mathematics

MDPI

Article

Transformative Noise Reduction: Leveraging a Transformer-Based Deep Network for Medical Image Denoising

Rizwan Ali Naqvi [1,†], Amir Haider [1,†], Hak Seob Kim [2], Daesik Jeong [3,*] and Seung-Won Lee [4,*]

[1] Department of AI and Robotics, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea; rizwanali@sejong.ac.kr (R.A.N.); amirhaider@sejong.ac.kr (A.H.)
[2] Korea Agency of Education, Promotion and Information Service in Food, Agriculture, Forestry and Fisheries, Sejong 30148, Republic of Korea; moon0626@eois.or.kr
[3] Division of Software Convergence, Sangmyung University, Seoul 03016, Republic of Korea
[4] School of Medicine, Sungkyunkwan University, Suwon 16419, Republic of Korea
* Correspondence: jungsoft97@smu.ac.kr (D.J.); swleemd@g.skku.edu (S.-W.L.)
† These authors contributed equally to this work.

Abstract: Medical image denoising has numerous real-world applications. Despite their widespread use, existing medical image denoising methods fail to address complex noise patterns and typically generate artifacts in numerous cases. This paper proposes a novel medical image denoising method that learns denoising using an end-to-end learning strategy. Furthermore, the proposed model introduces a novel deep–wider residual block to capture long-distance pixel dependencies for medical image denoising. Additionally, this study proposes leveraging multi-head attention-guided image reconstruction to effectively denoise medical images. Experimental results illustrate that the proposed method outperforms existing qualitative and quantitative evaluation methods for numerous medical image modalities. The proposed method can outperform state-of-the-art models for various medical image modalities. It illustrates a significant performance gain over its counterparts, with a cumulative PSNR score of 8.79 dB. The proposed method can also denoise noisy real-world medical images and improve clinical application performance such as abnormality detection.

Keywords: medical image denoising; deep–wider residual block; multi-head attention; multi-modal denoising; deep learning

MSC: 68T07

Citation: Naqvi, R.A.; Haider, A.; Kim, H.S.; Jeong, D.; Lee, S.-W. Transformative Noise Reduction: Leveraging a Transformer-Based Deep Network for Medical Image Denoising. *Mathematics* **2024**, *12*, 2313. https://doi.org/10.3390/math12152313

Academic Editor: James Chung-Wai Cheung, Yanping Huang

Received: 2 July 2024
Revised: 18 July 2024
Accepted: 20 July 2024
Published: 24 July 2024

1. Introduction

Noise is widespread in medical images because of the characteristics of medical imaging, including the image acquisition process and respiratory movement. Such arbitrary modifications to the acquired images can significantly degrade the perceptual quality by incorporating numerous artifacts and obscuring salient details. Consequently, the performance of image analysis algorithms, such as segmentation, registration, and classification, are affected. Additionally, these degraded images directly affect the decision-making processes of medical practitioners. Despite its numerous real-world implications, medical image denoising (MID) is challenging, as it necessitates the preservation of crucial diagnostic information while effectively reducing noise [1–3].

MID, which is a challenging topic, is widely investigated by the vision community. Initially, classical image processing techniques such as non-local self-similarity [4], sparse coding [5], and filter-based approaches [6–8] were employed for MID. However, the current state-of-the-art denoising methods involve deep learning using two learning strategies: learning denoising as image-to-image translation and learning residual noise from noisy images. Although these learning-based denoising methods have shown promising results compared with classical approaches, their performance remains limited, and they fail in extreme cases (i.e., Gaussian noise at sigma = 50), as shown in Figure 1.

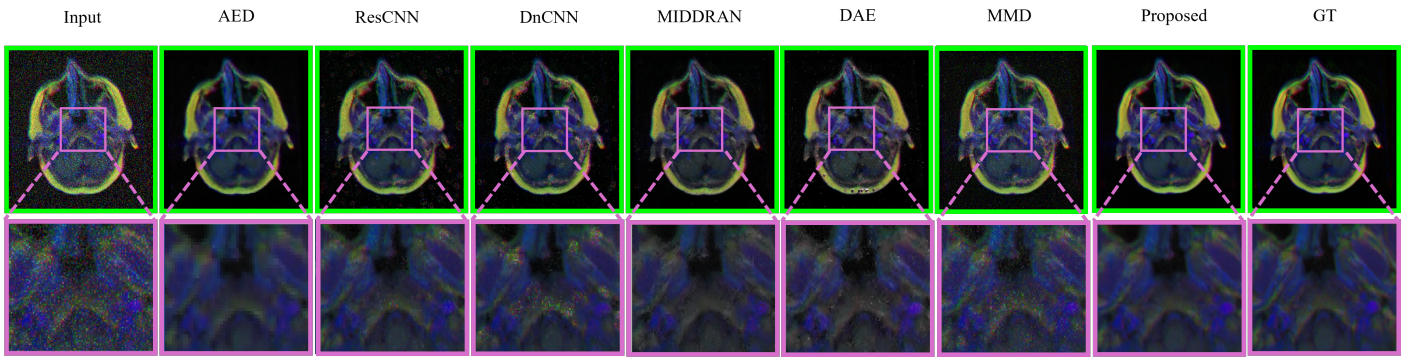


Figure 1. Comparison between existing MID and the proposed method. Existing denoising methods typically yield smooth denoising results with visual artifacts. The proposed method can clean noisy medical images and address the limitations of existing methods. Left to right: noisy input, AED [9], ResCNN [10], DnCNN [11], MIDDRAN [12], DAE [13], MMD [3], the proposed method, and the reference image.

The current image-to-image denoising methods can effectively remove noise from images. However, these methods typically result in oversmoothing in cases involving complex spatial distributions. Consequently, essential features and details are not preserved, which is a critical issue in medical imaging. Meanwhile, residual denoising strategies are exempt from smooth spatial representations. However, these strategies may generate visually disturbing artifacts with desaturated complex structures. In general, both existing MID methods fail to reconstruct detailed, natural-appearing images similar to the reference ground truths. This underscores the urgency and significance of our study for accurately preserving the details of medical images.

Despite the severe limitations of existing denoising methods, MID presents numerous real-world implications. The widespread applicability and importance of MID motivated us to develop an effective method for managing the diverse and high levels of imaging noise present in numerous medical imaging modalities. Additionally, the recent success of transformer [14,15] models inspired us to investigate transformer-based attention in MID to shift the paradigm of medical image denoising research.

This study proposes a novel deep method to effectively learn MID and address its limitations, thus steering MID research in a new direction. The proposed method introduces a novel deep and wide residual (DWR) block to learn the underlying noise in medical images. Despite being deep in architecture, the proposed block leverages a dilated convolution operation to capture long-distance intrapixel dependencies while rendering the block more efficient. The proposed DWR block addresses smoothing artifacts, which are widespread in existing image-to-image translation-based denoising methods. Additionally, the proposed method leverages a multihead attention (MHA) block [14] to reconstruct denoised images in the latter half of the proposed network. The proposed MHA addresses the limitations of residual denoising approaches by reconstructing plausible, high-quality images without generating visual artifacts. The proposed method was extensively investigated and compared with existing MID methods on different medical imaging modalities. The practicability of the proposed method was evaluated using noisy real-world medical images. The contributions of the current study are as follows:

- A novel transformer-attention-based deep architecture is proposed that can address the limitations of existing MID methods.
- A novel DWR module is proposed to learn long-distance pixel dependencies in order to perform MID efficiently. Additionally, this study proposes to leverage MHA in the decoder to mitigate artifacts from denoised images.
- Dense experiments conducted on numerous medical modalities show that the proposed method substantially outperforms existing MID methods based on qualitative and quantitative comparisons.

- The effectiveness of the proposed method is investigated based on real-world noisy medical images, and its practicability is analyzed for real-world usage.

The remainder of this paper is organized as follows: Section 2 reviews the related studies, Section 3 details the data simulation and learning strategy, Section 4 presents an analysis of the experimental results, and Section 5 concludes this paper.

2. Related Studies

MID is considered one of the most challenging enhancement tasks in medical imaging. Hence, numerous novel approaches for addressing MID have been introduced in recent years. However, learning-based methods are superior to their classical counterparts. This section briefly reviews the learning-based approaches.

2.1. Image-to-Image Translation

Deep learning is widely used in MID [3,16,17]. The most recent studies have considered denoising to be an image-to-image translation task. Gondara et al. proposed a convolutional autoencoder (CAE [9]), and Walid et al. proposed a denoising autoencoder (DAE [13]) to learn additive denoising. Chen et al. and Fan et al. proposed a residual encoder–decoder convolutional neural network [18] and a quadratic autoencoder [19], respectively, to denoise low-dose computed tomography (CT) images. Hyun et al. used U-Net denoising and k-space correction simultaneously to denoise magnetic resonance images [20]. Similarly, Kidoh et al. designed a shrinkage convolutional neural network (SCNN) and a deep-learning-based reconstruction (dDLR) network to denoise brain magnetic resonance images [21]. Rawat et al. proposed a feature-guided denoising convolutional neural network for learning additive noise reduction [22].

A few recent studies have leveraged adversarial training strategies. For example, Ghahremani et al. [23] comprehensively investigated MID using a U-Net and adversarial guidance. Zhou et al. proposed a unified motion correction and denoising adversarial network [24]. Li et al. utilized a conditional generative adversarial network to reduce random noise in CT images [25]. Similarly, Jianning et al. [26] proposed a multilevel discriminator to denoise CT images. Notably, such image-to-image translation approaches typically yield smoother outputs with less-detailed edges and textures compared with their conventional counterparts.

2.2. Residual-Noise Estimation

Recent studies have addressed the challenge of denoising noisy medical images by learning the underlying noise patterns instead of relying solely on image-to-image translation techniques. Jiang et al. [11] employed a denoising convolutional neural network (DnCNN) [27] that was specifically designed for magnetic resonance image denoising. Their main aim was to enhance image quality by effectively removing noise while preserving important image features. Based on this study, Jifara et al. and Walid et al. [3,10] further improved the performance of a DnCNN by modifying the network architecture such that it can manage the complex noise patterns inherent in medical images more effectively. Similarly, Kokil et al. [28] proposed the use of a residual learning network to address speckle noise in medical images. Their method effectively mitigated noise artifacts by leveraging residual connections while maintaining the image details. More recently, Sharif et al. [12] introduced a dynamic residual attention network (DRAN) designed to learn residual noise patterns from multimodal medical images. This approach adapts attention mechanisms to focus on the relevant image regions, thereby enabling accurate noise removal across different imaging modalities.

