applied sciences

MDPI

Article

A Masked-Pre-Training-Based Fast Deep Image Prior Denoising Model

Shuichen Ji [1], Shaoping Xu [2,*], Qiangqiang Cheng [3], Nan Xiao [2], Changfei Zhou [2] and Minghai Xiong [2]

[1] School of Information Engineering, Nanchang University, Nanchang 330031, China; 5903121065@email.ncu.edu.cn
[2] School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China; 406100210085@email.ncu.edu.cn (N.X.); 409100220054@email.ncu.edu.cn (C.Z.); 419100220136@email.ncu.edu.cn (M.X.)
[3] School of Mechanical and Electronic Engineering, Gandong University, Fuzhou 344000, China; chengqq@nchu.edu.cn
* Correspondence: xushaoping@ncu.edu.cn

Abstract: Compared to supervised denoising models based on deep learning, the unsupervised Deep Image Prior (DIP) denoising approach offers greater flexibility and practicality by operating solely with the given noisy image. However, the random initialization of network input and network parameters in the DIP leads to a slow convergence during iterative training, affecting the execution efficiency heavily. To address this issue, we propose the Masked-Pre-training-Based Fast DIP (MPFDIP) Denoising Model in this paper. We enhance the classical Restormer framework by improving its Transformer core module and incorporating sampling, residual learning, and refinement techniques. This results in a fast network called FRformer (Fast Restormer). The FRformer model undergoes offline supervised training using the masked processing technique for pre-training. For a specific noisy image, the pre-trained FRformer network, with its learned parameters, replaces the UNet network used in the original DIP model. The online iterative training of the replaced model follows the DIP unsupervised training approach, utilizing multi-target images and an adaptive loss function. This strategy further improves the denoising effectiveness of the pre-trained model. Extensive experiments demonstrate that the MPFDIP model outperforms existing mainstream deep-learning-based denoising models in reducing Gaussian noise, mixed Gaussian–Poisson noise, and low-dose CT noise. It also significantly enhances the execution efficiency compared to the original DIP model. This improvement is mainly attributed to the FRformer network's initialization parameters obtained through masked pre-training, which exhibit strong generalization capabilities for various types and intensities of noise and already provide some denoising effect. Using them as initialization parameters greatly improves the convergence speed of unsupervised iterative training in the DIP. Additionally, the techniques of multi-target images and the adaptive loss function further enhance the denoising process.

Keywords: deep; image prior; execution efficiency; masked pre-training; network parameter initialization; iterative convergence rate; denoising effect

check for updates

Citation: Ji, S.; Xu, S.; Cheng, Q.; Xiao, N.; Zhou, C.; Xiong, M. A Masked-Pre-Training-Based Fast Deep Image Prior Denoising Model. *Appl. Sci.* **2024**, *14*, 5125. https://doi.org/10.3390/app14125125

Academic Editor: Douglas O'Shaughnessy

Received: 15 May 2024
Revised: 6 June 2024
Accepted: 7 June 2024
Published: 12 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Image noise has a negative impact not only on image quality but also on subsequent high-level visual tasks. These tasks encompass image segmentation [1,2], classification [3,4], super-resolution [5,6], and low-light enhancement [7,8]. Consequently, researchers have been engaged in addressing the longstanding challenge of image denoising in the field of image processing for decades. Traditional denoising algorithms, such as Block Matching and 3D Filtering (BM3D) [9], Non-Local Mean Filter (NLM) [10], Non-Local Centralized Sparse Representation (NCSR) [11], and Weighted Nuclear Norm Minimization (WNNM) [12], utilize a variety of approaches for noise reduction. Specifically, the key principle underlying

