compared static and depthwise convolution in the encoder. Depthwise convolution reduced parameters from 585 to 120 k while improving PSNR from 29.74 to 29.88 dB. This demonstrates superior efficiency without sacrificing performance, making it preferable for lightweight design.

3) Effects of DM/DT Ratio: Since no ground-truth datasets exist for real industrial samples, we generated synthetic data by combining a DM and DT. As shown in Table IV, using only DM or only DT degrades performance, particularly for DT alone (27.49 dB). The best results are achieved with a 7:3 DM/DT ratio, yielding 31.66 dB, highlighting the importance of jointly modeling degradation and domain-specific characteristics to accurately simulate real industrial samples.

4) Effects of Diffusion Steps: We varied diffusion steps from 6 to 12. PSNR remained within a narrow range (31.53–31.61 dB), with the best result at 10 steps (31.61 dB). Increasing steps beyond this did not improve performance and sometimes reduced it due to redundancy. Thus, an intermediate step size offers the best tradeoff between quality and efficiency.

VI. CONCLUSION

This article presented a diffusion-based denoising framework tailored for deep-layer SAM, addressing long-standing challenges arising from severe depth-dependent noise and the lack of ground-truth data. By utilizing hierarchical knowledge transfer from top-layer images, the proposed method effectively restored structural details in highly degraded deep-layer SAM images. A physics-guided synthetic data generation strategy enabled robust model training without paired annotations, while the hybrid Transformer–Mamba diffusion architecture captured both spatial structures and long-range acoustic dependences. Comprehensive evaluations on both benchmark and real industrial datasets demonstrated that the proposed method significantly outperformed existing approaches. Our model achieved improvements of up to 33.11 and 29.41 dB in PSNR, as well as 0.89 and 0.73 in SSIM, respectively. It also attained the unsupervised NIQA score of 80.72, substantially exceeding the second-best result of 44.27. When integrated with downstream defect analysis, the proposed framework further yields pronounced gains, reaching 0.96 in DSC and 0.91 in IoU. Overall, this work bridged the gap between acoustic imaging physics and modern generative models, providing a scalable and practical solution for enhancing SAM image quality and reliability in industrial NDT applications. Furthermore, as part of future work, we planned to investigate model compression techniques, including pruning and knowledge distillation, to further reduce computational cost while maintaining denoising performance. In addition, the development of a cloud-based deployment framework is part of our ongoing efforts. Future work will also incorporate physics-informed priors into the diffusion process to better align restoration behavior with acoustic propagation principles. In addition, we aim to build larger, multimaterial SAM datasets to more rigorously validate the model's generalization across diverse materials and imaging conditions. Exploring k-fold cross-validation could also be considered to further enhance model stability and ensure more reliable generalization across diverse sample types.

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] L. P. Bauermann et al., "Scanning acoustic microscopy as a non-destructive imaging tool to localize defects inside battery cells," *J. Power Sources Adv.*, vol. 6, 2020, Art. no. 100035.

[2] H. Yu, "Scanning acoustic microscopy for material evaluation," *Appl. Microsc.*, vol. 50, no. 1, Art. no. 25, 2020.

[3] P. Gong et al., "Noise suppression for ultrasound attenuation coefficient estimation based on spectrum normalization," *IEEE Trans. Ultrasonics, Ferroelectr., Freq. Control*, vol. 68, no. 8, pp. 2667–2674, Aug., 2021.

[4] F. Bertocci, A. Grandoni, and T. Djuric-Rissner, "Scanning acoustic microscopy (SAM): A robust method for defect detection during the manufacturing process of ultrasound probes for medical imaging," *Sensors*, vol. 19, no. 22, 2019, Art. no. 4868.

[5] C. Weng, X. Gu, and H. Jin, "Coded excitation for ultrasonic testing: A review," *Sensors*, vol. 24, no. 7, 2024, Art. no. 2167.

[6] M. I. Silva, E. Malitckii, T. G. Santos, and P. Vilaça, "Review of conventional and advanced non-destructive testing techniques for detection and characterization of small-scale defects," *Prog. Mater. Sci.*, vol. 138, 2023, Art. no. 101155.