Although these residual denoising methods are promising for generating sharper and cleaner images, their potential limitations must be acknowledged, particularly in extreme cases where they may inadvertently introduce visual artifacts. Thus, studies are being actively conducted to further refine these algorithms so that noise reduction can

be achieved simultaneously with the preservation of crucial image features in medical imaging applications.

The proposed method further extends the possibility of a generic denoising method that can perform multipattern denoising on multiple imaging modalities. Notably, a generic denoising method can offer many advantages, such as the sharing of domain knowledge among numerous imaging modalities. Table 1 shows a comparison between existing methods and the proposed method.

Table 1. Comparison between existing denoising methods and proposed two-stage network.

Method	Learning Strategy	Strengths	Weaknesses
Image-to-image translation	Translates noisy image into clean image	• Can outperform conventional (non-deep-learning) approaches • Easy to train and infer	• Tends to yield smooth images with fewer details • Limited to specific noise types/modality
Residual denoising	Learns underlying noise from noisy image	• Can achieve sharper images • Well-known for Gaussian denoising	• Yields visual artifacts in extreme cases (high noise) • Can estimate only a specific noise pattern
Proposed method	Denoises medical images utilizing DWR and MHA	• Outperforms existing MID methods in visual and quantitative comparison • Modality-independent deep denoiser that can manage real and synthetic data • Computationally lightweight	• Optimized for desktop-class hardware

3. Method

This section describes the process of preparing the data for learning MID. Additionally, insights into the proposed novel deep model and its components are presented.

3.1. Data Preparation

The preparation of large-scale data samples to learn MID is challenging. Only a few real-world data samples are available for open MID research. Therefore, in this study, we obtained large-scale MID data samples and simulated Gaussian noise to learn generic MID.

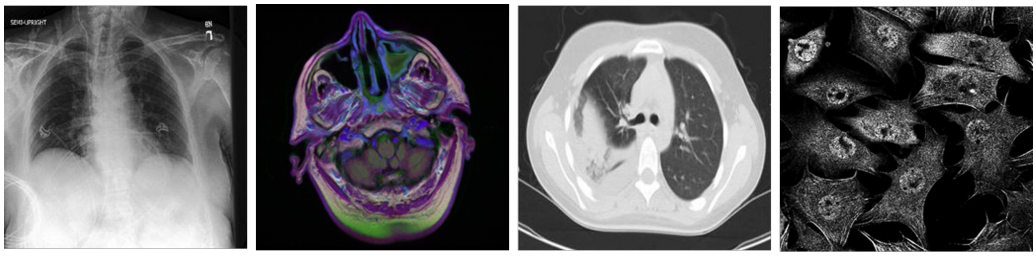
3.1.1. Data Acquisition

One of the main motivations of this study was to generalize and illustrate the practicability of deep denoising in diverse medical imaging modalities. Therefore, we investigated the following modalities to learn MID efficiently:

- X-ray imaging is widely used for diagnosing bone fractures, joint problems, lung conditions, dental issues, etc. This study leverages the well-known Chexpert [29] dataset to represent X-ray images.
- Magnetic resonance imaging (MRI) is an effective medical imaging technique that uses magnetic fields and radio waves to generate detailed images of the body's internal structures. It is crucial for diagnosing various conditions from brain tumors to joint injuries. This study leverages the dataset presented in [30] to learn MID for MRI.
- CT is a diagnostic imaging method that uses X-rays to create cross-sectional images of the body, thus providing detailed views of internal structures and aiding in the detection and diagnosis of various medical conditions such as fractures, tumors, and internal bleeding. The scan dataset presented in [31] was used to learn MID in CT images.
- Microscopy provides high-resolution images that reveal the intricate details of minute biological structures, cells, tissues, and microorganisms, and it is essential for advancing our understanding of biology, medicine, and various scientific disciplines.

Furthermore, microscopic images typically contain Gaussian noise, which exhibits various pixel intensities, thus complicating accurate analyses and interpretations in fields such as biology and materials science. Thus, protein atlas scans [32] were used to investigated MID.

We obtained 20,000 random images for training and 1000 images for validation while learning MID. An additional 4000 samples from the obtained data (1000 images from each modality) were used for an extensive evaluation with various noise factors. Figure 2 shows the representative images of the imaging modalities used in this study. Notably, a fixed noise deviation was used to test the deep model in the testing phase and to realize an unbiased comparison among the deep models. However, noise was randomly generated during training to diversify the data and avoid overfitting. We simulated Gaussian noise in these samples to learn MID and to analyze the performance of the deep networks. Notably, the proposed study leveraged only imaging modalities that are commonly used for diagnoses and incorporated Gaussian noise.



(a) **(b)** **(c)** **(d)**

Figure 2. Representative images obtained via each imaging modality: (**a**) X-ray; (**b**) MRI; (**c**) CT; (**d**) microscopy.

In addition to simulating noise for extensive evaluation, we incorporated noisy medical images to illustrate its practicability in actual applications.

3.1.2. Noise Simulation

Noise in medical images is typically considered an additive factor and can be represented as

$$n_s \sim \mathcal{N}(I_R|\mu, \sigma^2) \tag{1}$$

Here, μ and σ^2 denote the mean and variance of the Gaussian distribution ($\mathcal{N}$), respectively.

Considering this basic principle, we added Gaussian noise to a clean image I_R. To learn MID efficiently, we simulated noise in the acquired data samples. Therefore, reference noisy image pairs must be formulated by contaminating them with artificial noise. In this study, a uniform noisy image I_N was generated from a noise-free image I_R as follows:

$$I_N = I_R + n_s \tag{2}$$

The illustration presented in Figure 3 shows an example of a noisy–clean image pair alongside the corresponding generated noise. Notably, the noise simulation process incorporates a crucial element: the random standard deviation of the noise distribution. We tuned n_s such that the standard deviation varied randomly from 0 to 75. This wide range of noise deviation allowed us to extensively evaluate the capability of deep models for a diverse range of noise patterns and levels.

This method was deliberately designed to introduce variability in the intensity of the generated noise. It aims to mimic the diverse spectra of noise encountered in real-world scenarios. This variability is essential for creating a realistic representation of noise that reflects the nuances observed in practical settings.

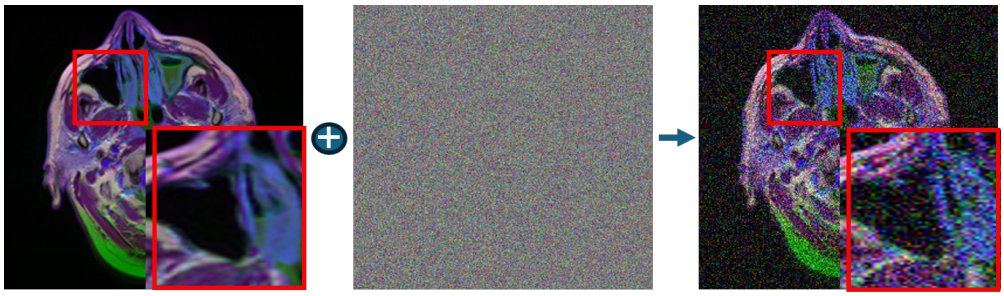


Figure 3. Gaussian noise simulation for learning medical image denoising. This study incorporated noise simulation to learn and evaluate MID methods using numerous medical imaging modalities. From left to right: clean image, random noise (simulated), and noisy image (clean image + generated noise).

3.2. Learning from Data

The proposed method introduces a novel deep architecture to learn MID effectively. The proposed network aims to learn MID as M : $I_N \rightarrow I_C$. Here, the mapping function (M) learns to generate a clean medical image (I_C) from a noisy input (I_N) as $I_C \in [0,1]^{H \times W \times 3}$. Meanwhile, H and W represent the height and width, respectively, of the input and output images.

3.2.1. Network Architecture

As shown in Figure 4, the proposed network regards MID as an image-to-image translation task. The proposed deep network is designed to leverage the advantages of a feature pyramid structure [33] with a DWR module and the MHA to obtain plausible images. The proposed DWR module allows the method to perceive long-distance pixel correlations to understand the spatial relations between neighboring pixels. Additionally, the proposed MHA enables the proposed method to exploit learned long-distance pixel dependencies while performing image reconstruction through decoding. Meanwhile, the features learned at different feature levels are propagated using a contextual gating mechanism to leverage spatial awareness and reduce the underlying noise of the encoder blocks. Notably, the early layer of the denoising networks encodes raw noise and salient features. Therefore, propagating such features to decode a clean image can result in noisy images, despite efficient feature encoding. This study leverages a feature gate to refine the encoded features and address this limitation.

Additionally, the proposed deep network presents a fully convolutional encoder–decoder architecture [34,35] that features convolutional skip connections. The initial layer of the generator transforms the input image (I_L) into a 64-depth feature map. This input convolutional layer employs a kernel size of 3×3, padding of 1, and stride of 1. Meanwhile, the encoder comprises four consecutive feature levels with alternating feature depths of $d = 64, 96, 128, 160$. Following each DWR block in the encoder, a convolutional downsampling layer is applied as follows:

$$F_{\downarrow} = C_{\downarrow}(X) \tag{3}$$

Here, $C_{\downarrow}$ represents a 3×3 convolution operation with a stride of 2.

In addition to the encoder, the proposed architecture includes a decoder that efficiently reconstructs noise-free images. The proposed decoder leverages a DWR block, followed by an MHA block and an upsampling block. The decoder section of the network mirrors the encoder in terms of the number of feature levels, with an upsampling layer following each residual block. The upsampling operation is implemented as follows:

$$F_{\uparrow} = C_{\uparrow}(X) \tag{4}$$

Here, $F_{\uparrow}$ involves a transpose convolution operation [36], which renders the model fully convolutional and results in effective restoration.

When traversing the feature levels, the decoder has the same dimensions as the encoder. Additionally, to propagate features between blocks of the same dimensions for efficient denoising, we leverage a convolutional gate to refine the features while propagating them for reconstruction. The convolutional gating mechanism is perceived as follows:

$$F_G = C_{1\times1}(X) \tag{5}$$

Here, $C_{1\times1}$ represents a point-wise convolutional operation with a kernel size of 1×1. Finally, the decoder portion culminates in a final convolutional layer, which yields a three-channel enhanced image based on a convolutional kernel size of 3×3, padding of 1, and a stride of 1. This final output layer is activated using a tanh function to obtain the final images within the $[0, 1]$ range.