Non-Local Means (NLMs) is the identification of image patches similar to the area around the current pixel. NLM computes the similarities between these patches and the target area, and then leverages these similarity measures to weight the pixels in the patches, ultimately averaging them to produce the denoised pixel value. Spatial domain denoising algorithms like NLM determine new pixel values by analyzing the relationships among pixels within a certain window surrounding the central pixel. BM3D and other transformation domain denoising algorithms process the image by first applying transformations (such as Fourier or orthogonal transformations), then denoising in the transformed domain, and finally reconstructing the denoised image through an inverse transformation. These algorithms, which operate in the transformation domain, stand in contrast to spatial domain denoising approaches. Sparse representation-based denoising algorithms, exemplified by NCSR, capitalize on the observation that an undistorted image can be sparsely represented, whereas noise cannot. These methods endeavor to find an optimal dictionary to represent the denoised image, thereby effectively removing noise from the original data [13]. The image denoising algorithms based on low rank treat the image as a matrix and utilize the property that undistorted images are of low rank, unlike noise. Gu et al.'s WNNM denoising algorithm is based on this principle of exploiting the low-rank characteristics of images. However, these traditional methods heavily rely on inherent image information and manually crafted priors. They typically require complex optimization algorithms to achieve satisfactory results, thereby incurring substantial computational costs. In contrast, Deep Neural Networks (DNNs) over the past decade have exhibited significant advantages in image denoising due to their learning capabilities and nonlinear mapping abilities. The Denoising Convolutional Neural Network (DnCNN) [14], which is rooted in Convolutional Neural Networks (CNNs), marks a series of advancements. Through the integration of intricate network architectures, residual learning, and batch normalization, the DnCNN model notably boosts the denoising efficacy. Nevertheless, the DnCNN model exhibits certain constraints in addressing noise images with unknown noise levels, prompting the necessity for training multiple denoising models to ensure practical noise reduction effectiveness. To address this concern, Zhang et al. further refined the model, introducing a Fast and Flexible Denoising Convolutional Neural Network (FFDNet) [15]. This innovation, building upon the foundation of the DnCNN model, incorporates a noise level map as an additional input, thereby augmenting the model's adaptability to diverse noise levels. Additionally, the Dilated Residual UNet (DRUNet) denoising model [16] combines residual learning with the UNet structure and incorporates a noise level map. This approach improves flexibility in handling different noise levels and outperforms most CNN models regarding denoising efficiency. However, DRUNet requires extensive training time and its performance is limited under extremely low Signal-to-Noise Ratio (SNR) conditions and complex real-world noise interference. Recently, Transformer models have gained prominence in image denoising [17]. The Google research team utilized the self-attention mechanism of the Transformer architecture to address the limitations of CNNs in global information modeling, resulting in an enhanced performance. Based on this, Liang et al. introduced the SwinIR (Image Restoration Using Swin Transformer) denoising network [18]. SwinIR effectively learns the mapping from noisy images to clean images through the integration of shallow and deep feature extraction and advanced image reconstruction techniques. This network significantly improves the restoration of image details and reduces noise interference. Although SwinIR demonstrates superior denoising capabilities, it requires extensive training, particularly for extremely low-quality images. Furthermore, Zamir et al. introduced the Restormer (Restoration Transformer) denoising model [19], which has garnered attention due to its efficiency. This network with a U-shaped architecture employs the Multi-Dconv Head Transposed Attention (MDTA) and the Gated-Dconv Feed-forward Network (GDFN) to establish a novel Transformer module, enhancing performance. Restormer demonstrates high efficiency and advanced denoising capabilities, although it still demands substantial computational resources. In addition, the Uformer model [20] employed an encoder–decoder architecture to enhance the network's ability to capture local informa-

tion through localized strengthened window Transformer modules. Li et al. introduced the Efficient Wavelet Transformer (EWT) [21] model, which combines the advantages of CNNs and Transformers through a specially designed dual-stream feature extraction block. EWT achieves comprehensive processing of different levels of image information. While EWT demonstrates a desirable balance between model performance, size, execution time, and GPU memory consumption, it underperforms in preserving image texture fidelity compared to some mainstream denoising models like SwinIR, particularly when handling specific noise levels and datasets. Additionally, Yuan et al. introduced the HCFormer model [22], extending the application of Transformer-based denoising models to low-dose CT image denoising. HCFormer incorporates a Neighborhood Feature Enhancement (NEF) module, replacing conventional Multiple-Layer Perceptron (MLP) layers, to efficiently extract channel-level features, thereby demonstrating superior performance in low-dose CT image denoising tasks. Overall, while supervised deep-learning-based denoising models offer significant advantages, their training requires extensive noisy image–clean image pairs, which can be challenging to assemble [23]. Moreover, if the noise distribution in test images significantly deviates from that in the training set, the denoising performance may degrade due to data bias issues, highlighting a limitation in generalization.

