MANet-CycleGAN: An Unsupervised LDCT Image Denoising Method Based on Channel Attention and Multi-Scale Features

Jinglong Tian
Minzu University of China
Beijing, China
22302022@muc.edu.cn

Tianze Zhao
Minzu University of China
Beijing, China
yuxhallin@163.com

Zhijun Fan
East China Jiaotong University
Jiangxi, China
fzjaimit@163.com

Linin Shen
Minzu University of China
Beijing, China
763944823@qq.com

Jieyao Wei
Minzu University of China
Beijing, China
weijieyao427@163.com

Qiumei Pu*
Minzu University of China
Beijing, China
puqiumei@muc.edu.cn

Abstract

Low-dose CT (LDCT) can significantly reduce health risks associated with radiation exposure compared to normal-dose CT (NDCT). However, the lower radiation dose may result in projection data being contaminated by noise, which can hinder the accurate identification of lesion details. Currently, most LDCT image denoising techniques employ supervised learning methods that rely on paired noisy and noise-free datasets for model training. In practical applications, however, obtaining such paired data is often challenging. To address this issue, we propose an unsupervised LDCT denoising method called MANet-CycleGAN, which can train a high-quality denoising model via unpaired data.Our design approach is as follows: 1. Eliminate the dependency of the denoising model training on paired data through a cyclic generative adversarial network architecture; 2. Apply the UNet architecture to generator for feature extraction and NDCT image generation, while using a PatchGAN discriminator to enhance the details of the generated images; 3. Introduce channel attention and multi-scale feature extraction capabilities through the Squeeze-and-Excitation (SE) module, Efficient Channel Attention (ECA), and Atrous Spatial Pyramid Pooling (ASPP) to improve image generation quality; 4. Utilize perceptual loss in training process to better preserve the structural features of the image while denoising. We conducted comparative experiments on the Mayo Clinic LDCT Grand Challenge dataset. The results demonstrate that the proposed method outperforms existing methods in both qualitative and quantitative aspects.

Keywords: Image Denoising, CycleGAN, Low-dose CT.

1. Introduction

Computed Tomography (CT) plays a crucial role in identifying subtle tissue abnormalities. In recent years, the application of CT technology has surged dramatically, particularly during the COVID-19 pandemic, when CT provided essential imaging assessments for the diagnosis and treatment of the disease [18]. Scanning the human body with high-dose X-ray beams can yield clear CT images, however, this radiation poses a risk of cellular and tissue damage, potentially adversely affecting human health [14]. LDCT significantly reduces harm compared to NDCT [16], but it typically results in lower image quality. The reconstructed images often contain substantial noise and artifacts, which can obscure lesion details and ultimately affect diagnostic accuracy [21]. Suppressing noise in LDCT images while preserving texture and pathological information is a highly challenging task, highlighting the significant research value of denoising methods for LDCT images.

Compared to NDCT, LDCT scans have a reduced radiation dose, resulting in a relatively lower number of photons detected by the detectors [12]. This decrease in photon count increases the likelihood of noise being introduced during the data acquisition process, leading to the appearance of white noise spots and blocky or waxy artifacts in the reconstructed images, which can adversely affect subsequent diagnoses. Traditional denoising methods used to enhance the clarity of LDCT images primarily fall into three categories: sine wave filtering, model-based statistical iterative reconstruction, and image domain denoising algorithms [17]. These denoising methods typically treat the noisy image as a simple superposition of a clean image and noise. These methods require a detailed estimation of the prior knowledge regarding the image and the noise distribution, resulting in a strong dependency on the noise model.

Consequently, such denoising algorithms often incur high computational costs and are prone to losing important image details.

Deep learning-based LDCT denoising algorithms do not require explicit modeling of noise, instead they can directly learn the mapping relationships between images through neural networks. These algorithms generate corresponding NDCT images from input LDCT images, offering stronger generalization capabilities compared to traditional methods. However, most current deep learning based image denoising approaches rely heavily on supervised learning, which requires the use of manually labeled data along with appropriate training tasks to optimize the parameters of the deep neural network. This necessitates a substantial amount of paired data consisting of noisy and noise-free images for model training. The primary challenge of applying supervised deep learning algorithms to LDCT image denoising is the high cost of acquiring medical images such as LDCT. Typically, only a single NDCT or LDCT image can be obtained for the same patient, making it difficult to gather sufficient paired image data for model training. To address this issue, recent studies often employ methods to artificially add noise to NDCT images in order to generate simulated LDCT images, thereby constructing paired datasets for model training.

In this study, we approach the image denoising problem from a novel perspective by utilizing an unsupervised learning method to eliminate the dependence on paired datasets for LDCT image denoising. We conceptualize the task of LDCT image denoising as an image translation problem, where the objective is to transfer images from one domain, A (the noisy image domain), to another domain, B (the noise-free image domain), without altering the primary content, such as the anatomical structures present in the CT images.

To implement this approach, we designed a denoising framework based on CycleGAN. We enhanced the UNet architecture, which serves as the generator, by integrating channel attention mechanisms and multi-scale feature extraction to improve the quality of the generated noise-free images. Additionally, we employed a patch-based discriminator to enhance image details. Throughout the training process, we utilized a perceptual loss mixed with other loss functions to better preserve the structural integrity of the images. As a result, we achieved precise LDCT denoising without the need for paired datasets. Our main contributions are summarized as follows:

- We propose a CycleGAN-based image denoising framework that facilitates the mutual transformation of LDCT images between the noisy image domain and the noise-free image domain through two cycles, while preserving the integrity of the image content to achieve effective denoising.
- UNet is employed as the image generator, utilizing its U-shaped architecture and skip connections to effectively preserve structural information in the images, thereby generating high-quality noise-free images.
- Based on the characteristics of convolutional neural networks, we introduce a channel attention mechanism into the backbone network to quantify the importance of different channels in the feature maps. This mechanism assigns appropriate weights to enhance significant features while suppressing redundant or ineffective ones.
- In the spatial dimension, we integrate features with different receptive fields to obtain more comprehensive multi-scale information, which is beneficial for addressing the various scales of anatomical structures present in CT images.
- Perceptual loss is applied during the training process of CycleGAN, combined with the pixel-wise L1 loss as hybrid loss function. This approach assesses the differences between the generated images and the real images at the patch scale, thereby enhancing the effectiveness of model training.