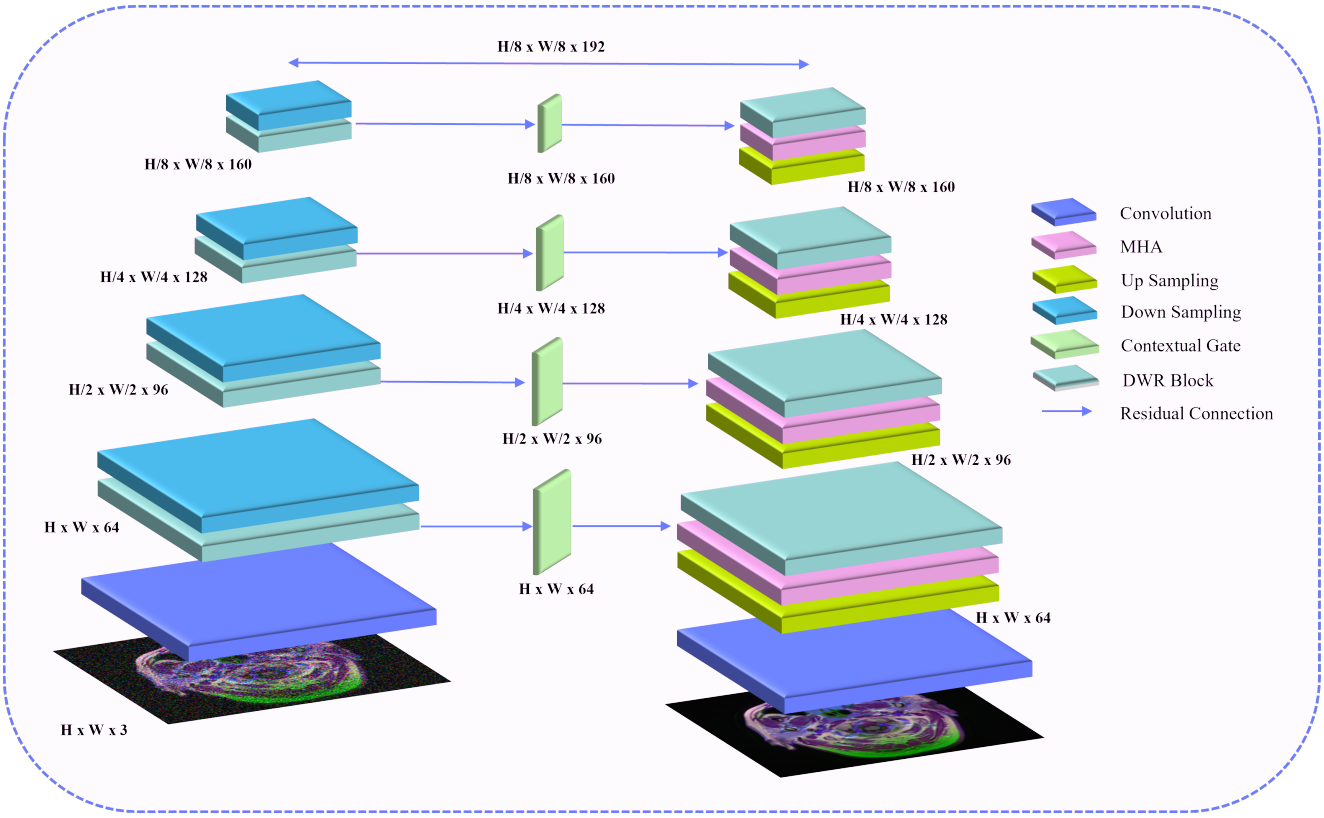


Figure 4. Overview of the proposed novel MID network. The proposed method allows the network to encode salient features in high-dimensional space and to learn to reconstruct clean images by decoding the encoded features. The proposed network incorporates a novel DWR module to capture long-distance pixel dependencies and an MHA block to perform effective reconstruction.

3.2.2. DWR Module

Residual blocks [37] have been proven to be efficient for learning image denoising. Typically, the residual block learns the input feature X using the following equation:

$$R = \mathrm{D}(X) + X \tag{6}$$

Here, $\mathrm{D}(\cdot)$ represents vanilla residual blocks with consecutive convolutional operations, which present a few notable limitations. For example, they cannot extract salient features using a shallow architecture. Therefore, recent studies pertaining to denoising using residual blocks have used consecutive blocks to create a deeper architecture to learn denoising effectively. However, we discovered that such an approach renders the convolu-

tional architecture computationally expensive. Despite their high complexity, conventional residual blocks cannot capture long-distance pixel dependencies because of their narrow receptive fields. Hence, we propose a novel residual block to address the mentioned problems. Figure 5 presents an overview of and a comparison between the proposed DWR and conventional residual blocks.

As shown in Figure 5, the proposed DWR module replaces the convolutional operation of Figure 5b with two consecutive dilated convolutions. Here, the proposed dilated convolution operation leverages a dilation of size 4. Significant dilation enables the proposed DWR module to encompass a wider receptive area and capture long-distance pixel-wise dependencies. Additionally, consecutive dilated convolutions allow the proposed network to traverse deeper without incorporating consecutive residual blocks. Apart from capturing long-distance dependencies, such an architecture can avoid gradient-diminishing problems, which are inherent in consecutive residual blocks. The proposed DWR module allows the deep architecture to traverse deeper with a wider receptive field without exponentially increasing the computational complexity. Based on the architectural modifications, Equation (6) can be derived as follows:

$$R' = \mathrm{D}'(X) + X \tag{7}$$

Here, D incorporates a 1×1 convolution and is followed by two convolutions comprising a kernel size of 3×3, stride size of 1, padding size of 4, and dilation size of 4. The final layer of the proposed DWR module is a point-wise convolution. Here, we used point-wise convolutions to reduce the computational complexity while introducing adaptive channel interactions to render the architecture more efficient.

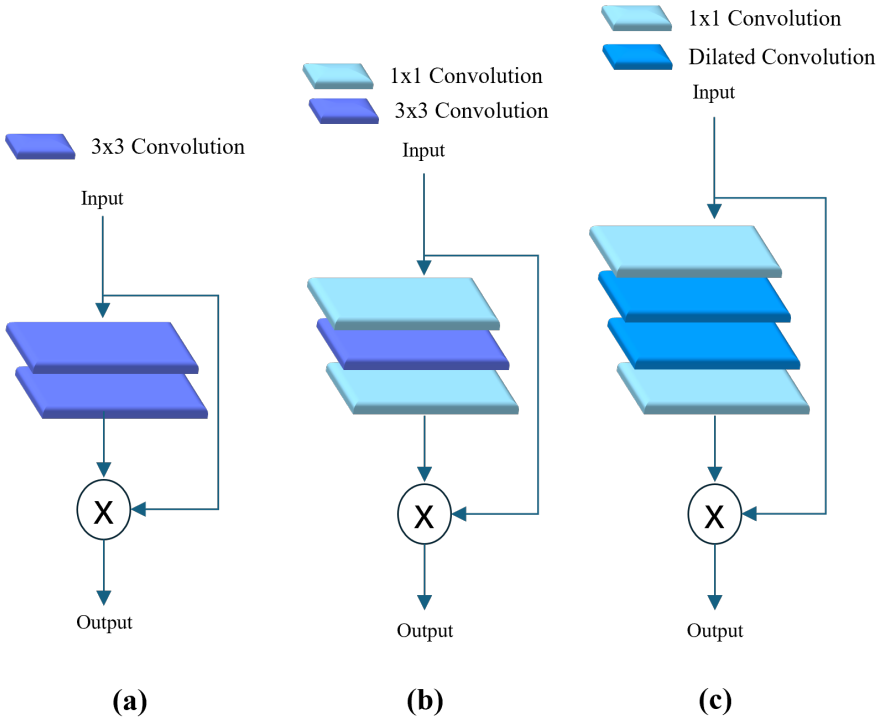


Figure 5. Comparison between vanilla residual blocks and proposed DWR. Proposed DWR block design captures long-distance pixel dependencies to learn efficient denoising. (**a**) Residual block; (**b**) bottleneck residual block; (**c**) proposed deep–wider residual block.

3.2.3. MHA

MHA is a pivotal mechanism in artificial intelligence that is prominently utilized across diverse domains such as natural language processing and computer vision [14,15]. It empowers models to focus concurrently on multiple segments of the input sequence,

thereby facilitating the capture of intricate dependencies and correlations within the tensor. At its core, the MHA block processes input embeddings through query, key, and value matrices and computes attention scores to ascertain the relevance of each element to the others in the sequence. Through a multistep process involving attention score computation, softmax normalization, and weighted value aggregation, the MHA enables the model to simultaneously attend to various aspects of the input, thus enhancing its ability to discern complex patterns and nuances within the specified tensor.

Considering the widespread success of the MHA, we suggested its incorporation into the proposed network architecture to efficiently process long-range dependencies and capture complex patterns extracted by the proposed DWR module. Figure 6 provides an overview of the MHA block. We conceptualize the MHA as follows:

$$MHA(Q, K, V) = concat(head_1, \ldots, head_h)W^O \tag{8}$$

$$\text{where} \quad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{9}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

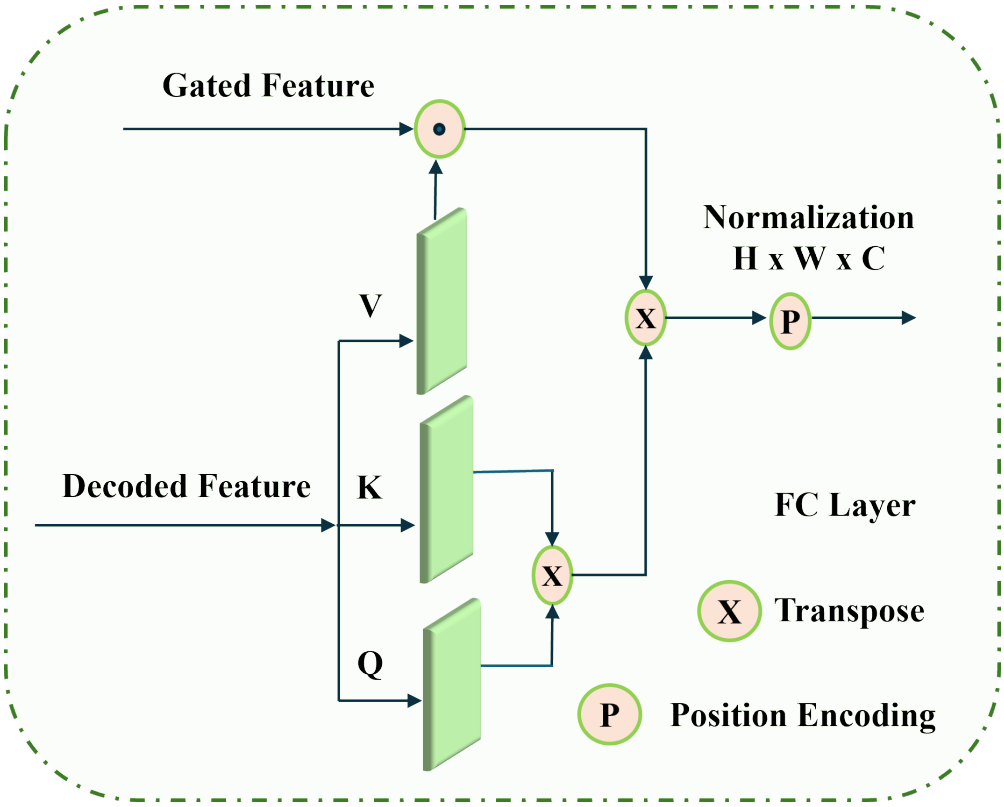


Figure 6. Overview of proposed MHA, which enables proposed network to reconstruct clean and artifact-free medical images while performing denoising.