In recent years, several methodologies have been developed to address the challenge of inadequate training datasets for supervised denoising models by utilizing noisy images alone, without the need for clean reference images (Ground Truth) during training. The self-supervised denoising model called Noise2Noise [24], introduced by Lehtinen et al. in 2018, represents a significant breakthrough in this field. Although Noise2Noise still requires two paired noisy images from the same scene for training, experimental results have shown that effective denoising is achievable solely through training with noisy images. Subsequently, Krull et al. introduced Noise2Void [25], which employs a blind-spot technique that allows training without paired noisy images, significantly reducing the data requirements for training denoising models. Similar methodologies are also employed within the Noise2Self [26] framework. Despite achieving commendable denoising results, these models necessitate training with noisy images that display notable correlation with the observed noisy data. This reliance poses practical challenges and escalates the expenses linked with data acquisition. Huang et al. developed Neighbor2Neighbor [27], which leverages noise image subsampling techniques to generate pairs of noisy images, enhancing the performance and reducing the influence of noise distribution on the denoising outcomes. However, the effectiveness of Neighbor2Neighbor depends on the availability of extensive noisy image datasets. In contrast, Quan et al. presented Self2Self [28], which employs Bernoulli sampling and local convolution techniques to train with just a single noisy image without any clean reference. However, the extended training duration of Self2Self limits its practical utility. Similarly, Ulyanov et al. introduced Deep Image Prior (DIP) [29], which utilizes the network architecture itself as implicit regularization. DIP relies solely on a single noisy image for online training, which doubles as the denoising process. This results in network parameters that are specifically tuned to the given noisy image, offering significant flexibility and practicality. Compared to Self2Self, DIP more effectively preserves image details and considerably reduces the time required for training (and denoising). However, the time DIP takes to complete the denoising process remains substantially longer than that required by supervised models. Generally, such unsupervised denoising approaches, which do not require reference images, exhibit stronger generalization capabilities and practicality compared to their supervised counterparts. Nevertheless, in scenarios where clean image–noisy image pairs are available, the denoising performance of unsupervised models, which still rely solely on noisy images, lags significantly behind that of supervised models.

While current deep-learning-based denoising models significantly outperform traditional techniques, both supervised and unsupervised approaches in this domain exhibit inherent limitations. Supervised models are typically limited to specific datasets, restricting their applicability across diverse scenarios. Conversely, unsupervised models, although less reliant on annotated data, still fall short regarding the denoising efficacy compared to

supervised ones [23]. In our previous works [30,31], we have significantly enhanced the DIP model's denoising capabilities. Nonetheless, the DIP model, known for its reliance on randomly initialized network parameters and network inputs during online training, boasts strong generalization capabilities but suffers from prolonged execution times due to excessive training iterations. This paper introduces a novel approach within the DIP framework, termed the Masked-Pre-training-Based Fast Deep Image Prior (MPFDIP) Denoising Model, aiming to preserve the DIP model's generalization capabilities and flexibility while reducing the number of iterations. Achieving this involves utilizing pre-trained model parameters to initialize the DIP model, thereby enhancing its execution efficiency. Specifically, the MPFDIP integrates a novel fast network model named FRformer (Fast Restormer). Built upon the core Transformer module from the established Restormer model, the FRformer incorporates techniques like sampling, residual learning, and refinement, replacing the traditional UNet backbone employed in the original DIP. Supported by masking technology, the pre-trained FRformer demonstrates a robust denoising performance across various noise types and intensities, considerably reducing both the number of model parameters and computational demand. By substituting DIP's UNet with FRformer, the convergence rate of unsupervised iterative training is accelerated, markedly improving the execution efficiency. Moreover, this study enhances the denoising performance further by refining the previously introduced multi-target image technology [30]. Extensive experimental evaluations confirm that the proposed MPFDIP model exhibits a superior denoising performance on synthetic and real-world noise, including Gaussian noise, mixed Gaussian–Poisson noise, and low-dose CT scans, surpassing existing mainstream supervised models and significantly outperforming the original DIP model in execution efficiency. The objective of the MPFDIP model is to achieve a higher denoising performance at the minimum time cost. The innovative contributions of this paper can be summarized as follows:

(1) Utilizing network parameters acquired from supervised pre-training to initialize the unsupervised DIP phases results in significantly accelerated convergence. This amalgamation of supervised and unsupervised learning enables the model to leverage prior knowledge during pre-training and fine-tune during online training, thereby yielding a superior performance across diverse image denoising scenarios.
(2) The integration of refined multi-target image techniques and adaptive loss functions further bolsters the denoising efficacy of MPFDIP. By leveraging refined multi-target image techniques, the model can adeptly capture and preserve crucial details while mitigating noise artifacts. Additionally, the adaptive loss functions facilitate the training process, enabling the model to concentrate on specific image areas necessitating denoising while upholding the overall image quality.
(3) The architectural design of the MPFDIP model demonstrates remarkable extensibility, facilitating the future incorporation of more advanced backbone networks to replace the proposed FRformer. This potential enhancement holds promise for further augmenting the denoising efficacy of the MPFDIP model.