2. Related Works

With the advancement of deep learning in recent years, numerous studies have demonstrated that many deep learning-based denoising methods significantly outperform traditional approaches. The image denoising problem can be represented by the degradation model$X = Y - N$, where X denotes the noisefree clean image, Y and N represent the noisy image and additive Gaussian white noise with a standard deviation of σ respectively. Image denoising methods without deep learning often rely on prior knowledge to estimate the noise in the images, which typically requires manual parameter selection for noise modeling and complex optimization algorithms to achieve satisfactory denoising results[7, 34, 19, 3, 25]. The key point of these methods lies in obtaining a detailed estimate of the noise distribution, however, in most realworld scenarios, the noise distribution is unknown, and significant variations often exist between the noise distributions of different datasets, greatly limiting the denoising performance and generalization ability of the models.

Deep learning-based denoising framework have gained widespread application in the field of image denoising due to their exceptional performance. Unlike traditional methods, deep learning algorithms do not rely on manual noise modeling, instead, they directly utilize data and specific training tasks to optimize the parameters of neural networks for end-to-end image denoising[22], These algorithms can

effectively learn the mapping from noisy images to noise-free images, resulting in high generalization capabilities.

The process of deep learning-based image denoising can be summarized as follows: it first involves dimensionality reduction and feature extraction from the noisy images, followed by the generation of noise-free images using the extracted features. Both feature extraction and image generation are accomplished through neural networks, and a loss function is designed to minimize the difference between the generated noise-free images and the real noise-free images, with network parameters optimized using gradient descent.

Convolutional Neural Networks (CNNS) have been widely used in various image processing tasks due to their simple and efficient network structure. The inherent inductive biases of CNNs enable them to effectively extract image features, many studies have attempted to use CNNs for image denoising with success. RED-CNN[2] proposed a convolutional neural network that integrates an autoencoder network with a residual structure to map LDCT images to their corresponding NDCT images in an end-to-end manner, the network is composed entirely of convolutional layers, allowing it to theoretically handle images of any size. It optimizes network parameters using a mean squared error loss function to achieve image denoising. Subsequent research introduced batch normalization techniques and dilated convolutions to enhance denoising capabilities while reducing model parameters[36]. In addition to 2D images, a denoising algorithm based on 3D ResNet[9] models the spatial distribution of noise through 3D convolutions, thereby achieving denoising of 3D CT images. In recent years, new CNN based architectures have continuously been proposed and applied in the field of image denoising, aiming to improve denoising performance through improve the network structures[35, 6, 8].

Despite the success of CNN-based methods in various image denoising tasks, using convolutional neural networks to generate noise-free images often leads to over-smoothing, presenting significant challenges in preserving image edges and texture details. The UNet architecture was originally applied to image segmentation tasks[20], where an encoder encodes and reduces the dimensionality of the input image, followed by a decoder that generates the mask image. The encoder and decoder form a U-shaped structure, with skip connections that concatenate feature maps at the same depth, facilitating information transfer. The presence of skip connections allows the network to leverage the feature maps from shallow encoders to recover structural features that may have been degraded by down-sampling, thus producing higher-quality generated images.

Inspired by this, some studies have attempted to use UNet to convert noisy images into noise-free images while utilizing skip connections to preserve structural details. RatUNet [31] replaces the UNet convolutional blocks with residual blocks to avoid performance saturation, while also improving the upsampling method in the decoder and the skip connection structure to better recover image details, significantly enhancing image clarity in denoising tasks. FEUNet [27] employs UNet as an image generator, reducing network performance loss and accelerating training by generating a residual map between noisy and noise-free images. Many related studies utilize UNet as a generator to produce high-quality noise-free images, exploring improvements to the network's encoder, decoder, and skip connection structure to enhance overall performance.[32, 33, 15, 4].

Optimizing network parameters through supervised learning can yield good image denoising results, however, this process often requires a large amount of paired data, which is challenging to obtain for medical images like LDCT images. An alternative approach is to use unsupervised learning methods for network training, thereby eliminating the dependence on paired data. Generative Adversarial Networks (GANs) do not directly compute the difference between generated images and real images, instead, they rely on a discriminator to assess the authenticity of the images. This provides a solution for unsupervised image denoising. GAN[5]employs a generator and a discriminator that learn the distribution of the image domain through adversarial training to achieve image generation. Subsequently, WGAN[1]improved GANs by utilizing Wasserstein distance, enhancing the stability of the learning process and addressing the issue of mode collapse. A WGAN-based deep learning method enforced cycle consistency using Wasserstein distance to establish a nonlinear end-to-end mapping from noisy input images to noise-free output images, achieving semi-supervised image denoising[30].

CycleGAN[38] offers a method for image transformation that does not rely on paired data for model training, enabling images to be converted between two different domains without altering their content. Some studies have already applied CycleGAN in the image denoising field[23, 37, 29, 26]. However, issues such as training instability, model collapse, and poor image generation quality still persist. To address these problems, we propose a CycleGAN-based image denoising framework that aims to improve the denoising performance by improve the generator network and loss function, while also incorporating a patch-based discriminator for unsupervised image denoising.

3. Proposed Method

3.1. CycleGAN Based Denoising Model

To achieve unsupervised LDCT image denoising, we designed an image denoising framework based on CycleGAN, as shown in Fig. 1. Fig. 1-A illustrates the training process

of CycleGAN, where LDCT and NDCT refer to LDCT images and NDCT images, respectively, images generated by the model are indicated with an asterisk superscript. We treat image denoising as an image translation task, where the goal is to convert images from the noisy image domain A to the noise-free image domain B without altering the content of the images. To achieve this goal, we simultaneously trained two image generators G_{AB} and G_{BA} as the mapping functions from domain A to domain B and from domain B to domain A respectively. We set up two adversarial discriminators, D_A and D_B, to train these two generators. The discriminators are used to differentiate between real images and generated images in domains A and B respectively. In addition, to strengthen the constraints on the mapping functions, we introduced two cycle consistency losses. These losses ensure that images, after being transformed from one domain to another, can be mapped back to the original domain while remaining as consistent as possible with their initial states. The training process consists of two parts: the forward cycle and the backward cycle, The forward cycle is illustrated by the blue arrows in Fig. 1-A, A randomly selected LDCT image A is processed through the generator G_{AB} to produce a synthetic NDCT image B^*, adversarial loss L_{GAN} is calculated through discriminator, and simultaneously employ the identity mapping loss $L_{Identity}$ to constrain the content difference between the generated synthetic image and the original image. This ensures that the content of the image remains undistorted after the domain transformation. Finally, we use the generator G_{BA} to map the synthetic image B^* back to image domain A and get the cycle consistency loss L_{Cycle}. The backward cycle, illustrated by the yellow arrows in Fig. 1-A, is the inverse process of the forward cycle. A randomly selected NDCT image B is used as input to generate a synthetic image and calculate the aforementioned three losses. Finally, all the losses are summed to perform backpropagation and optimize the network parameters. The three losses, along with the final overall loss computed, are shown in Eq. 1 to Eq.4, In this context, we set two hyperparameters λ_{GAN} and $\lambda_{Identity}$ to adjust the proportions of the adversarial loss and the identity mapping loss, respectively. Through experiments, we found that set $\lambda_{GAN} = 10$ and $\lambda_{Identity} = 5$ yields better results. The function $Err(x, y)$ measures the difference between images x and y, and we chose to use the L1 loss function for this purpose.