In our approach, we employ three crucial matrices: Q, K, and V, which represent the query, key, and value matrices, respectively. These matrices contribute significantly to capturing various aspects of the input data. Each attention head, denoted by head_i, yields an output, thus enabling the model to focus on different input components simultaneously. To derive the final output, the outputs from all the attention heads are concatenated along the feature dimension and then multiplied by the output weight matrix W^O. Moreover, each attention head possesses its own set of learnable weight matrices, i.e., W_i^Q, W_i^K, and W_i^V, thus enabling the model to learn distinct representations for different attention heads.

The number of attention heads, denoted by h, and the dimensionality of the key vectors, d_k, are hyperparameters that affect the capacity of the model to capture intricate dependencies within the data. Finally, the softmax activation function is used to compute the attention scores, thereby facilitating the weighted aggregation of values based on their relevance to the queries.

3.2.4. Learning Objective

We applied pixel-wise reconstruction loss to steer our deep model through a coarse-to-refined reconstruction process. The L1 or L2 distance typically serves as a pixel-wise loss function. Whereas both options are commonly used, the L2 loss is directly related to the peak signal-to-noise ratio (PSNR) and typically yields smoother images [38,39]. However, because low-light images contain significant sensor noise, we selected the L1 objective function as the reconstruction loss.

The reconstruction loss can be represented as follows:

$$\mathcal{L}_D = \| I_R - I_C \|_1 \tag{11}$$

Here, I_C represents the output obtained via M, and I_R denotes the reference clean image. This loss function quantifies the absolute differences between the corresponding pixels in the output and reference images, thereby allowing the network to minimize these differences during training.

3.3. Learning Details

The proposed deep network for effective MID was implemented with the PyTorch framework [40]. We optimized our method in the learning phase with an Adam optimizer [41]. We tuned its hyperparameters as $\beta_1 = 0.9$, $\beta_2 = 0.99$. Initially, we set the learning rate for the model as $\eta_i = 1 \times 10^{-4}$. The proposed method was trained for 50,000 steps with a batch size of 24 with synthesized data (for extensive comparison). We adjusted the learning rate with the `ReduceLROnPlateau` scheduler [40] by reducing η_i by a factor of 0.1. For the training phase, we utilized image patches with dimensions of $128 \times 128 \times 3$. All experiments were executed on low-end hardware featuring an AMD Ryzen 3200G central processing unit (CPU) operating at 3.6 GHz, complemented by 32 GB of random-access memory and an NVIDIA GeForce GTX 3060 (12 GB) graphical processing unit (GPU).

In addition to tuning the hyperparameters and setting up the hardware, we leveraged a sophisticated training strategy to ensure the proposed model's convergence with noisy data. As Algorithm 1 shows, the proposed method was trained by generating random noise between 0 and 75 for each mini-batch from the training set D_{train}. Such random noise generation helps the proposed method avoid overfitting while learning denoising. The training process iterated over 50,000 training attempts, adjusting the learning rate every 2500 steps to facilitate convergence. Additionally, objective loss was computed for each mini-batch, and the model weights were updated using the Adam optimizer. Every 5000 steps, the model weights were saved as checkpoints.

During the training, we observed the convergence of the proposed method by observing the training loss and the PSNR score for each step. Figure 7 illustrates the training process of the proposed method. It can be seen that the proposed method learned to address the noise more precisely with each training step. Besides minimizing the objective loss, the proposed network improved its PSNR performance. This ensured the convergence of the proposed method on the given MID data. In the experiment, we found that the proposed method had converged by 50,000 steps, and training beyond that did not drastically improve the proposed method's performance. It took less than 24 h to train the proposed method on our hardware.

Algorithm 1 Training algorithm of the proposed method

Input: Training set D_{train}, validation set D_{val}
Output: Trained deep model M
Initialize CNN model M with random weights
Initialize learning rate η_0, initial batch size B_0, number of steps N_{steps}, learning rate decay factor α
$\sigma = 75$—maximum standard deviation of Gaussian noise
Initialize step counter $e = 1$
for $i = 1$ to N_{epochs} **do**
  if $i \mod 2500 = 0$ **then**
    Update learning rate: $\eta_i = \alpha \cdot \eta_{i-1}$
  Sample mini-batches B_{train} from D_{train} with augmentation
  $R_N \leftarrow uniform(0, 75)$
  $B_{train} \leftarrow noise(D_{train}, R_N)$
  for each mini-batch B_{train} **do**
    Compute loss L on B_{train}
    Update weights of M
  if $i \mod 2 = 5000$ **then**
    Save current weights of M as current weights: $current\_weights = M$.get_weights()

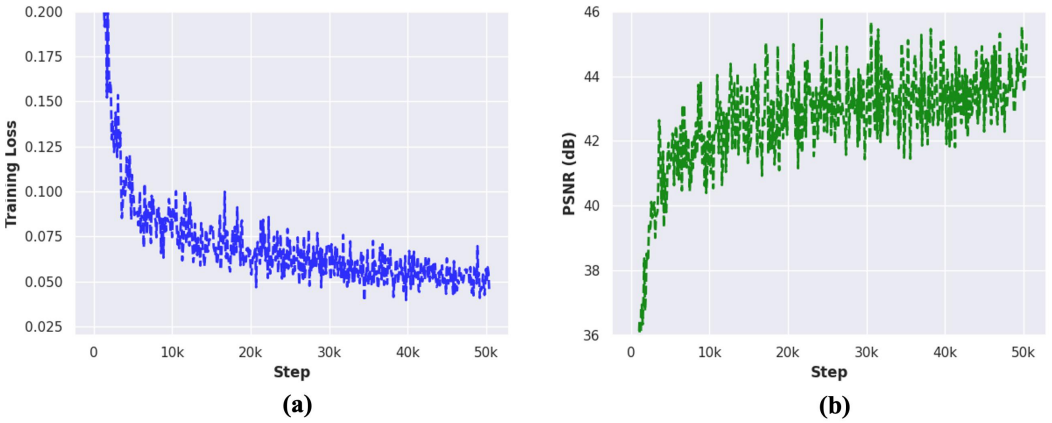


Figure 7. Learning process of proposed network. Proposed method was trained for 50,000 steps. Convergence was determined by considering training loss and PSNR scores. (**a**) Training loss vs. steps; (**b**) PSNR vs. steps.

4. Experiments

The proposed method was evaluated and compared with existing MID methods to determine its practicability for diverse medical imaging modalities. The performance of the proposed method was qualitatively and quantitatively evaluated using noisy real-world medical images. Furthermore, we evaluated the practicability of the proposed components and analyzed the inference performance of the proposed method via sophisticated experiments.

4.1. Comparison with State-of-the-Art Methods

This section presents a comparison of existing MID methods with the proposed deep method. The proposed method was evaluated using noisy real-world medical images, and its parameters were analyzed to demonstrate its practicability for real-world usage.

4.1.1. Comparison Setup

The MID methods were evaluated using four imaging modalities: MRI, X-ray imaging, microscopy, and CT. We incorporated Gaussian and speckle noise with distinct noise levels (i.e., 20, 25, 50, and 75) into each image sample and summarized the performance using the following evaluation metrics:

- PSNR: This is commonly used in image denoising to measure the quality of denoised images based on a comparison with the original noisy image. Higher PSNR scores represent better visual quality of generated images. Equation (12) presents the derivation of the PSNR.

$$\text{PSNR}(I_1, I_2) = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(I_1, I_2)} \right) \tag{12}$$

  where H, W, I_G, and I_C represent the image height, image width, ground truth image, and reconstructed image, respectively. The term c is an index of the image channels.
- SSIM: This is a widely used metric for image quality assessment. This study utilized the SSIM to compare the structural information of generated and ground truth images. A higher SSIM score represents better structural reconstruction. We calculated the SSIM score as follows:

$$\text{SSIM}(I_G, I_C) = \frac{(2\mu_{I_G}\mu_{I_C} + c_1)(2\sigma_{I_G I_C} + c_2)}{(\mu_{I_G}^2 + \mu_{I_C}^2 + c_1)(\sigma_{I_G}^2 + \sigma_{I_C}^2 + c_2)} \tag{13}$$

  where I_G and I_C represent the ground truth and denoised images, respectively; μ_x and μ_y are the mean values of I_G and I_C, respectively; σ_x^2 and σ_y^2 are the variances of I_G and I_C, respectively; $\sigma_{I_G I_C}$ is the covariance of I_G and I_C.
- LLIPS: In addition to the standard quantitative metrics, we used another well-known perceptual metric, i.e., the LLIPS, to summarize the performance of the deep models in terms of perceptual perspective. Specifically, we leveraged the LLIPS with Alexnet pretrained weights. The reference and denoised images were compared quantitatively by calculating the LLIPS as follows:

$$\mathcal{L}_{LLIPS} = \| I_G - I_H \|_1 \tag{14}$$

We compared the performances of state-of-the-art residual denoising methods (i.e., ResCNN [10], DnCNN [11], MMD [3], MID-DRAN [12]) and image-to-image translation medical image denoising methods (i.e., CAE [9], DAE [13]). It is worth noting that most of the existing medical image denoising methods are not publicly available. Therefore, based on their available implementation information, we implemented the existing denoising methods. Further, we trained all these methods using their suggested hyperparameters and their suggested datasets to cross-check our implementation. We added only those methods for which we were able to reproduce their reported results for a fair comparison. Later, we trained all these methods using the same data samples with their suggested hyperparameters. We tested the performance of these methods and compared them with the proposed network for numerous medical image modalities and noise levels. We summarized the performance of these methods using standard metrics. In addition to the quantitative comparison, we compared the MID methods with visual comparisons. Therefore, the strengths and weaknesses of each method can be quantified with visual observations.