2. Related Work

2.1. Deep Image Prior

To address the limited generalization capabilities in supervised denoising models due to biased data, the DIP denoising model effectively completes the denoising task using only a single noisy image. Leveraging the implicit feature extraction capabilities of deep learning and the prior knowledge embedded in the input noisy image, the DIP model transforms the denoising process into an optimization task focused on finding the optimal network parameters. The DIP model employs an encoder–decoder architecture similar to a U-net, where the input image undergoes multiple downsampling and upsampling operations for feature extraction. The input tensor z for the DIP model has the same spatial resolution as the output of the network, $f_\theta(z)$, which typically has a fully convolutional architecture. The spatial dimensions of the input are represented as $R^{C\times W\times H}$, where R represents the range of pixel intensity values, C indicates the number of input channels,

and W and H correspond to width and height, respectively. Additionally, the network uses skip connections to concatenate features of varying scales before feeding them into the next processing module. It primarily utilizes a residual learning approach, where a residual connection at the end adds the output back to the noisy input image to produce the final denoised image. In terms of the convolutional architecture, the DIP uses various kernel sizes to process the input images through the network, commonly employing sizes such as 3×3 or 5×5 depending on the specific layer and setup. This optimization is conducted through an online unsupervised training approach, where the process of optimizing network parameters can be described as follows:

$$(\hat{\theta}) = \arg\min_{\theta} E(f_\theta(z); y), \quad \hat{x} = f_\theta(z) \tag{1}$$

where $\hat{x}$ represents the denoised image, y denotes the noisy input image, θ denotes the network parameters, and z represents a randomly generated network input; the loss function E is defined as either the L1 or L2 distance between the network's output image, $f_\theta(z)$, and the noisy image y, which serves as the target image. The denoising process of the DIP model (i.e., the online training process) essentially establishes a complex nonlinear mapping between the random input tensor z and the target noisy image. Starting with random initialization, the model parameters are iteratively updated via a training method that minimizes the loss function, thereby generating a reconstructed image highly similar to the noisy image. As illustrated in Figure 1, the DIP model typically employs a UNet architecture as its backbone network. Theoretically, after infinite iterations of training, the model's output would converge towards the target image, failing to achieve the denoising task. However, the DIP model, trained in an unsupervised manner, exhibits characteristic noise repulsion. This trait optimizes the model's output (the reconstructed image) to the low-frequency signals of the noisy image. Consequently, by employing the early stopping technique, training is halted after a certain number of iterations, at which point the model's output is considered the denoising result.

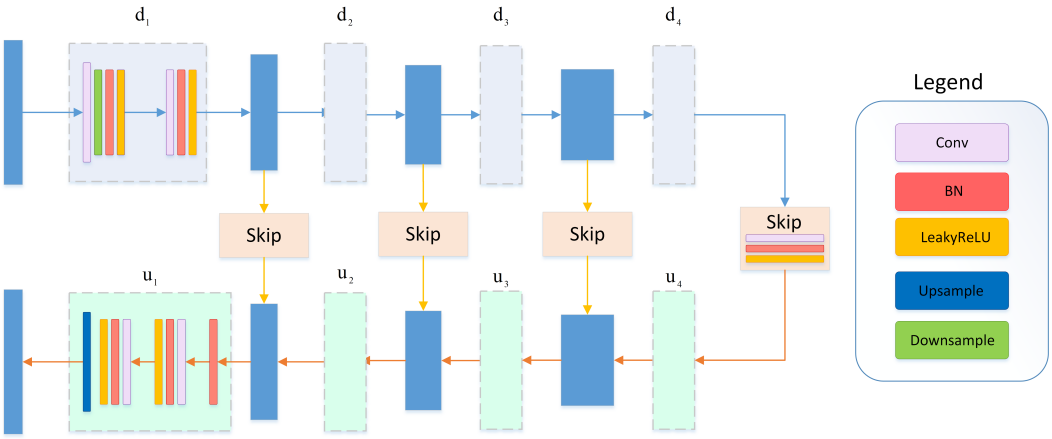

Figure 1. DIP network architecture. DIP is based on a decoder–encoder architecture, adding skip connections between corresponding layers of the encoder and decoder to minimize losses caused by upsampling and downsampling. In particular, the blue arrows indicate downsampling, the orange arrows represent upsampling, and the yellow arrows denote skip connections. In this case, d_i and u_i, respectively, represent the downsampling and upsampling operations performed at the i-th layer of the network. The network input is a randomly initialized tensor z, and the corresponding network output image is $f_\theta(z)$.

Compared to mainstream supervised denoising models, the DIP denoising model offers a key advantage with its unsupervised training mode, which allows for precise fine-tuning of the output image. This capability is particularly beneficial in scenarios with limited training data, showcasing a robust generalization ability. However, the

unsupervised online training process often leads to significantly longer execution times compared to supervised models. Several primary reasons contribute to this:

1. Random initialization of inputs and model parameters enhances the model's generalization capabilities. However, adjusting these parameters to adapt to specific noisy images and produce high-quality outputs necessitates numerous iterations, resulting in prolonged training durations.
2. The model's nonlinear mapping capability needs enhancement. During the iterative process, DIP employs a four-layer UNet network to achieve nonlinear mapping between the random input tensor z and the target (noisy) image. This approach's ability to capture global image information is inferior compared to Transformer-based backbone networks [20].
3. Using low-quality noisy images as target images leads to a broad search range for the network output image, resulting in suboptimal denoising outcomes and extended convergence times (for specific details, refer to ablation experiments).