$$L_{GAN} = E_{a\sim p(A)}[(D_B(G(a)) - 1)^2] + E_{b\sim p(B)}[(D_A(G(b)) - 1)^2] \quad (1)$$

$$L_{Identity} = E_{a\sim p(A)}Err(G_{AB}(a), a) + E_{b\sim p(B)}Err(G_{BA}(b), b) \quad (2)$$

$$L_{Cycle} = E_{a\sim p(A)}Err(G_{BA}(G_{AB}(a)), a) + E_{b\sim p(B)}Err(G_{AB}(G_{BA}(b)), b) \quad (3)$$

$$Loss = \lambda_{GAN}L_{GAN} + \lambda_{Identity}L_{Identity} + L_{Cycle} \quad (4)$$

To balance image generation quality and computational complexity, we adopted a classic UNet architecture as the backbone generator, which consists of both an encoder and a decoder. The encoder reduces the dimensionality of the image and extracts features through multiple convolutional layers and downsampling operations. The decoder transforms the extracted low-dimensional high-level semantic information through upsampling and convolution, gradually restoring it to the original size to generate the target image. During the upsampling process in the decoder, skip connections are used to integrate feature maps from the corresponding depths of the encoder, recovering information lost during downsampling and ensuring that the output image maintains complete structural content. We improved the UNet architecture by incorporating channel attention and multi-scale feature extraction to enhance network performance. The network structure and details are illustrated in Fig. 1-B to Fig. 1-E.

For the discriminator, we adopted the PatchGAN design approach, where the discriminator assesses image authenticity at the scale of image patches rather than the entire image. This allows the discriminator to capture finer textures and local details, thereby guiding the generator to produce higher-quality images. We employed a 70×70 PatchGAN discriminator, as shown in Fig. 2. The discriminator network is entirely composed of convolutional layers, allowing it to theoretically process images of any size and produce corresponding output at that size. By adjusting the convolutional kernel size and stride, we control the receptive field of the final output feature map to be 70×70 pixels. This is equivalent to using a sliding window of 70 pixels in width across the original image, generating a judgment result for each position. Compared to producing a single judgment result for the entire image, this approach effectively captures the content of different regions in the spatial dimensions, thereby obtaining more comprehensive information.

3.2. Channel Attention in Convnet

Traditional convolution and pooling processes struggle to effectively recognize the importance differences among various feature channels. The Squeeze-and-Excitation (SE) module enhances the network's expressive power by modeling the interdependencies between the feature channels produced by the convolutional layers. The fundamental principle of the SE module is to explicitly weight the different channels in the network's feature maps to distinguish their

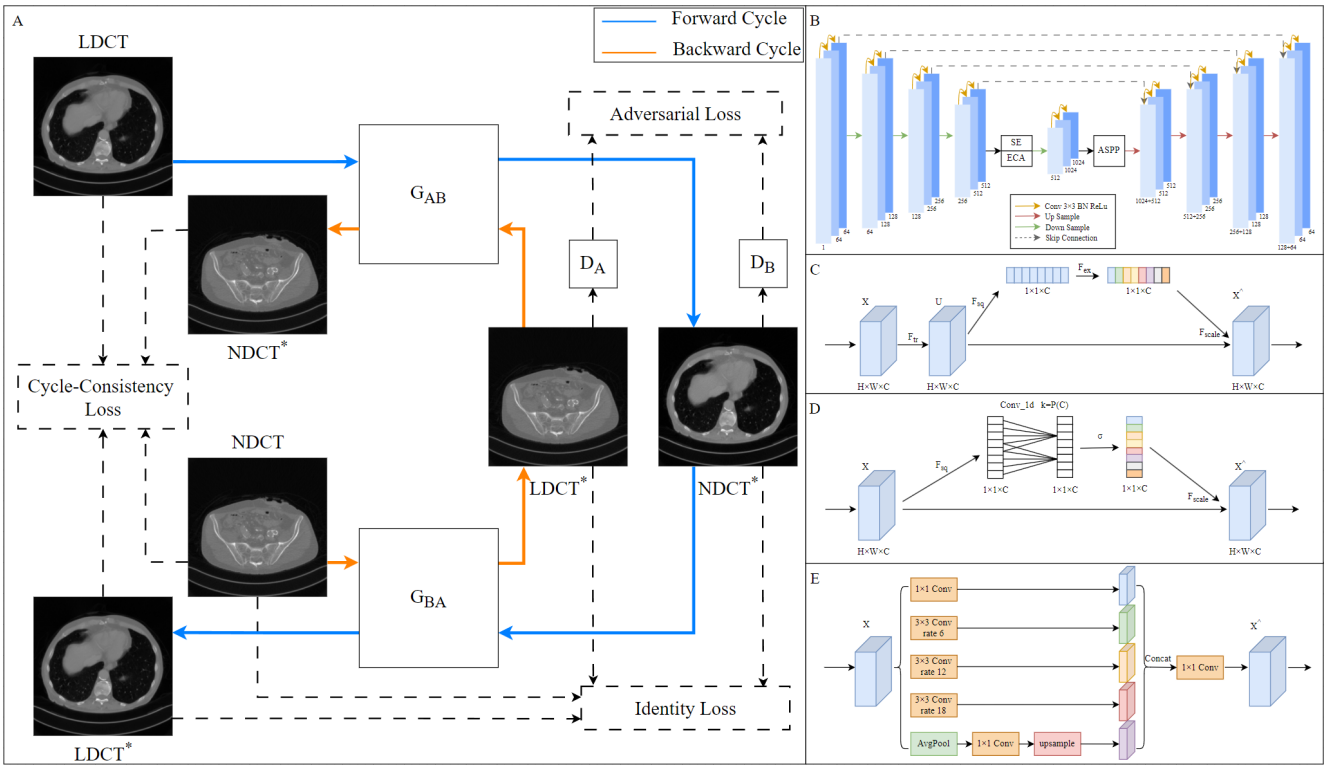

Figure 1. The overall structure of MANet-CycleGAN. A: CycleGAN based image denoising framework; B: The proposed MANet generator network structure, which uses a UNet as the backbone. SE, ECA, and ASPP modules are added at the end of the encoder and the beginning of the decoder to enhance image generation performance; C: Structure of the SE module; D: Structure of the ECA module; E: Structure of the ASPP module.