4.1.2. Quantitative Evaluation

Table 2 presents a quantitative comparison between the state-of-the-art MID models and the proposed method. As shown, the proposed method outperformed existing techniques substantially for numerous medical imaging modalities. Notably, the proposed method demonstrated consistency for all noise levels. In addition to the conventional evaluation metrics, such as the PSNR and SSIM, the proposed method is superior in terms of the perceptual evaluation metrics. The proposed method can achieve a higher fidelity

score than existing methods. Compared with its counterpart MID model, it achieved higher PSNR, SSIM, and LLIPs values by 8.79 dB, 0.07, and 0.09, respectively. The significant improvement of the proposed method compared with existing methods for numerous modalities and noise levels confirm its practicability for generic cases.

Table 2. Quantitative comparison between existing MID models and the proposed deep network. The proposed method outperforms the existing models by a large margin for MID. Notably, the performance of the proposed method is consistent overall when comparing noise levels and imaging modalities.

Model	σ	Chexpert			CT			MRI			Microscopy			Combined		
		PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓
AED	10	30.43	0.9178	0.1078	27.39	0.8882	0.1361	33.93	0.9375	0.0680	32.07	0.9094	0.0784	30.95	0.9132	0.0976
DnCNN		26.19	0.7812	0.2786	23.29	0.6763	0.2260	26.53	0.7131	0.1697	30.34	0.8660	0.0933	26.59	0.7592	0.1919
ResCNN		24.77	0.7455	0.3324	23.92	0.7214	0.1859	26.68	0.7517	0.1610	30.64	0.8710	0.1012	26.50	0.7724	0.1951
DRAN		33.35	0.9236	0.0622	36.72	0.9624	0.0162	35.15	0.9442	0.0409	37.11	0.9693	0.0330	35.58	0.9499	0.0381
MMD		27.63	0.8537	0.1771	25.05	0.7498	0.1582	24.88	0.6848	0.2211	29.55	0.8514	0.1224	26.78	0.7849	0.1697
DAE		24.10	0.8383	0.1968	19.00	0.8008	0.1752	29.72	0.8028	0.1475	27.61	0.6417	0.2169	25.11	0.7709	0.1841
Proposed		**37.19**	**0.9685**	**0.0130**	**41.55**	**0.9856**	**0.0037**	**42.55**	**0.9819**	**0.0062**	**43.13**	**0.9892**	**0.0053**	**41.11**	**0.9813**	**0.0070**
AED	25	30.51	0.9150	0.1046	27.09	0.8709	0.1488	33.72	0.9318	0.0710	31.95	0.9053	0.0795	30.82	0.9057	0.1010
DnCNN		28.01	0.8299	0.2074	25.22	0.7448	0.1743	27.82	0.7689	0.1757	29.95	0.8541	0.1246	27.75	0.7994	0.1705
ResCNN		29.04	0.8603	0.1676	26.50	0.8010	0.1261	28.84	0.7704	0.1993	30.63	0.8449	0.1379	28.75	0.8192	0.1578
DRAN		30.84	0.8828	0.1110	32.01	0.8950	0.0657	34.38	0.9270	0.0728	34.91	0.9408	0.0509	33.03	0.9114	0.0751
MMD		30.28	0.8749	0.1336	26.84	0.7929	0.1367	27.98	0.7709	0.1780	31.14	0.8759	0.1080	29.06	0.8287	0.1391
DAE		24.67	0.8389	0.1806	19.17	0.7938	0.1727	29.01	0.7954	0.1857	27.60	0.6542	0.2008	25.11	0.7706	0.1849
Proposed		**36.94**	**0.9670**	**0.0145**	**40.07**	**0.9825**	**0.0046**	**40.83**	**0.9787**	**0.0082**	**41.25**	**0.9841**	**0.0076**	**39.77**	**0.9781**	**0.0087**
AED	50	30.22	0.9071	0.1108	27.27	0.8778	0.1431	33.28	0.9211	0.0802	31.73	0.8993	0.0860	30.63	0.9013	0.1051
DnCNN		28.10	0.8335	0.2166	26.55	0.8134	0.1330	26.83	0.7171	0.2805	27.20	0.7543	0.2151	27.17	0.7796	0.2113
ResCNN		29.27	0.8781	0.1589	27.65	0.8506	0.1111	26.75	0.6155	0.3075	27.06	0.6417	0.2501	27.68	0.7465	0.2069
DRAN		26.27	0.7800	0.2542	27.94	0.8229	0.1366	32.80	0.8756	0.1512	33.00	0.8785	0.0876	30.00	0.8393	0.1574
MMD		27.94	0.8589	0.1907	25.67	0.7745	0.1628	26.33	0.6137	0.2789	26.80	0.6347	0.2208	26.68	0.7205	0.2133
DAE		24.03	0.8055	0.2080	18.89	0.7720	0.1897	28.00	0.7492	0.2720	27.54	0.6510	0.2029	24.62	0.7444	0.2182
Proposed		**36.75**	**0.9659**	**0.0160**	**39.20**	**0.9804**	**0.0056**	**39.83**	**0.9757**	**0.0110**	**39.64**	**0.9793**	**0.0104**	**38.85**	**0.9753**	**0.0107**
AED	75	29.95	0.9000	0.1195	27.42	0.8853	0.1376	32.85	0.9105	0.0912	31.52	0.8938	0.0940	30.44	0.8974	0.1106
DnCNN		26.53	0.8146	0.2441	25.61	0.8176	0.1458	21.17	0.3179	0.4531	21.28	0.3096	0.4473	23.65	0.5649	0.3226
ResCNN		27.05	0.8392	0.2312	26.57	0.8261	0.1395	20.53	0.2831	0.4906	20.70	0.2860	0.4903	23.71	0.5586	0.3379
DRAN		23.89	0.6918	0.3738	27.32	0.8158	0.1595	31.03	0.7946	0.2494	31.75	0.8151	0.1388	28.50	0.7794	0.2304
MMD		26.09	0.8039	0.2760	25.21	0.7915	0.1700	21.13	0.3085	0.4490	21.29	0.3065	0.4303	23.43	0.5526	0.3313
DAE		22.92	0.7681	0.2595	18.49	0.7507	0.2143	27.55	0.7130	0.3309	27.09	0.6142	0.2350	24.01	0.7115	0.2599
Proposed		**36.57**	**0.9653**	**0.0164**	**38.66**	**0.9789**	**0.0063**	**38.97**	**0.9703**	**0.0139**	**38.79**	**0.9763**	**0.0124**	**38.25**	**0.9727**	**0.0122**
AED	Avg.	30.28	0.9100	0.1107	27.29	0.8806	0.1414	33.45	0.9252	0.0776	31.82	0.9020	0.0845	30.71	0.9044	0.1035
DnCNN		27.21	0.8148	0.2366	25.17	0.7630	0.1698	25.58	0.6292	0.2698	27.19	0.6960	0.2201	26.29	0.7258	0.2241
ResCNN		27.53	0.8308	0.2225	26.16	0.7998	0.1407	25.70	0.6052	0.2896	27.26	0.6609	0.2449	26.66	0.7242	0.2244
DRAN		28.59	0.8196	0.2003	31.00	0.8740	0.0945	33.34	0.8854	0.1286	34.19	0.9009	0.0776	31.78	0.8700	0.1252
MMD		27.98	0.8478	0.1943	25.69	0.7771	0.1569	25.08	0.5945	0.2818	27.20	0.6671	0.2204	26.49	0.7216	0.2133
DAE		23.93	0.8127	0.2112	18.89	0.7793	0.1880	28.57	0.7651	0.2340	27.46	0.6403	0.2139	24.71	0.7493	0.2118
Proposed		**36.87**	**0.9667**	**0.0150**	**39.87**	**0.9819**	**0.0050**	**40.54**	**0.9767**	**0.0098**	**40.70**	**0.9822**	**0.0089**	**39.50**	**0.9769**	**0.0097**

In contrast to the proposed method, existing MID models are inconsistent when confronted with diverse noise levels. Furthermore, their performance can vary depending on the imaging modality. For instance, DRAN demonstrates promising performance at low noise levels (i.e., $\sigma = 10, 25$). It dominates the existing MID models for such noise levels. However, AED outperformed DRAN at extreme noise levels (i.e., $\sigma = 50, 75$). These experimental results further confirm the limitations of existing MID methods under diverse noise patterns. By contrast, our proposed method can manage realistic diverse noise patterns, thus outperforming its counterparts.

4.1.3. Qualitative Evaluation

We extensively evaluated the MID models to quantify their strengths and weaknesses for numerous medical imaging modalities. Figure 8 shows a subjective comparison among MID models. As shown, existing image-to-image translation models tend to yield blurry images in Gaussian denoising. By contrast, the residual models demonstrate color discretion. In general, the existing models yield implausible images with visually disturbing artifacts. The proposed method addresses both limitations via an effective denoising network. Notably, the proposed DWR module and MHA-guided reconstruction enable the proposed model to yield sharp, clear, and plausible medical images. The proposed model

is superior for all compared modalities. It can denoise medical images without generating visual artifacts, even under a high noise proportion (i.e., $\sigma = 50$). A qualitative comparison confirmed the practicability of the proposed method for generic MID in real-world applications. Notably, the performance of the proposed method was consistent across numerous imaging modalities. This consistency indicates that the proposed method can be leveraged for any imaging modality, particularly those that incorporate noise patterns such as Gaussian noise.