Therefore, exploring backbone networks with stronger nonlinear mapping capabilities and effective techniques for network parameter initialization is crucial for enhancing the denoising performance of the DIP model.

2.2. Transformer Network Modules

Unlike traditional CNN architectures, Transformer network modules leverage global self-attention mechanisms and have been extensively utilized in various computer vision tasks in recent years due to their superior capability to capture long-distance pixel relationships ([32–34]). Here, we illustrate the Transformer module of the Restormer denoising model proposed by Zamir et al. This model's overarching network architecture closely resembles U-net, wherein the input image undergoes multiple downsampling and upsampling operations for feature extraction. Moreover, the network integrates skip connections to concatenate image features of varying scales before feeding them into the subsequent processing module. Primarily, the network adopts a residual learning strategy, where a residual connection at the network's terminus combines the network output image with the original noisy input image to generate the final denoised image. An overview of the overall structure is presented in Figure 2. Specifically, the model consists of two sub-modules, MDTA and the GDFN, as illustrated in Figure 3. In the MDTA module, the input data first undergo preprocessing through a normalization layer, followed by expansion via a point-wise convolution (PwConv) layer [35], which triples the number of channels. The primary function of the PwConv layer is to enhance the model's feature representation capacity, enabling the network to learn richer image features to tackle the complexities in image denoising. Subsequently, the input is processed by a depth-wise convolution (DwConv) layer [36], which groups data, effectively reducing the model's parameters and computational complexity, and generates the Q, K, and V matrices. The combination of point-wise and depth-wise convolution has been widely adopted in lightweight network designs such as MobileNet [37] and ShuffleNet [38], which demonstrates their capacity to enhance network performance while preserving model efficiency. After the depth-wise convolution process, the generated Q, K, and V matrices participate in the core self-attention mechanism computation of the Transformer:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK}{\alpha}\right)V \tag{2}$$

where SoftMax denotes the softmax function, and Q, K, and V represent the Query, Key, and Value matrices in the Transformer, respectively, with α being a learnable parameter of the network.

In the GDFN, the input first passes through a normalization layer and is subsequently expanded to four times the number of original channels via a PwConv layer. The data are then processed by a DwConv layer, generating matrices x_1 and x_2 with a reduced

computational cost. Subsequently, x_1 undergoes processing via the Gaussian Error Linear Unit (GELU) activation function, is element-wise multiplied by x_2, and then passed through another PwConv layer to restore the channel count to its original size. The output from this processing is then combined with the original input to produce the final output image. The processing can be represented as follows:

$$x_1 = \text{Dwconv1}(\text{Pwconv1}(x)), \quad x_2 = \text{Dwconv2}(\text{Pwconv2}(x)) \tag{3}$$

$$\text{GDFN}(x) = \phi(x_1) \otimes x_2 \tag{4}$$

where ϕ denotes the GELU activation function, PwConv represents point-wise convolution operations, and DwConv signifies depth-wise convolution operations. The use of both point-wise and depth-wise convolutions enhances the network's nonlinear mapping capabilities, strengthening its ability to handle complex scenarios. It should be noted that despite the computational efficiency optimizations in the core Transformer module of the Restormer denoising model, it still involves numerous matrix operations, and there remains room for improvement in computational efficiency.