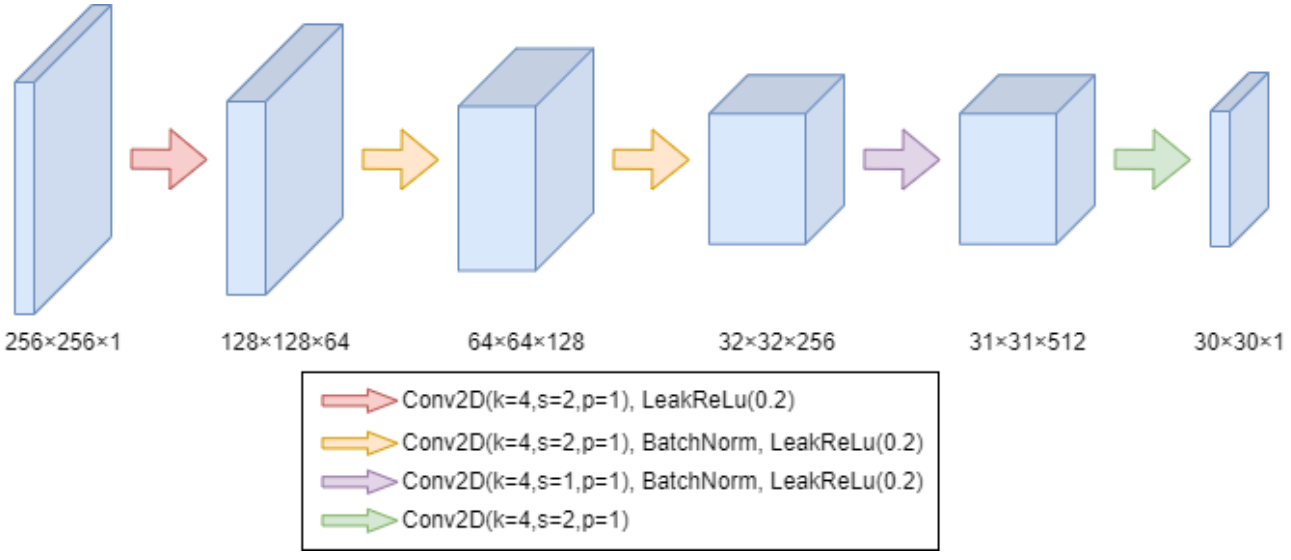

Figure 2. The structure of the PatchGAN discriminator

importance. This is achieved by adding an attention mechanism along the channel dimension, as shown in Fig. 1-C, F_{tr} represents a standard convolution operation. The key components of the SE module lie in the two steps: Squeeze and Excitation, F_{sq} represents the Squeeze step, which essentially performs a global average pooling operation (Eq. 5), By calculating the mean across the spatial dimensions of the feature map, each channel is compressed from the original feature map into a single scalar, resulting in a compact feature representation. F_{ex} represents the

Excitation step, which is used to accurately model the dependencies between feature channels. This is accomplished by concatenating two fully connected layers along with an activation function, as shown in Eq. 6, First, the first fully connected layer g transforms and reduces the dimensionality, followed by applying the activation function δ to introduce non-linearity, The activation function the ReLU is selected as δ, the output is then passed through a second fully connected layer f to restore the original dimensionality. This bottleneck structure design effectively reduces computational complexity. Finally, a Sigmoid activation function (denoted as σ in the figure) compresses the output values into the range $[0, 1]$ to serve as the channel attention scores. These scores are then applied to each channel through element-wise multiplication F_{scale} assigning the corresponding attention weights.

$$F_{sq}(X) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{ij} \tag{5}$$

$$F_{ex}(X) = \sigma(f(\delta(g(X)))) \tag{6}$$

The SE module adds attention weights along the channel dimension of the feature map, which enhances the model's response to important information in the input features. However, an analysis of its computation reveals two key issues. First, while the bottleneck structure of the two consecutive fully connected layers reduces computational complexity, it may lead to information loss, adversely affecting the precise prediction of channel attention. Second, using fully connected layers only captures global information along the channel dimension, resulting in a limited receptive field that can create redundant features. To address these issues, the ECA module replaces the fully connected layers of the SE module with a one-dimensional convolution of adaptive kernel size, as shown in Fig. 1-D, F_{sq} and F_{scale} remain the same as in the SE module, representing global average pooling and element-wise multiplication respectively. The function σ denotes the Softmax function, The channel attention weights are computed using a one-dimensional convolution, with the kernel size determined by the number of input feature map channels C. The calculation method is detailed in Eq. 7, the parameters γ and b are used to adjust the kernel size, where we set $\gamma = 2$, $b = 1$ in this study, $|\cdot|_{odd}$ indicates rounding up while ensuring that the kernel size k is an odd number.

$$P(C) = \left| \frac{log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{7}$$

3.3. Multi Scale Feature Extraction

In the process of image feature extraction using convolutional layers, different sizes of convolutional kernels can capture information at various scales in the spatial dimensions of the image. Larger convolutional kernels have a greater receptive field, allowing them to capture the macro structure of the image content, while smaller kernels have a relatively smaller receptive field, making them more effective for extracting fine details. The ASPP module integrates multi-scale information from the image, enabling the network to obtain a more comprehensive view for noise identification and removal while preserving the integrity of image details and structure. The structure of the ASPP module is shown in Fig. 1-E, It first employs multiple dilated convolutions with different dilation rates to obtain feature maps with varying receptive fields. Then, pointwise convolutions are used for feature fusion while adjusting the output channel count. The use of dilated convolutions allows the ASPP module to capture multi-scale information without introducing additional computational complexity, resulting in a more comprehensive feature representation.

3.4. Perceptual Loss

In the training phase of image translation tasks based on deep learning (such as image denoising, style transfer, and image super-resolution), the Mean Squared Error (L1) loss is often used to minimize the pixel-wise error between the input image and the target image. However, L1 loss can lead to blurry images and result in detail distortion or loss. Unlike L1 loss, which compares images on a pixel-by-pixel basis, perceptual loss first extracts high-level semantic information from images using a pre-trained convolutional neural network before calculating the error between the output vectors. Perceptual loss aligns more closely with human perception of image quality, placing greater emphasis on semantic information and being less sensitive to minor differences in pixel values. The calculation formulas for L1 loss and perceptual loss are given in Eq. 8 and Eq. 9, where N is the number of pixels in the image; $||\cdot||_F$ denotes the Frobenius norm; G represents the denoising network; ϕ is the pre-trained network used to extract image features, and dim is the dimension of the image feature vector encoded by ϕ.