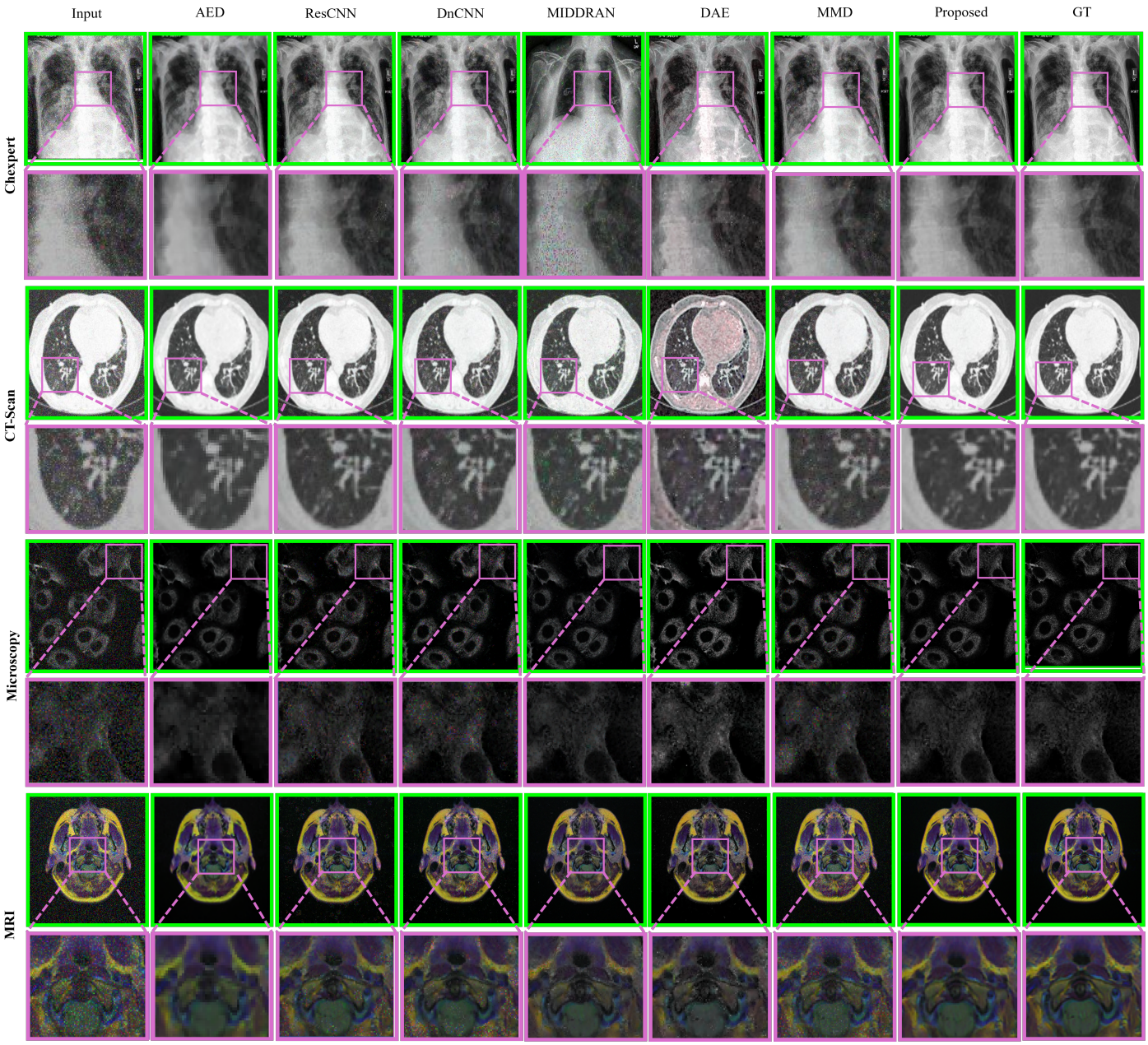


Figure 8. Comparison between deep medical image denoising methods. Existing denoising methods tend to yield smooth denoising results with visual artifacts. Proposed method can clean noisy medical images and address limitations of existing methods. Left to right: noisy input, AED [9], ResCNN [10], DnCNN [11], MIDDRAN [12], DAE [13], MMD [3], proposed method, and reference image.

4.2. Real-World MID

We evaluated our method using real-world noisy CT images [33,42,43] and synthesized noisy images. Therefore, the model was tuned to noisy real-world medical images. The proposed method was retrained by leveraging transfer learning with low-dose sharp

kernel CT images. We used 15,824 images from the dataset presented in [42] to tune the proposed method for real-world denoising. Notably, the training samples comprised sharp and soft kernel images with 1 and 3 mm capture settings. We regarded the full-dose images as the reference clean images and the quarter-dose images as noisy inputs, as suggested for previous methods. Additionally, we used 500 samples from each kernel (sharp and soft) and their subcategories (1 and 3 mm) to perform quantitative and qualitative comparisons.

Table 3 presents a quantitative comparison between images yielded by the proposed method and input low-dose images. The proposed method effectively denoised noisy images and significantly improved their quality. Across all the subcategories, the proposed method substantially improved the quality of the noisy images. In particular, the proposed method improved the low-dose noisy images in terms of the PSNR, SSIM, and LLIPS by 5.88 dB, 0.15, and 0.02, respectively. The proposed method not only performed denoising but also significantly improved the structural quality, as indicated by the metrics. In addition to structural improvements, the proposed method can improve the perceptual quality of noisy real-world images. The notable performance of the proposed method confirms its practical usage in widespread medical applications and diagnostic processes.

Table 3. Quantitative performance of proposed model for real-world noisy MID. Proposed method substantially improved quality of noisy real-world images.

Kernel	Wavelength	Method	PSNR↑	SSIM↑	LLIPS↓
Soft	1 mm	Input	36.31	0.8799	0.0802
		Proposed	**40.71**	**0.9543**	**0.0431**
	3 mm	Input	36.29	0.8832	0.0777
		Proposed	**40.80**	**0.9556**	**0.0414**
Sharp	1 mm	Input	28.53	0.6768	0.1342
		Proposed	**34.90**	**0.8462**	**0.1180**
	3 mm	Input	28.55	0.6751	0.1354
		Proposed	**34.77**	**0.8459**	**0.1175**
Combine	1 mm	Input	32.42	0.7783	0.1072
		Proposed	**37.81**	**0.9003**	**0.0806**
	3 mm	Input	32.42	0.7791	0.1066
		Proposed	**37.79**	**0.9007**	**0.0795**
Average	1 mm/3 mm	Input	30.47	0.7275	0.1207
		Proposed	**36.35**	**0.8732**	**0.0993**

In addition to a quantitative comparison, we performed a visual comparison, as shown in Figure 9. The proposed method proved to be superior for real-world MID. In particular, it yielded clearer and more visually plausible images than the inputs for all subcategories. In complex spatial regions, it maintained the salient information. Additionally, it generated cleaner images and ensured perceptual quality. The performance demonstrated by the proposed method in real-world MID confirms its applicability beyond synthetic datasets. The proposed method can be leveraged in real-world applications, including computer-aided diagnosis (CAD) applications [12], to shift medical imaging to a new paradigm.

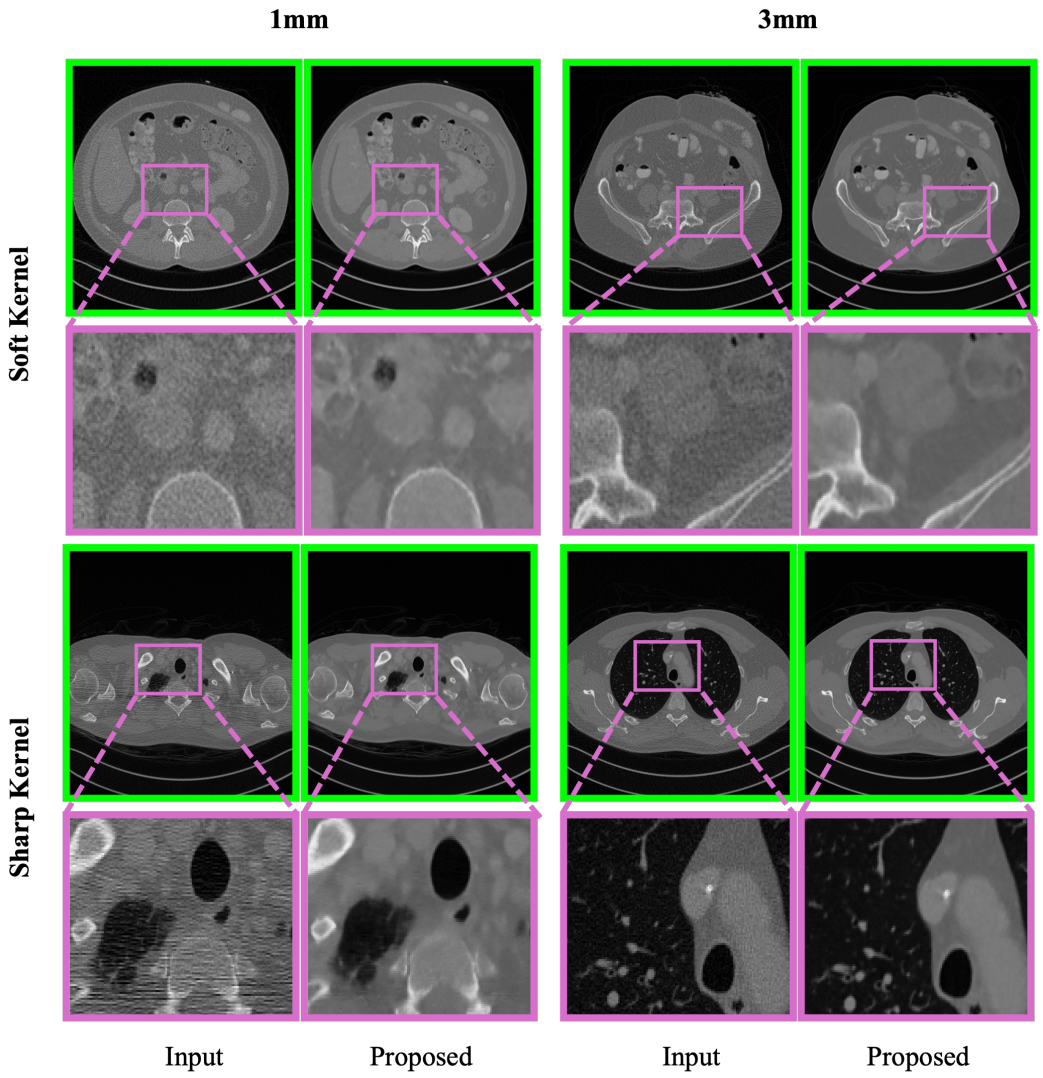


Figure 9. Performance of proposed method in real-world noisy MID. Proposed method can manage real-world noise. In each pair, left represents noisy input and right represents image denoised by proposed method.

4.3. Real-World Application

Medical image denoising has many real-world implications. A sophisticated denoising method can substantially improve the diagnosis process for medical experts by enhancing the medical images. In addition, the common noise in medical images can deteriorate the performance of computer-aided diagnosis systems such as segmentation, detection, etc. To further confirm the application of denoising in CAD application, we incorporated our proposed method to improve real-world noisy medical images. Further, we studied the state-of-the-art detection method (i.e., Yolo-V8 [44]) on red blood cells (RBCs), white blood cells (WBCs), and platelets in blood cell images. Table 4 illustrates the performance of Yolo-V8 on the RBC and WBC blood cell detection dataset [45]. It can be seen that the proposed method significantly improves the detection performance by reducing noise from the original images. Overall, the proposed method improves the performance of the Yolo-V8 (small) model in the RBC and WBC blood cell detection dataset. The performance improves by 0.06 for a mean average precision of 50.