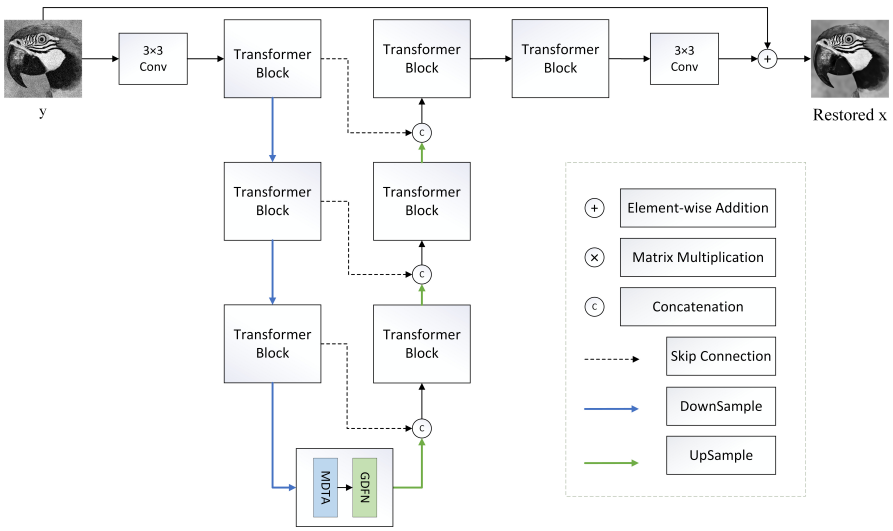

Figure 2. The framework of the Restormer denoising model. The Restormer showcases a multiscale hierarchical design integrating efficient Transformer blocks, meticulously crafted to maximize performance. The pivotal constituents of these Transformer blocks are MDTA, which fosters enriched query–key feature interactions across channels instead of spatial dimensions, thereby augmenting the depth of feature processing, and GDFN, which orchestrates controlled feature transformation, enabling seamless propagation of crucial information through the network layers.

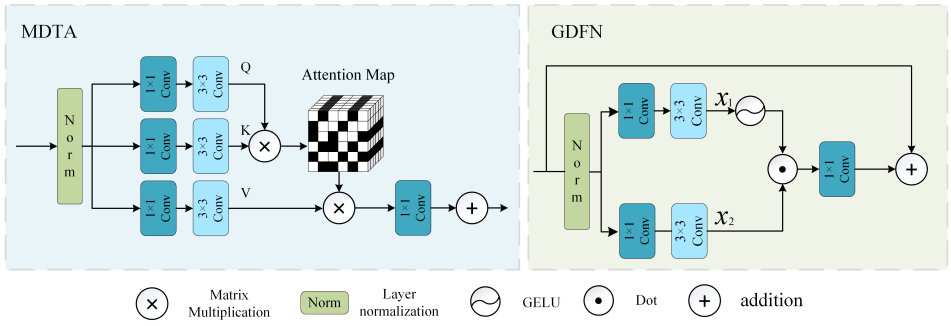

Figure 3. Structure of the core transformer module in the Restormer denoising model. The MDTA enables query–key feature interactions primarily across channels rather than in spatial dimensions. The GDFN enables controlled feature transformations, facilitating the further propagation of useful information.

3. Methodology

3.1. Basic Idea

As mentioned above, while the unsupervised DIP model exhibits advantages in the field of image denoising by offering a highly flexible (data-bias-free) denoising solution, its effectiveness and convergence speed require further enhancement. To address these limitations, this paper introduces three significant enhancements aimed at boosting the execution efficiency and enhancing the denoising performance of the model:

1. Development of a network structure with enhanced nonlinear mapping capabilities. To enhance the denoising performance, the DIP model's UNet backbone is replaced with the FRformer network, centered around a core Transformer module. The FRformer network utilizes an efficient attention mechanism and adopts a topology akin to the SwinIR model. This enhancement not only improves nonlinear mapping capabilities but also significantly reduces the network's parameter count, ensuring a higher execution efficiency.
2. Improvement in the DIP network parameter initialization. To address slow convergence due to random parameter initialization in the DIP, a supervised training approach initially maps noisy images to clean images to derive the necessary parameters. These parameters are subsequently used as the initial parameters for the unsupervised training of the DIP model, changing the input from a tensor z to a noisy image, facilitating a seamless transition between the pre-training and online training phases. To ensure the general applicability of the pre-trained network parameters, masking training techniques are employed so that the parameters are not limited to a specific noise type and intensity (ensuring robustness), thereby accelerating the convergence speed of the DIP's unsupervised training.
3. Addition of high-quality preprocessed images as target images. Several mainstream denoising models with better complementarity are used to process a given noisy image to obtain high-quality denoised images (i.e., preprocessed images) as the target images for the DIP model. This approach narrows the search range for the network's output image under the unsupervised training mode of the DIP model, thus ensuring better denoising results and enhancing the convergence speed.

3.2. Framework

As illustrated in Figure 4, the implementation of the MPFDIP model is structured into two phases: masked pre-training and DIP unsupervised training (fine-tuning). During the pre-training phase, the model leverages the proposed FRformer network, which possesses enhanced nonlinear mapping capabilities, to initially learn the mapping relationship from noisy images to clean images via a masking training strategy. In the DIP unsupervised training phase, a selection of denoising methods from mainstream complementary deep-learning-based denoising models is applied to the given noisy image, generating denoised images $\{\text{Denoiser}_i(y) \mid i \in \{1, 2, \ldots, n\}\}$ (preprocessed images), which serve as multi-target images. Here, the noisy image remains the target image. With the noisy image as input, an adaptive loss function [29] is constructed based on the network output image, several complementary preprocessed images, and the noisy image itself [31]. Under the constraints of this loss function, the DIP unsupervised iterative training method and early stopping techniques are utilized to conduct online training (denoising).