During the training process of CycleGAN, we replaced the identity mapping loss $L_{Identity}$ with a mixed loss function that combines perceptual loss and L1 loss, ϕ in the perceptual loss refers to the pre-trained VGG16 network. The mixed loss function is formed by encoding images with the VGG network to obtain semantic features, which are then combined with pixel-wise L1 loss.

$$L_{L1}(G) = E_{(x,z)} \left[\frac{1}{N} ||G(z) - x||_1^2 \right] \tag{8}$$

$$L_{Perceptual}(G) = E_{(x,z)} \left[\frac{1}{dim} ||\phi(G(z)) - \phi(x)||_F^2 \right] \tag{9}$$

4. Experiments and results

4.1. Data and Details of Implementation

4.1.1 Dataset for experiments

This study evaluates the performance of the proposed LDCT image denoising method using the clinical CT dataset released by the Mayo Clinic for the "2016 NIH-AAPM-Mayo Clinic LDCT Grand Challenge" [13]. This dataset serves as a standard reference for assessing CT reconstruction and denoising techniques, covering X-ray projection images and reconstructed images of the head, chest, and abdomen from 10 anonymized patients. Each case includes paired NDCT images (NDCT) and simulated LDCT images at 25% of the normal dose. The images have a thickness of 3 mm and a resolution of 512×512 pixels.

4.1.2 Comparison Metric

For quantitative analysis, this study employs four evaluation metrics to assess the image denoising performance: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Gradient Magnitude Similarity Deviation (GMSD), and Root Mean Square Error (RMSE).

RMSE quantifies the difference between images by calculating the root mean square error on a pixel-by-pixel basis, making it the most direct method for comparing pixel value differences between two images. The calculation formula is shown in Eq. 10, where A and B represent the original noisy image and the denoised image, respectively, and M and N denote the height and width of the images. PSNR is commonly used to compare the differences between the denoised image and the original image to assess denoising effectiveness. A higher PSNR value indicates a greater ratio of retained information to suppressed noise in the denoised image, suggesting better denoising performance [24], as shown in Eq. 11.

$$RMSE = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(A_{ij}-B_{ij}\right)^2} \tag{10}$$

$$PSNR = 10 \times lg\left(\frac{\left(2^n-1\right)^2}{RMSE^2}\right) \tag{11}$$

SSIM is an important metric for assessing image similarity, with the calculation formula provided in Eq. 12. In this formula, μ_x and μ_y represent the mean values of the images, serving as estimates of image brightness, σ_x and σ_y denote the standard deviations of the images, providing estimates of image contrast, and σ_{xy} indicates the covariance between the two images, also related to their contrast. By considering the three factors of brightness, contrast, and structure, SSIM compares the differences between two images. A value closer to 1 indicates greater similarity between the images.

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{12}$$

The gradient magnitude of an image reflects the structural information of its content. GMSD (Gradient Magnitude Similarity Deviation) quantifies the perceptual quality of an image by analyzing the variations in pixel-level gradient magnitude similarity between the reference image and the denoised CT image, using standard deviation as a measure. This approach effectively captures changes in local image quality. The calculation formulas are provided in Eq. 13 and Eq. 14, where N represents the total number of pixels in the image, GMS denotes the gradient magnitude of the image (which can be computed using the Sobel operator), and $GMSM$ is the mean of the gradient magnitudes. The value of GMSD reflects the range of distortion severity in the image. The higher the GMSD score, the greater the distortion range, and the lower the perceived image quality. For more detailed information about this evaluation metric, refer to[28].

$$GMSM = \frac{1}{N}\sum_{i=1}^{N} GMS(i) \tag{13}$$

$$GMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(GMS\left(i\right) - GMSM\right)^2} \tag{14}$$

4.1.3 Experimental setup

We randomly selected 1,000 pairs of data from the Mayo abdominal CY dataset as the training set and 200 pairs as the validation set. The training data, consisting of NDCT and LDCT images, was shuffled to create unpaired dataset. To enhance training speed, we randomly cropped 256×256 pixel regions as input images during the network training phase, with a batch size set to 4, using the Adam optimizer. The validation set consists of paired NDCT and LDCT images, each with a resolution of 512×512pixels.

Due to the inconsistency between the loss function and image generation quality in GAN-based image generation methods, this study computes the RMSE between the network output and the NDCT images using the validation set after each training epoch. This serves as an evaluation metric to quantify the network's denoising performance. During the training phase, we employed an early stopping strategy to prevent overfitting, setting the patience of the early stopping mechanism to 5. This means that training will stop if the validation set RMSE does not decrease for 5 consecutive epochs, and the parameters with the minimum error will be retained as the optimal parameters.

4.2. Results

First, we validated the effectiveness of the improvements made to the generator network through ablation experiments. Using the same dataset, denoising framework, and experimental configuration, we tested different generator networks, all composed of UNet along with SE, ECA, and ASPP modules. We quantitatively assessed and compared the performance of each network using four evaluation metrics: PSNR, SSIM, GMSD, and RMSE. The experimental results are presented in Table 1. Ours indicates training MANet using a perceptual loss function to replace $L_{Identity}$, while the other methods employed L1 loss. From the table, it is evident that the generator using our improved network, MANet, performs better in denoising compared to the original UNet. Additionally, training with perceptual loss leads to further improvements across multiple evaluation metrics, with the most significant enhancement observed in PSNR. While RMSE directly reflects the differences between the generated denoised image and the true noise-free image by comparing them pixel-by-pixel, providing the most straightforward indication of their disparity, the other three metrics focus on the structural content of the images and the distribution characteristics of the noise. Our proposed MANet achieves the best performance in PSNR, SSIM and RMSE, indicating its effectiveness in preserving the structural integrity and details of the image content.

Through the visual results shown in Fig. 3, rows 1, 3, 5, and 7 clearly demonstrate that the images generated by MANet exhibit sharper edges and more detailed structures. Rows 2, 4, 6, and 8 present the residual heatmaps between the generated noise-free images and the true noise-free images, these heatmaps are obtained by taking the absolute difference between each image and its corresponding label image, followed by normalization. Pixels in the images that appear more red indicate greater differences from the label image at those locations, while pixels that are more blue signify smaller differences. It is important to note that due to normalization, the pixel values in the heatmaps only reflect relative differences within the images and do not indicate absolute differences in values between different residual maps. By comparing the residual heatmaps in columns D, E, and F, it is evident that using both the SE and ECA modules together yields better results. It is worth noting that when the SE module is used alone in the original UNet network, there is a slight improvement in the PSNR metric but a higher GMSD score. However, when both the SE module and ECA module are used together, it is possible to achieve a higher PSNR while reducing the score of GMSD. This demonstrates the effectiveness of adding channel attention mechanisms to the network. Furthermore, the results after adding the ECA module show that performance can be further improved by optimizing the channel attention mechanism.