Table 4. The proposed method was used to determine the performance of the Yolo-V8 (small) model on the RBC and WBC blood cell detection dataset on real-world noisy medical images and denoised images. The proposed method can significantly improve the detection accuracy of the Yolo-V8 model by reducing the noise that commonly contaminates medical images.

Input	Class	Box (Precision)	Recall	mAP (50)	mAP (50–95)
Original [45]	Platelets	0.8240	0.8150	0.8550	0.4620
	RBC	0.7480	0.7440	0.7870	0.5770
	WBC	0.9830	0.8840	0.9140	0.7880
	All	0.8510	0.8140	0.8520	0.6090
Enhanced	Platelets	0.8730	0.8380	0.9120	0.4720
	RBC	0.7490	0.8260	0.8590	0.6200
	WBC	0.9800	0.9830	0.9840	0.8290
	All	**0.8680**	**0.8820**	**0.9180**	**0.6400**

4.4. Inference Analysis

The results presented in Table 5 provide a comprehensive overview of the computational efficiency and performance of the proposed two-stage denoising method, which showcases its superiority over existing denoising techniques while remaining highly computationally efficient. Featuring only 12.54 million trainable parameters, the proposed method balances model complexity and computational overhead, thus rendering it a promising solution for practical deployment.

Table 5. The inference and parameter analysis of the proposed network. In addition to illustrating a significant performance improvement over existing methods, the proposed method is also computationally efficient. It comprises only 12.54 million trainable parameters and takes less than 10 ms to denoise a medical image on mid-level hardware.

Dimension	$128 \times 128 \times 3$	$256 \times 256 \times 3$	$512 \times 512 \times 3$
Flops (G)	17.42	69.68	278.74
Gmacs	16.22	64.90	259.59
Parameters (M)		**12.54**	
Inference Time (ms)	9.56	31.80	119.99

The fact that the proposed method has 12.54 million trainable parameters indicates its ability to process data efficiently without imposing an excessive computational burden. This efficiency translates into real-time performance, as evidenced by the mere 9.56 ms to denoise an input image measuring $128 \times 128 \times 3$ pixels. Notably, the proposed method is fully convolutional. Therefore, it does not incorporate pre- or post-processing. Consequently, the inference time is expected to remain constant for similar hardware (such as ours). However, network optimization techniques and more efficient hardware will allow the proposed network to operate faster and more efficiently for specific use cases.

These results underscore the feasibility of integrating the proposed method into real-world applications, where performance and computational efficiency are paramount. By offering a significant performance gain over existing denoising methods while maintaining computational frugality, the proposed approach is promising for various applications ranging from image processing in consumer electronics to medical imaging [12]. This balance between efficacy and efficiency renders the method a compelling option for addressing real-world denoising challenges.

4.5. Ablation Study

An ablation study was performed to evaluate the practicability of the novel component and the proposed MHA-guided reconstruction mechanism. Therefore, the proposed DWR

block was replaced with a vanilla residual block, and the MHA was removed from the proposed architecture. Table 6 presents the performance of the proposed network with and without its novel components. The base model incorporates the vanilla residual block without MHA in the decoder. The DWR variants of the proposed network incorporate a DWR block into the encoder and decoder. As shown, the proposed DWR block significantly improved the performance of deep networks with residual blocks. Additionally, the proposed MHA block in the decoder further enhanced the performance of the proposed model. Notably, the contributions of the proposed modules and mechanisms were independent of the imaging modalities. Consequently, the practicability of these modules and mechanisms, independent of the imaging type, can improve MID performance for any imaging modality.

Table 6. Quantitative evaluation of proposed components for different medical imaging modalities. Proposed modules substantially improved performance of proposed deep network, thus allowing proposed network perform consistently across numerous imaging modalities.

Model	σ	Chexpert			CT			MRI			Microscopy			Combined		
		PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓	PSNR↑	SSIM↑	LLIPS↓
Base	10	20.61	0.9125	0.0591	18.42	0.8761	0.0614	31.21	0.6852	0.1111	30.45	0.6284	0.1352	25.17	0.7756	0.0917
DWR		35.90	0.9588	0.0312	36.75	0.9643	0.0183	38.33	0.9433	0.0182	40.21	0.9416	0.0116	37.80	0.9520	0.0198
Proposed		**37.19**	**0.9685**	**0.0130**	**41.55**	**0.9856**	**0.0037**	**42.55**	**0.9819**	**0.0062**	**43.13**	**0.9892**	**0.0053**	**41.11**	**0.9813**	**0.0070**
Base	25	20.64	0.8230	0.2419	18.18	0.7610	0.1697	36.02	0.9168	0.0338	23.28	0.3492	0.4364	24.53	0.7125	0.2204
DWR		35.37	0.9524	0.0346	35.69	0.9544	0.0202	24.28	0.3925	0.3393	37.80	0.9036	0.0122	33.28	0.8007	0.1016
Proposed		**36.94**	**0.9670**	**0.0145**	**40.07**	**0.9825**	**0.0046**	**40.83**	**0.9787**	**0.0082**	**41.25**	**0.9841**	**0.0076**	**39.77**	**0.9781**	**0.0087**
Base	50	19.38	0.6432	0.5316	17.22	0.6318	0.3511	17.84	0.2048	0.6275	17.64	0.1963	0.7461	18.02	0.4190	0.5641
DWR		34.11	0.9431	0.0476	33.82	0.9399	0.0306	33.77	0.8714	0.0698	34.92	0.8505	0.0244	34.15	0.9012	0.0431
Proposed		**36.75**	**0.9659**	**0.0160**	**39.20**	**0.9804**	**0.0056**	**39.83**	**0.9757**	**0.0110**	**39.64**	**0.9793**	**0.0104**	**38.85**	**0.9753**	**0.0107**
Base	75	17.93	0.5380	0.6938	16.22	0.5622	0.4754	14.71	0.1395	0.7954	14.51	0.1344	0.8834	15.84	0.3435	0.7120
DWR		20.59	0.6200	0.5963	20.94	0.6804	0.3626	15.40	0.1494	0.7511	15.21	0.1455	0.8582	18.04	0.3988	0.6420
Proposed		**36.57**	**0.9653**	**0.0164**	**38.66**	**0.9789**	**0.0063**	**38.97**	**0.9703**	**0.0139**	**38.79**	**0.9763**	**0.0124**	**38.25**	**0.9727**	**0.0122**
Base	Avg.	19.64	0.7292	0.3816	17.51	0.7078	0.2644	24.95	0.4866	0.3919	21.47	0.3271	0.5503	20.89	0.5626	0.3971
DWR		31.49	0.8685	0.1774	31.80	0.8847	0.1079	27.94	0.5892	0.2946	32.03	0.7103	0.2266	30.82	0.7632	0.2016
Proposed		**36.87**	**0.9667**	**0.0150**	**39.87**	**0.9819**	**0.0050**	**40.54**	**0.9767**	**0.0098**	**40.70**	**0.9822**	**0.0089**	**39.50**	**0.9769**	**0.0097**

In addition to a quantitative evaluation, we visually compared the effects of the proposed components. Figure 10 shows the denoising comparison between the proposed network and its variants. As shown, the proposed DWR block facilitated the proposed method to mitigate imaging noise as compared with its vanilla counterpart. Additionally, the proposed MHA-guided reconstruction enabled the proposed method to leverage the salient features of the DWR module for reconstructing visually plausible images. In general, the proposed modules substantially improved the performance of the proposed network and addressed the limitations of conventional MID methods.

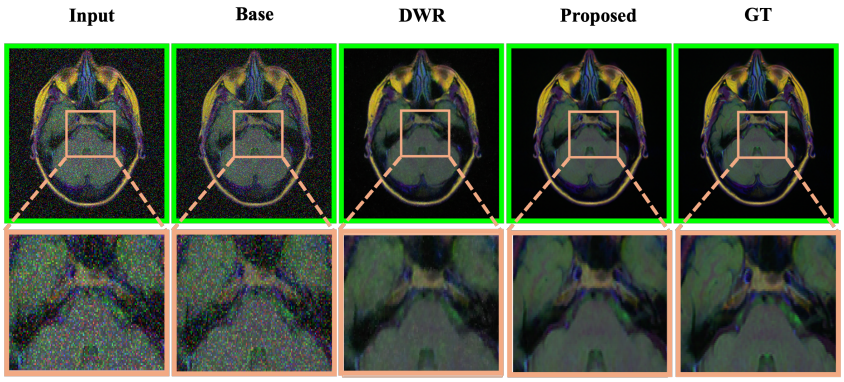


Figure 10. Ablation study on proposed network. Proposed DWR facilitates deep network to learn to mitigate noise by leveraging long-distance pixel dependencies. Proposed MHA block aims to reconstruct plausible, clean images by exploiting salient features extracted by proposed DWR module. From left to right, the Input image, base network (without DWR + MHA), DWR network (without MHA block), the proposed deep network (DWR + MHA), and the reference image.

4.6. Discussion

The proposed method reveals the limitations of existing MID models. It proposes a novel DWR module and a transformer attention module to achieve effective multimodal MID. Additionally, it demonstrates that a transformer-based attention module with a vanilla CNN can be extremely effective for multimodal MID. In addition to its significant improvements over conventional deep models, the proposed model is computationally effective. It comprises only 12.54 million trainable parameters and requires only 31 ms for denoising a medical image using mid-level hardware. Notably, the proposed model was utilized without incorporating any model optimization, such as quantization and pruning. Therefore, the inference speed of the proposed method can be substantially improved by leveraging optimization techniques.

In addition to learning and testing the proposed method on x64 architectures, it can be deployed on edge devices [46]. Such deployment and the leveraging of edge devices for MID can further advance MID research. An efficient MID method for edge platforms can significantly facilitate CAD applications [47]. Additionally, such an optimized network should allow edge vision developers to develop optimized and portable medical image enhancement devices [48]. The evaluation and optimization of the proposed method for edge devices would be an interesting research direction.

In addition to edge optimization, the proposed method focuses on the most common noise patterns in two-dimensional images. However, the proposed model can be adjusted to manage three-dimensional (3D) medical images, thus resulting in more sophisticated MID. Future studies are planned to apply the proposed method to 3D MID.