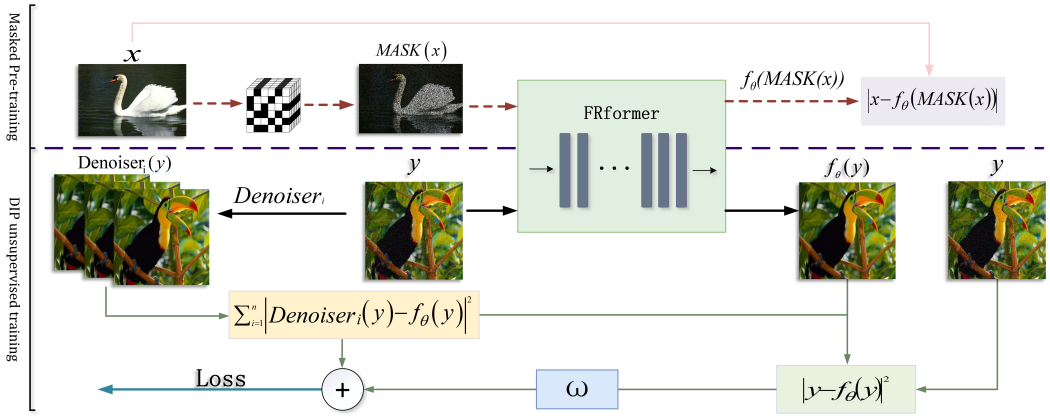

Figure 4. The flowchart of the framework of MPFDIP model. In the pre-training phase, the FRormer network is pre-trained using masking techniques. During the unsupervised online training phase, noisy images serve as network input, with pre-trained model parameters loaded as initial parameters for the unsupervised training phase of the DIP. This involves constructing multi-target images by utilizing both preprocessed and noisy images, and developing an adaptive loss function to optimize training outcomes.

3.3. FRformer Backbone Network

The proposed FRformer aims to address the high computational complexity inherent in Transformer-based denoising models for image denoising tasks. Illustrated in Figure 5, the network consists primarily of three parts: input processing, feature extraction, and image reconstruction. This structure mirrors the topology of the SwinIR model while diverging from the U-shaped topology of the Restormer model, leading to significant reductions in model parameters. During the input processing stage, the input image initially undergoes a random masking module, then is downsized by a downsampling module to half its original size, thereby reducing computational demands. Subsequently, the image channels are expanded via a convolution layer, employing the Pixel-unshuffle downsampling technique [39], which effectively minimizes the model's parameter count. The feature extraction part mainly comprises multiple Residual Transformer Convolutional Blocks (RTCBs) and a convolution layer. Each RTCB includes several basic Transformer layers (BTLs) and a convolution layer, integrating a residual learning structure. In the image reconstruction segment, feature maps are refined by the Feature Refinement Module (FRM), followed by a convolution layer that adjusts the channel count. A Pixel-shuffle upsampling module then restores the image to its original size, combining the output with the damaged image from the masking module to produce the final image. The design of the FRformer network topology optimizes network performance while significantly reducing reliance on computational resources, achieving a balance between efficient performance and resource consumption. In the two-phase execution sequence of the MPFDIP denoising model proposed in this paper, the FRformer network operates in both supervised and unsupervised modes.