The ASPP module utilizes convolutional kernels of different sizes for feature extraction, allowing it to effectively handle structures of various scales within the image. Experimental results indicate that adding the ASPP module significantly enhances performance, resulting in clearer edges and detailed structures in the generated images. However, when the ASPP module is used alone, noticeable noise appears in the edge regions of the images. This issue is effectively mitigated when the ASPP module is paired with the other two channel attention modules, as shown in columns G and H of row 8 in Fig. 3.

Generator	PSNR	SSIM	GMSD	RMSE
LDCT	26.932	0.967	0.026	12.169
U-Net	30.486	0.968	0.028	7.751
U-Net+SE	31.137	0.974	0.037	7.150
U-Net+ECA	31.612	0.979	0.027	6.951
U-Net+SE_ECA	31.734	0.985	**0.022**	6.939
U-Net+ASPP	32.076	0.980	0.023	6.506
MANet	32.990	0.985	0.026	5.875
Ours	**34.239**	**0.988**	0.023	**5.137**

Table 1. Comparison Study of the Generator

Ablation experiments demonstrate that using a mixed loss function composed of perceptual loss and L1 loss as the identity mapping loss during network training can achieve better denoising results. To further investigate the impact of the mixed loss function on network performance, we used two parameters, α and β, to adjust the ratio of perceptual loss to L1 loss, as shown in Eq.15. Based on this, we conducted two sets of comparative experiments: i) setting $\beta = 1$ and gradually increasing α from 0 to 1 in increments of 0.1; ii) setting $\alpha = 1$ and gradually increasing β from 0 to 1 in increments of 0.1. We trained the denoising network using different combinations of the identity mapping loss function and evaluated the network performance. The experimental results are shown in Fig.4.

We observed that using both perceptual loss and ontology mapping loss simultaneously for network training yields better denoising results compared to using either one alone. The denoising performance is optimal when $\alpha = 0.8$ and $\beta = 1$. The trend of the curves in the figure indicates that the L1 loss plays a dominant role during training, so changing the proportion of L1 loss alone leads to more unstable network performance. However, adding an appropriate amount of perceptual loss improves the denoising performance.

$$L_{Identity} = \alpha L_{Perceptual}(G) + \beta L_{L1}(G) \tag{15}$$

Additionally, we compared our proposed MANet-CycleGAN denoising framework with six advanced LDCT denoising methods, including two conventional CNN-based algorithms: UNet and RED-CNN, three GAN-based meth-

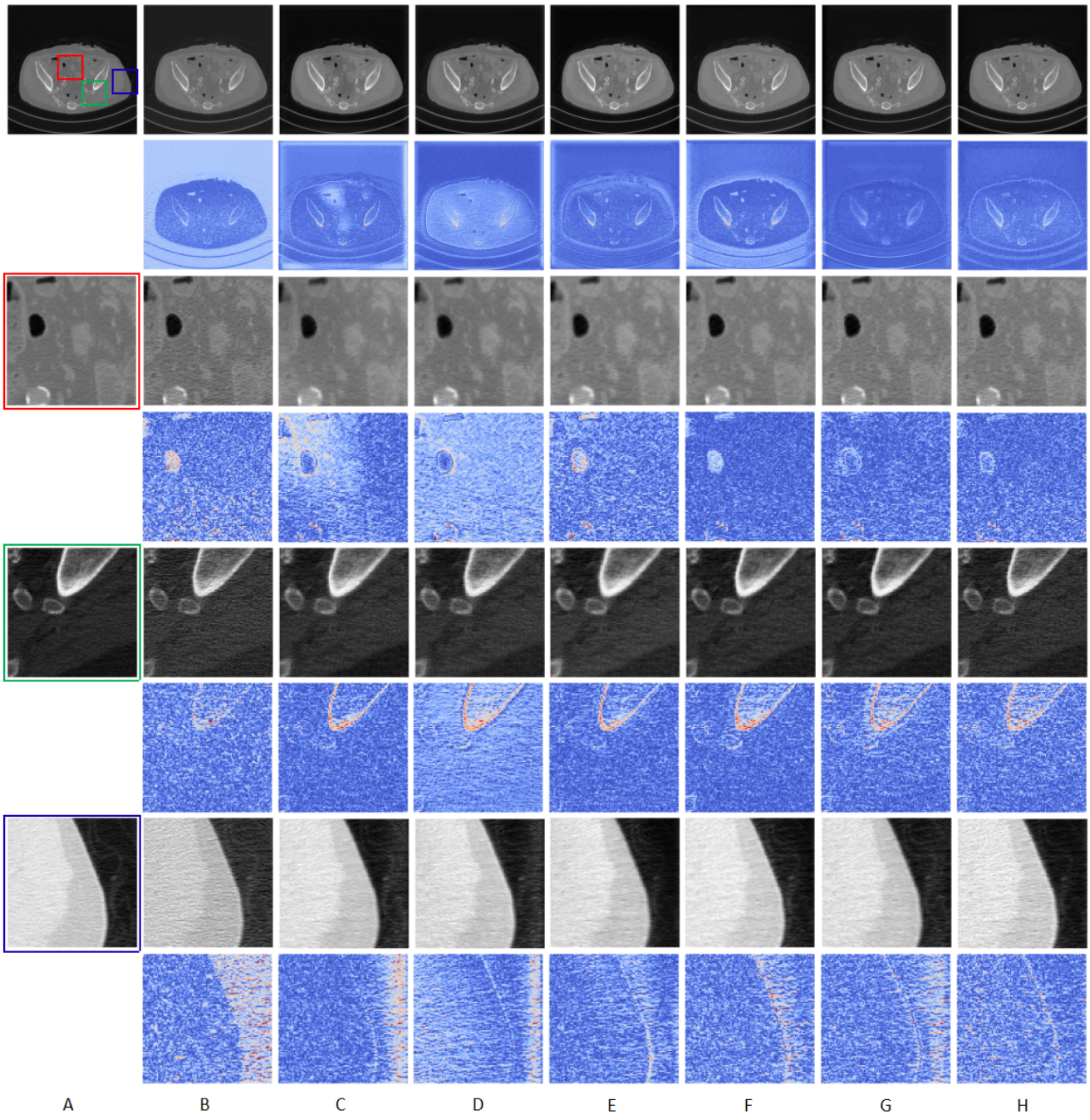

Figure 3. The visual comparison of denoising results for each model is presented as follows: Rows 1-2 show the CT images and the residual maps between the images and the noise-free images. Rows 3-4, 5-6, and 7-8 display the images for three selected ROI (Regions of Interest). Column A represents the NDCT images, column B represents the LDCT images, and columns C-H illustrate the denoising results using different models as encoders, C: UNet; D: UNet+SE; E: UNet+ECA; F: UNet+SE+ECA; G:UNet+ASPP; H: Our proposed MANetUNet, which combines UNet with all the aforementioned modules.