5. Conclusions

In this study, a novel MID method that leverages end-to-end deep learning to perform denoising in diverse medical imaging modalities was proposed. The proposed method incorporates an efficient residual block with dilation to capture long-distance pixel-wise dependencies and mitigate extreme noise from medical images. Additionally, it proposes utilizing MHA to leverage the salient features extracted by the proposed DWR module to obtain plausible images. The proposed method can denoise medical images without generating visual artifacts and can yield clean images that are similar to the reference images. The practicability of the proposed method was extensively evaluated using synthesized and noisy real-world medical images. The proposed method outperformed existing methods based on both quantitative and qualitative comparisons. Studies have been planned to investigate the practicability of the proposed method for 3D medical images and edge devices.

Author Contributions: Conceptualization, R.A.N. and A.H.; methodology, R.A.N., D.J., and S.-W.L.; validation, A.H. and H.S.K.; investigation, A.H. and H.S.K.; resources, D.J. and S.W.L; writing—original draft preparation, R.A.N., A.H., and S.-W.L.; writing—review and editing, H.S.K. and D.J.; supervision, D.J.; project administration, D.J., and S.-W.L.; funding acquisition, R.A.N. and H.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a National Research Foundation (NRF) grant funded by the Ministry of Science and ICT (MSIT), Republic of Korea, through the Development Research Program (NRF2022R1G1A1010226 and NRF2021R1I1A2059735).

Data Availability Statement: Data are contained within the article.

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

References

1. Lee, G.; Fujita, H. *Deep Learning in Medical Image Analysis: Challenges and Applications*; Springer: Cham, Switzerland, 2020; Volume 1213.
2. Kulathilake, K.S.H.; Abdullah, N.A.; Sabri, A.Q.M.; Bandara, A.R.; Lai, K.W. A review on self-adaptation approaches and techniques in medical image denoising algorithms. *Multimed. Tools Appl.* **2022**, *81*, 37591–37626. [CrossRef]

3. El-Shafai, W.; Mahmoud, A.; Ali, A.; El-Rabaie, E.; Taha, T.; Zahran, O.; El-Fishawy, A.; Soliman, N.; Alhussan, A.; Abd El-Samie, F. Deep cnn model for multimodal medical image denoising. *Comput. Mater. Contin* **2022**, *73*, 3795–3814. [CrossRef]
4. Wang, J.; Guo, Y.; Ying, Y.; Liu, Y.; Peng, Q. Fast non-local algorithm for image denoising. In Proceedings of the 2006 International Conference on Image Processing, Atlanta, GA, USA, 8–11 October 2006; pp. 1429–1432.
5. Elad, M.; Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* **2006**, *15*, 3736–3745. [CrossRef] [PubMed]
6. Arif, A.S.; Mansor, S.; Logeswaran, R. Combined bilateral and anisotropic-diffusion filters for medical image de-noising. In Proceedings of the 2011 IEEE Student Conference on Research and Development, Cyberjaya, Malaysia, 19–20 December 2011; IEEE: New York, NY, USA, 2011; pp. 420–424.
7. Bhonsle, D.; Chandra, V.; Sinha, G. Medical image denoising using bilateral filter. *Int. J. Image Graph. Signal Process.* **2012**, *4*, 36. [CrossRef]
8. Dabov, K.; Foi, A.; Katkovnik, V.; Egiazarian, K. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **2007**, *16*, 2080–2095. [CrossRef]
9. Gondara, L. Medical image denoising using convolutional denoising autoencoders. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 241–246.
10. Jifara, W.; Jiang, F.; Rho, S.; Cheng, M.; Liu, S. Medical image denoising using convolutional neural network: A residual learning approach. *J. Supercomput.* **2019**, *75*, 704–718. [CrossRef]
11. Jiang, D.; Dou, W.; Vosters, L.; Xu, X.; Sun, Y.; Tan, T. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. *Jpn. J. Radiol.* **2018**, *36*, 566–574. [CrossRef] [PubMed]
12. Sharif, S.; Naqvi, R.A.; Biswas, M. Learning medical image denoising with deep dynamic residual attention network. *Mathematics* **2020**, *8*, 2192. [CrossRef]
13. El-Shafai, W.; El-Nabi, S.A.; El-Rabaie, E.S.M.; Ali, A.M.; Soliman, N.F.; Algarni, A.D.; El-Samie, A.; Fathi, E. Efficient Deep-Learning-Based Autoencoder Denoising Approach for Medical Image Diagnosis. *Comput. Mater. Contin.* **2022**, *70*, 6107–6125. [CrossRef]
14. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17683–17693.
15. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
16. Suganyadevi, S.; Seethalakshmi, V.; Balasamy, K. A review on deep learning in medical image analysis. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 19–38. [CrossRef]
17. Patil, R.; Bhosale, S. Medical image denoising techniques: A review. *Int. J. Eng. Sci. Technol. (IJonEST)* **2022**, *4*, 21–33. [CrossRef]
18. Chen, H.; Zhang, Y.; Kalra, M.K.; Lin, F.; Chen, Y.; Liao, P.; Zhou, J.; Wang, G. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans. Med Imaging* **2017**, *36*, 2524–2535. [CrossRef]
19. Fan, F.; Shan, H.; Kalra, M.K.; Singh, R.; Qian, G.; Getzin, M.; Teng, Y.; Hahn, J.; Wang, G. Quadratic autoencoder (Q-AE) for low-dose CT denoising. *IEEE Trans. Med. Imaging* **2019**, *39*, 2035–2050. [CrossRef] [PubMed]
20. Hyun, C.M.; Kim, H.P.; Lee, S.M.; Lee, S.; Seo, J.K. Deep learning for undersampled MRI reconstruction. *Phys. Med. Biol.* **2018**, *63*, 135007. [CrossRef]
21. Kidoh, M.; Shinoda, K.; Kitajima, M.; Isogawa, K.; Nambu, M.; Uetani, H.; Morita, K.; Nakaura, T.; Tateishi, M.; Yamashita, Y.; et al. Deep learning based noise reduction for brain MR imaging: Tests on phantoms and healthy volunteers. *Magn. Reson. Med. Sci.* **2020**, *19*, 195. [CrossRef]
22. Rawat, S.; Rana, K.; Kumar, V. A novel complex-valued convolutional neural network for medical image denoising. *Biomed. Signal Process. Control* **2021**, *69*, 102859. [CrossRef]
23. Ghahremani, M.; Khateri, M.; Sierra, A.; Tohka, J. Adversarial distortion learning for medical image denoising. *arXiv* **2022**, arXiv:2204.14100.
24. Zhou, B.; Tsai, Y.J.; Chen, X.; Duncan, J.S.; Liu, C. MDPET: A unified motion correction and denoising adversarial network for low-dose gated PET. *IEEE Trans. Med. Imaging* **2021**, *40*, 3154–3164. [CrossRef] [PubMed]
25. Li, Y.; Zhang, K.; Shi, W.; Miao, Y.; Jiang, Z. A Novel Medical Image Denoising Method Based on Conditional Generative Adversarial Network. *Comput. Math. Methods Med.* **2021**, *2021*, 9974017. [CrossRef] [PubMed]
26. Chi, J.; Wu, C.; Yu, X.; Ji, P.; Chu, H. Single low-dose CT image denoising using a generative adversarial network with modified U-Net generator and multi-level discriminator. *IEEE Access* **2020**, *8*, 133470–133487. [CrossRef]
27. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef] [PubMed]
28. Kokil, P.; Sudharson, S. Despeckling of clinical ultrasound images using deep residual learning. *Comput. Methods Programs Biomed.* **2020**, *194*, 105477. [CrossRef]
29. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597.

30. Buda, M.; Saha, A.; Mazurowski, M.A. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Comput. Biol. Med.* **2019**, *109*, 218–225. [CrossRef]
31. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. Covid-ct-dataset: A ct scan dataset about covid-19. *arXiv* **2020**, arXiv:2003.13865.
32. Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; et al. Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* **2010**, *28*, 1248–1250. [CrossRef] [PubMed]
33. Sun, H.; Peng, L.; Zhang, H.; He, Y.; Cao, S.; Lu, L. Dynamic PET image denoising using deep image prior combined with regularization by denoising. *IEEE Access* **2021**, *9*, 52378–52392. [CrossRef]
34. Gao, F.; Wu, T.; Chu, X.; Yoon, H.; Xu, Y.; Patel, B. Deep Residual Inception Encoder–Decoder Network for Medical Imaging Synthesis. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 39–49. [CrossRef] [PubMed]
35. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
36. Sharif, S.; Naqvi, R.A.; Ali, F.; Biswas, M. DarkDeblur: Learning single-shot image deblurring in low-light condition. *Expert Syst. Appl.* **2023**, *222*, 119739. [CrossRef]
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
38. Kınlı, F.; Menteş, S.; Özcan, B.; Kıraç, F.; Timofte, R.; Zuo, Y.; Wang, Z.; Zhang, X.; Zhu, Y.; Li, C.; et al. AIM 2022 challenge on Instagram filter removal: Methods and results. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 27–43.
39. Sharif, S.; Naqvi, R.A.; Loh, W.K. Two-Stage Deep Denoising With Self-guided Noise Attention for Multimodal Medical Images. *IEEE Trans. Radiat. Plasma Med. Sci.* **2024**, *8*, 521–531. [CrossRef]
40. Pytorch. PyTorch Framework Code. 2016. Available online: https://pytorch.org/ (accessed on 24 April 2024).
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
42. McCollough, C. TU-FG-207A-04: Overview of the low dose CT grand challenge. *Med Phys.* **2016**, *43*, 3759–3760. [CrossRef]
43. Ma, Y.; Wei, B.; Feng, P.; He, P.; Guo, X.; Wang, G. Low-dose CT image denoising using a generative adversarial network with a hybrid loss function for noise learning. *IEEE Access* **2020**, *8*, 67519–67529. [CrossRef]
44. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 16 July 2024).
45. TFG. YOLO Dataset. 2022. Available online: https://universe.roboflow.com/tfg-2nmge/yolo-yejbs (accessed on 14 July 2024).
46. Sharif, S.; Mobin, I.; Mohammed, N. Augmented quick health. *Int. J. Comput. Appl.* **2016**, *134*, 1–6. [CrossRef]
47. Dong, G.; Ma, Y.; Basu, A. Feature-guided CNN for denoising images from portable ultrasound devices. *IEEE Access* **2021**, *9*, 28272–28281. [CrossRef]
48. Sakib, S.; Fouda, M.M.; Al-Mahdawi, M.; Mohsen, A.; Oogane, M.; Ando, Y.; Fadlullah, Z.M. Deep learning models for magnetic cardiography edge sensors implementing noise processing and diagnostics. *IEEE Access* **2021**, *10*, 2656–2668. [CrossRef]