[7] S. K. Gupta, R. Pal, A. Ahmad, F. Melandsø, and A. Habib, "Image denoising in acoustic microscopy using block-matching and 4D filter," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 13212.

[8] Á. Makra, W. Bost, I. Kalló, A. Horváth, M. Fournelle, and M. Gyöngy, "Enhancement of acoustic microscopy lateral resolution: A comparison between deep learning and two deconvolution methods," *IEEE Trans. Ultrasonics, Ferroelectr., Freq. Control*, vol. 67, no. 1, pp. 136–145, 2019.

[9] S.-Y. Zhang et al., "HFormer: Highly efficient vision transformer for low-dose CT denoising," *Nucl. Sci. Techn.*, vol. 34, no. 4, 2023, Art. no. 61.

[10] J. Zhang, Z. Shangguan, W. Gong, and Y. Cheng, "A novel denoising method for low-dose CT images based on transformer and CNN," *Comput. Biol. Med.*, vol. 163, 2023, Art. no. 107162.

[11] M. Diwakar and P. Singh, "CT image denoising using multivariate model and its method noise thresholding in non-subsampled Shearlet domain," *Biomed. Signal Process. Control*, vol. 57, 2020, Art. no. 101754.

[12] H. Zhao, L. Qian, Y. Zhu, and D. Tian, "Low dose CT image denoising: A comparative study of deep learning models and training strategies," *AI Med.*, vol. 1, pp. 7–7, 2024.

[13] P. Singh and A. Shankar, "A novel optical image denoising technique using convolutional neural network and anisotropic diffusion for real-time surveillance applications," *J. Real-Time Image Process.*, vol. 18, no. 5, pp. 1711–1728, 2021.

[14] M. Diwakar et al., "Low-dose COVID-19 CT image denoising using CNN and its method noise thresholding," *Curr. Med. Imag. Rev.*, vol. 19, no. 2, pp. 182–193, 2023.

[15] M. Diwakar, P. Singh, and D. Garg, "Edge-guided filtering based CT image denoising using fractional order total variation," *Biomed. Signal Process. Control*, vol. 92, 2024, Art. no. 106072.

[16] S. Katta, P. Singh, D. Garg, V. Ravi, and M. Diwakar, "A dual CT image denoising approach using guided filter and method-based noise in the NSST domain," *Open Bioinf. J.*, vol. 18, no. 1, 2025.

[17] L. Zhang and J. Zhang, "Ultrasound image denoising using generative adversarial networks with residual dense connectivity and weighted joint loss," *PeerJ Comput. Sci.*, vol. 8, 2022, Art. no. e873.

[18] S. Cammarasana, P. Nicolardi, and G. Patanè, "Real-time denoising of ultrasound images based on deep learning," *Med. Biol. Eng. Comput.*, vol. 60, no. 8, pp. 2229–2244, 2022.

[19] C. Qiu, Z. Huang, C. Lin, G. Zhang, and S. Ying, "A despeckling method for ultrasound images utilizing content-aware prior and attention-driven techniques," *Comput. Biol. Med.*, vol. 166, 2023, Art. no. 107515.

[20] K. F. Graff, *Wave Motion in Elastic Solids*. New York, NY, USA: Dover, 1975.

[21] J. L. Rose, *Ultrasonic Waves in Solid Media*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[22] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[23] T. Dao and A. Gu, "Transformers are SSMS: Generalized models and efficient algorithms through structured state space duality," *Proc. Mach. Learn. Res.*, vol. 235, pp. 10041–10071, 2024. [Online]. Available: https://arxiv.org/abs/2405.21060

[24] A. Gu et al., "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*. [Online]. Available: https://arxiv.org/abs/2312.00752

[25] Y. Zhang et al., "LBFormer: Scene perception segmentation transformer based on local block," *IEEE Trans. Ind. Inform.*, vol. 21, no. 2, pp. 1389–1398, Feb., 2025.

[26] Y. Liu et al., "VMamba: Visual state space model," in *Proc. 38th Conf. Neural Inf. Process. Syst.*, 2024, pp. 103031–103063.