ods: WGAN, WGAN-VGG [11] and the original CycleGAN network, and a Diffusion-based method: Dn-Dp[10]. In this comparison, UNet, RED-CNN, and the two WGAN-based methods utilized their original network structures and employed paired data during training. The CycleGAN generator first extracted features using 9 residual blocks and then upsampled the images through transposed convolutions. The discriminator used the standard CycleGAN discriminator network and operated by unpaired dataset. Dn-Dp utilize two interconnected diffusion models: one for de-

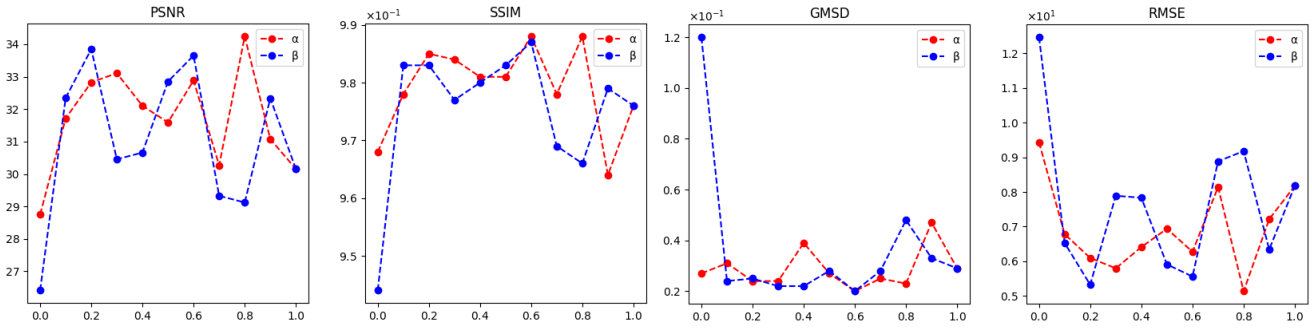

Figure 4. Analyze the impact of the proportion of perceptual loss and L1 loss in the change of the identity mapping loss function on network performance. The red and blue curves represent the effects of gradually increasing α with a fixed $\beta = 1$ and gradually increasing β with a fixed $\alpha = 1$, respectively, on denoising performance.r

noising low-resolution images and the other for converting low-resolution images to high-resolution images, The training can be accomplished using only NDCT images.

The results are shown in Table 2. Our proposed unpaired data denoising method achieved optimal results in PSNR, SSIM, and RMSE, even outperforming some algorithms trained with paired data. The improvement compared to the original CycleGAN algorithm, which also used unpaired data, was particularly significant, demonstrating the effectiveness of our proposed solution.

Model	PSNR	SSIM	GMSD	RMSE
LDCT	26.932	0.967	0.026	12.169
UNet	29.857	0.978	0.025	8.721
RED-CNN	32.332	0.985	**0.021**	6.532
WGAN	30.339	0.973	0.033	7.924
WGAN-VGG	31.244	0.976	0.024	7.083
CycleGAN	28.551	0.967	0.043	9.956
Dn-Dp	28.582	0.916	0.047	11.421
Ours	**34.239**	**0.988**	0.023	**5.137**

Table 2. Comparison of Various Denoising Methods

Model	PSNR	SSIM	GMSD	RMSE
LDCT	26.358	0.964	0.030	13.199
UNet	29.123	0.974	0.026	9.798
RED-CNN	31.452	0.983	0.024	7.493
WGAN	27.477	0.961	0.047	11.732
WGAN-VGG	30.417	0.976	0.031	8.342
CycleGAN	27.477	0.961	0.047	11.732
Dn-Dp	28.311	0.955	**0.021**	10.191
Ours	**32.798**	**0.985**	0.027	**6.470**

Table 3. Comparison of Various Denoising Methods (The first 200 pairs of images)

To further validate the model's generalization ability, we added more test data and conducted experiments on multiple different test sets. We tested each model using the first 200 and the last 200 pairs of NDCT and LDCT images,

Model	PSNR	SSIM	GMSD	RMSE
LDCT	27.725	0.964	0.024	11.367
UNet	30.292	0.975	0.026	8.415
RED-CNN	33.047	0.983	**0.019**	6.109
WGAN	32.150	0.970	0.027	6.523
WGAN-VGG	34.203	0.981	0.027	5.139
CycleGAN	29.384	0.954	0.045	9.012
Dn-Dp	29.644	0.917	0.042	10.478
Ours	**34.584**	**0.983**	0.024	**4.931**

Table 4. Comparison of Various Denoising Methods (The last 200 pairs of images)

which were not included in the training data, as test sets. The results are shown in Table 3 and Table 4. Overall, the supervised method training by paired data outperforms the method training by unpaired data. Our approach achieves the best results across different test sets and demonstrates relatively stable performance.

5. Conclusion

This study proposes an improved framework, MANet-CycleGAN, to address the denoising problem of LDCT images, successfully tackling the issue of training with unpaired data. We selected UNet as the backbone network for the GAN generator and enhanced it using a channel attention mechanism and an ASPP module that employs convolutional kernels of varying sizes for multi-scale feature extraction, thereby improving denoising performance.

For the GAN discriminator, inspired by PatchGAN, we calculated the loss function using image patches rather than mapping the entire image to a single value. Additionally, we employed a mixed loss function that combines perceptual loss from a pre-trained VGG network and L1 loss for training the network. Through ablation studies and comparative experiments, we demonstrated the effectiveness of our approach.

We found that even without paired data, satisfactory de-

noising results can be achieved, which is significant for scenarios where paired data is difficult to obtain. Training the model directly on unpaired real data, instead of relying on simulated paired data, enhances the model's denoising performance in practical applications. Our future work will continue to explore image denoising methods based on unpaired data to further improve model performance and training stability.

Resource availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Qiumei Pu (puqiumei@muc.edu.cn). All data reported in this paper will be shared by the lead contact upon request. The source code of our work can be found in: https://github.com/JinglingTian/CycleGAN_LDCT_Denoising.

Author contributions.

J.Tian. and L.Shen. designed this study. J.Tian., T.Zhao, L.Shen., and Q.Pu. analyzed and drafted the manuscript. J.Tian., T.Zhao and L.Shen. completed numerical experiments. T.Zhao., Z.Fan., J.Wang. and J.Wei. revised the manuscript. All authors were involved in explaining the concept and results of the data. All authors have reviewed and approved the final version of the manuscript.

Disclosure of Interests

All authors declare no competing interests.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, december 2017. *arXiv preprint arXiv:1701.07875*, 2017. 3

[2] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017. 3

[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2

[4] C.-M. Fan, T.-J. Liu, and K.-H. Liu. Sunet: Swin transformer unet for image denoising. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2333–2337. IEEE, 2022. 3

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[6] S. Herbreteau and C. Kervrann. Dct2net: An interpretable shallow cnn for image denoising. *IEEE Transactions on Image Processing*, 31:4292–4305, 2022. 3

[7] P. Li, J. Liang, M. Zhang, W. Fan, and G. Yu. Joint image denoising with gradient direction and edge-preserving regularization. *Pattern Recognition*, 125:108506, 2022. 2

[8] G. Liu, M. Dang, J. Liu, R. Xiang, Y. Tian, and N. Luo. True wide convolutional neural network for image denoising. *Information Sciences*, 610:171–184, 2022. 3

[9] J. Liu, Y. Zhang, Q. Zhao, T. Lv, W. Wu, N. Cai, G. Quan, W. Yang, Y. Chen, L. Luo, et al. Deep iterative reconstruction estimation (dire): approximate iterative reconstruction estimation for low dose ct imaging. *Physics in Medicine & Biology*, 64(13):135007, 2019. 3

[10] X. Liu, Y. Xie, C. Liu, J. Cheng, S. Diao, S. Tan, and X. Liang. Diffusion probabilistic priors for zero-shot low-dose ct image denoising. *Medical Physics*, 52(1):329–345, 2025. 9

[11] Y. Ma, B. Wei, P. Feng, P. He, X. Guo, and G. Wang. Low-dose ct image denoising using a generative adversarial network with a hybrid loss function for noise learning. *IEEE Access*, 8:67519–67529, 2020. 9

[12] I. Mastora, M. Remy-Jardin, C. Suess, C. Scherf, J.-P. Guillot, and J. Remy. Dose reduction in spiral ct angiography of thoracic outlet syndrome by anatomically adapted tube current modulation. *European Radiology*, 11:590–596, 2001. 1

[13] C. McCollough. Tu-fg-207a-04: overview of the low dose ct grand challenge. *Medical physics*, 43(6Part35):3759–3760, 2016. 7

[14] M. Mcdonnell, J. Das, D. O'Toole, A. Aldrahani, B. Verdon, M. Wilcox, J. Pearson, J. Lordan, A. De Soyza, N. Sousi, et al. Effects of gastro-oesophageal reflux and pulmonary micro-aspiration in bronchiectasis, 2018. 1

[15] D. Mehta, D. Padalia, K. Vora, and N. Mehendale. Mri image denoising using u-net and image processing techniques. In *2022 5th International Conference on Advances in Science and Technology (ICAST)*, pages 306–313. IEEE, 2022. 3

[16] D. P. Naidich, C. H. Marshall, C. Gribbin, R. S. Arams, and D. I. McCauley. Low-dose ct of the lungs: preliminary observations. *Radiology*, 175(3):729–731, 1990. 1

[17] Q. Pu, L. Shen, J. Tian, and J. Wei. Overview of deep learning-based denoising methods for low-dose ct images. *Chinese Journal of Stereology and Image Analysis*, (004):028, 2023. 1

[18] Z. Qinghua, F. Yaguang, Q. Youlin, Z. Guozhen, and S. Yan. Guidelines for low-dose ct screening of lung cancer in china (2023 edition). *Chinese journal of lung cancer*, 26(1):1–9, 2023. 1

[19] H. Rabbani. Image denoising in steerable pyramid domain based on a local laplace prior. *Pattern Recognition*, 42(9):2181–2193, 2009. 2

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[21] S. V. M. Sagheer and S. N. George. A review on medical image denoising algorithms. *Biomedical signal processing and control*, 61:102036, 2020. 1

[22] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020. 2

[23] T. Wang, Y. Lei, Z. Tian, X. Dong, Y. Liu, X. Jiang, W. J. Curran, T. Liu, H.-K. Shu, and X. Yang. Deep learning-based image quality improvement for low-dose computed tomography simulation in radiation therapy. *Journal of Medical Imaging*, 6(4):043504–043504, 2019. 3

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[25] Z. Wang and D. Zhang. Progressive switching median filter for the removal of impulse noise from highly corrupted images. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 46(1):78–80, 1999. 2

[26] S. Wu, C. Dong, and Y. Qiao. Blind image restoration based on cycle-consistent network. *IEEE Transactions on Multimedia*, 25:1111–1124, 2022. 3

[27] W. Wu, G. Lv, S. Liao, and Y. Zhang. Feunet: a flexible and effective u-shaped network for image denoising. *Signal, Image and Video Processing*, 17(5):2545–2553, 2023. 3

[28] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2013. 7

[29] Z. Yin, K. Xia, S. Wang, Z. He, J. Zhang, and B. Zu. Unpaired low-dose ct denoising via an improved cycle-consistent adversarial network with attention ensemble. *The Visual Computer*, 39(10):4423–4444, 2023. 3

[30] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE transactions on medical imaging*, 39(1):188–203, 2019. 3

[31] H. Zhang, Q. Lian, J. Zhao, Y. Wang, Y. Yang, and S. Feng. Ratunet: residual u-net based on attention mechanism for image denoising. *PeerJ Computer Science*, 8:e970, 2022. 3

[32] J. Zhang, Y. Niu, Z. Shangguan, W. Gong, and Y. Cheng. A novel denoising method for ct images based on u-net and multi-attention. *Computers in Biology and Medicine*, 152:106387, 2023. 3

[33] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, D.-P. Fan, R. Timofte, and L. V. Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023. 3

[34] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 2

[35] Q. Zhang, J. Xiao, C. Tian, J. Chun-Wei Lin, and S. Zhang. A robust deformed convolutional neural network (cnn) for image denoising. *CAAI Transactions on Intelligence Technology*, 8(2):331–342, 2023. 3

[36] Y. Zhang, J. Yang, and B. Yi. Improved residual autoencoder network for low-dose ct image denoising. *Journal of Shanghai Jiaotong University*, 53(8):7, 2019. 3

[37] S. Zhou, J. Yang, K. Konduri, J. Huang, L. Yu, and M. Jin. Spatiotemporal denoising of low-dose cardiac ct image sequences using recyclegan. *Biomedical physics & engineering express*, 9(6):065006, 2023. 3

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3