[27] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1692–1700.

[28] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul., 2017.

[29] Eastman Kodak Company, "Kodak lossless true color image suite," 1993. [Online]. Available: http://r0k.us/graphics/kodak/

[30] L. Zhang, X. Wu, A. Buades, and X. Li, "Color demosaicking by local directional interpolation and non-local adaptive thresholding," *J. Electron. Imag.*, vol. 20, no. 2, Apr.–Jun. 2011, Art. no. 023016, doi: 10.1117/1.3600632.

[31] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5197–5206.

[32] X. Kong, K. Li, Q. Yang, L. Wenyin, and M.-H. Yang, "A new image quality metric for image auto-denoising," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2888–2895.

[33] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[34] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1, 2007, pp. I-313–I-316.

[35] K. Zhang, W. Zuo, L. Zhang, W. Zhang, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," in *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep., 2018.

[36] L. Li, X. Yu, Z. Jin, Z. Zhao, X. Zhuang, and Z. Liu, "FDNCNN-based image denoising for multi-labfel localization measurement," *Measurement*, vol. 152, 2020, Art. no. 107367.

[37] S. Guo, D. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1712–1722.

[38] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided CNN for image denoising," *Neural Netw.*, vol. 124, pp. 117–129, 2020.

[39] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3929–3938.

[40] Y. Liu et al., "Invertible denoising network: A light solution for real noise removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13365–13374.

[41] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1833–1844.

[42] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5728–5739.

[43] J. Zhang, Y. Zhang, J. Gu, J. Dong, L. Kong, and X. Yang, "Xformer: Hybrid x-shaped transformer for image denoising," 2024. [Online]. Available: https://arxiv.org/abs/2303.06440

[44] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "MambaIR: A simple baseline for image restoration with state-space model," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 222–241.

[45] Y. Cui and A. Knoll, "Enhancing local–global representation learning for image restoration," *IEEE Trans. Ind. Inform.*, vol. 20, no. 4, pp. 6522–6530, Apr., 2024.



Trong Nhan Nguyen received the B.S. degree in information technology from Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2022, and the M.S. degree in Industry 4.0 convergence bionics engineering from Pukyong National University, Busan, South Korea, in 2025.

He has authored and coauthored several research papers published in reputable international journals and conferences, such as *Engineering Applications of Artificial Intelligence* and the IEEE International Conference on Industrial Technology. His current research aims to develop advanced learning-based methods for improving image quality in industrial and biomedical imaging applications. His research interests include image restoration and computer vision, including superresolution, image inpainting, image denoising, low-light image enhancement, image quality assessment, and image recognition.

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.



Vu Hoang Minh Doan received the B.S. degree in mechatronics engineering from the Ho Chi Minh City University of Technology (Bach Khoa University), Ho Chi Minh City, Vietnam, in 2016, and the M.S. and Ph.D. degrees in biomedical engineering from Pukyong National University, Busan, Republic of Korea, in 2019 and 2022, respectively.

He is currently a Research Professor with the Smart Gym-Based Translational Research Center for Active Senior's Healthcare, Pukyong National University. He has authored and coauthored several research articles published in well-known international journals, including *Scientific Reports*, *Engineering Applications of Artificial Intelligence*, and *ACS Biomaterials Science & Engineering*. His research interests include the application of artificial intelligence in biomedical engineering, automation control systems, and intelligent healthcare technologies.



Tan Hung Vo received the M.S. and Ph.D. degrees in Industry 4.0 convergence bionics engineering from Pukyong National University, Busan, Republic of Korea, in 2020 and 2024, respectively.

He is currently working as a Researcher with Pukyong National University. His research interests include intelligent control systems based on neural networks and digital image processing, with applications in advanced engineering and intelligent systems.



Dang Khoa Pham received the B.S. degree in information technology from the University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam, in 2022. He is currently working toward the Ph.D. degree with North Carolina A&T State University, Greensboro, NC, USA.

During his graduation, he developed a strong background in computer science and data-driven technologies. During his academic career, he has been involved in research projects related to image analysis, pattern recognition, and intelligent systems. His current work aims to advance robust and efficient AI models for medical data understanding and interdisciplinary applications, contributing to both theoretical development and practical implementation. His research interests mainly include computer vision, natural language processing, and medical imaging, with an emphasis on applying artificial intelligence and deep learning techniques to real-world healthcare and engineering problems.



Dinh Dat Vu received the M.S. and Ph.D. degrees in Industry 4.0 convergence bionics engineering from Pukyong National University, Busan, South Korea, in 2021 and 2025, respectively.

His research focuses on the development of electrical and electronic applications for scanning acoustic microscopy systems. He has designed and implemented key system components, including high-voltage pulsers and related electronic modules, contributing to the advancement of high-resolution ultrasonic imaging technologies.



Thi Thu Ha Vu received the M.S. and Ph.D. degrees in Industry 4.0 convergence bionics engineering from Pukyong National University, Busan, Republic of Korea, in 2022 and 2025, respectively.

Her research focuses on the application of artificial intelligence, machine learning, and deep learning techniques to image processing and intelligent vision systems. She has authored and coauthored several research papers published in reputable international journals, including IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING and *Applied Sciences*. Her work contributes to the advancement of data-driven methods for high-precision imaging and industrial applications. She is actively involved in interdisciplinary research that bridges artificial intelligence with engineering and applied sciences.



Truong Tien Vo received the B.E. and M.Eng. degrees in electrical engineering from the International University, Vietnam National University, Ho Chi Minh City, Vietnam, in 2011 and 2015, respectively. He is currently working toward the Ph.D. degree in Industry 4.0 convergence bionics engineering with the Department of Biomedical Engineering, Pukyong National University, Busan, Republic of Korea.

His research interests include embedded systems, machine learning, signal processing, and biosensors. He has authored and coauthored several research papers published in international journals and conferences, including IEEE INTERNET OF THINGS JOURNAL, IEEE Sensor conference, etc., contributing to the advancement of intelligent sensing and connected healthcare systems.



Quoc Dung Nguyen received the B.S. degree from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, and the M.S. degree in Industry 4.0 convergence bionics engineering from Pukyong National University, Busan, Republic of Korea.

During his graduate studies, his research focused on the integration of mechanical engineering with smart manufacturing technologies, automation, and interdisciplinary engineering approaches inspired by Industry 4.0. He developed a strong foundation in mechanical engineering and industrial systems. His academic interests include mechanical system design, intelligent manufacturing, and the application of advanced engineering methods to industrial problems. He has participated in various research projects and academic activities related to engineering innovation and technology convergence. His current research aims to contribute to the development of efficient, intelligent, and sustainable engineering solutions for modern industrial applications.



Jaeyeop Choi received the B.S. degree in biomedical engineering and the M.S. and Ph.D. degrees, all from Pukyong National University, Busan, Republic of Korea, in 2017, 2019, and 2022, respectively.

He is currently a Research Professor with the Smart Gym-Based Translational Research Center for Active Senior's Healthcare, Pukyong National University. His research interests include scanning acoustic microscopy and the fabrication of high-frequency ultrasonic transducers, with a focus on advanced biomedical imaging systems and translational healthcare applications.



Junghwan Oh (Member, IEEE) received the B.S. degree in mechanical engineering from Pukyong National University, Busan, Republic of Korea, in 1992, and the M.S. and Ph.D. degrees in biomedical engineering from The University of Texas at Austin, Austin, TX, USA, in 2003 and 2007, respectively.

He is currently a Professor in Industry 4.0 convergence bionics engineering program with the Department of Biomedical Engineering, Pukyong National University. He contributes extensively to interdisciplinary research in biomedical and industrial engineering. His research interests include the development of ultrasonic-based diagnostic imaging modalities for biomedical engineering applications, including advanced imaging systems and signal processing techniques.

Authorized licensed use limited to: Pukyong National Univ. Downloaded on June 16,2026 at 16:26:39 UTC from IEEE Xplore. Restrictions apply.