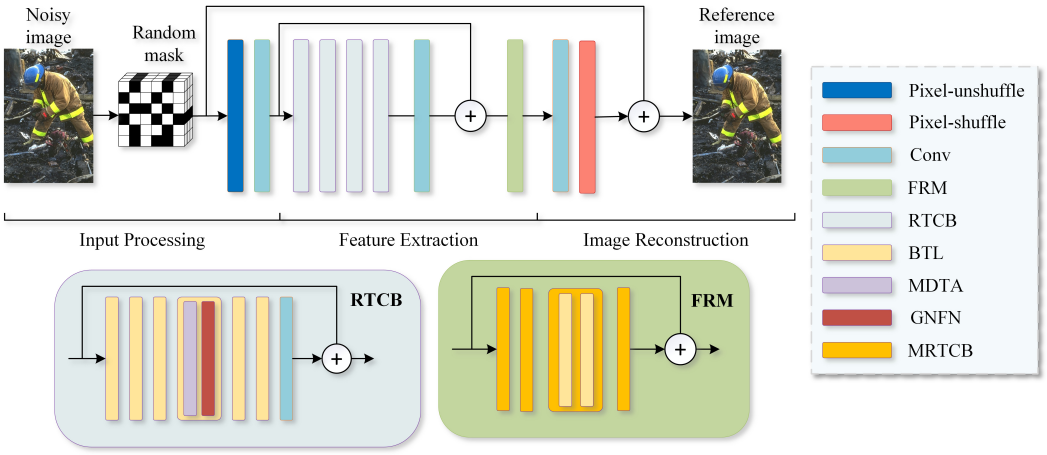

Figure 5. FRformer network structure. The FRformer network is divided into three parts: input processing, feature extraction, and image reconstruction. The RTCB is the core of feature extraction, combining six BTLs and a convolution layer using a residual learning structure. Each BTL consists of MDTA and a GDFN, while the FRM comprises four mini-residual Transformer convolutional blocks (MRTCBs).

3.4. Masked Pre-Training

Supervised deep-learning-based denoising models often struggle with generalization capabilities, particularly when trained on specific noise distributions such as Gaussian noise, leading to suboptimal performance on other noise types. Recently, Masked Image Modeling (MIM) [40] has been introduced as an innovative strategy to enhance the network performance. The central concept involves training the model to predict pixels missing due to masking operations, effectively generating noisy images. This approach leverages the intrinsic structure of the images to enhance the model's representational capacity. In this work, we adopt a supervised masked pre-training approach to improve the generalization capability of the FRformer backbone network model. This allows the pre-trained network parameters to adapt to specific noisy images with minimal iterative updates, addressing the slow convergence issue associated with random parameter initialization in the DIP model. Specifically, in the pre-training phase, the masked noisy image serves as the network input, with the corresponding clean image serving as the target image. Training is conducted using the L1 loss function, also known as Least Absolute Deviation (LAD):

$$\text{L1 Loss} = |x - f_\theta(\text{MASK}(x))| \tag{5}$$

where x is the clean image, θ are the network parameters, and MASK represents the random masking process.

Compared to traditional supervised deep learning denoising models, the masked training strategy significantly enhances the generalization capability of the denoising model when handling different types and intensities of noise. The network model parameters pre-trained with masking are already capable of mapping noisy images to clean images. Utilizing these pre-trained parameters to initialize the DIP model provides a robust foundation for its subsequent unsupervised training phase, thereby significantly enhancing both its denoising efficacy and execution efficiency. Relevant ablation experiments are discussed in Section 4.

3.5. DIP Unsupervised Training with Multi-Target Images

Although single supervised denoising algorithms, hindered by data bias, do not consistently outperform other methods across all application scenarios, they do exhibit complementary characteristics among each other. In the DIP model, the search space for the network output is constrained, and the adjustment of network parameters is directed by the target image [41]. The valuable information contained in noisy images and its

complementarity to preprocessed images have been revealed in the study conducted in [30]. Consequently, in this work, we adopt a multi-target image strategy, which entails using high-quality preprocessed images alongside retaining the noisy image as a target image. This approach effectively constrains the search space of the network output image, aiming to guarantee the quality of the final network output image. In the unsupervised training phase, the FRformer backbone network undergoes training using the L2 loss function, also referred to as the Least Squares Error (LSE):

$$\text{L2 loss} = \sum_{i=1}^{n} [\text{Denoiser}_i(y) - f_\theta(y)]^2 + \omega |f_\theta(y)|^2 \quad (6)$$

where y is the noisy image, θ are the network parameters, ω represents the variable weight of the loss function for the noisy image part, and Denoiser_i represents the i-th denoising algorithm chosen for its complementarity. Given that the amount of usable information in a noisy image diminishes as the noise level rises, the weight coefficient ω can be adjusted accordingly using specific settings as defined in [31].

4. Experimental Results

4.1. Datasets and Experimental Setup

For Gaussian noise image denoising experiments, widely recognized test datasets include Set12 [42] and BSD68 [43]. Additionally, to evaluate the model's adaptability and robustness, the experiments include denoising tests for images with mixed Gaussian–Poisson noise, real noisy images, and low-dose CT images. Real noisy image denoising experiments utilize datasets from Nam [44] and PolyU [45]. Low-dose CT image denoising uses images from the 2016 NIH-AAPM-Mayo Clinic Low-Dose CT Grand Challenge (LDCT) [46], with regular-dose CT images as a reference from the same dataset. Eight models in total were included in the evaluation: CNN-based models (DnCNN, FFDNet, DRUNet), Transformer-based methods (SwinIR, Restormer), masked training (MT) [47], traditional methods like BM3D, and the unsupervised DIP model. The training of the MPFDIP denoising model consisted of two phases. In the masked pre-training phase, training sets such as the Berkeley segmentation dataset (BSD) [48], Waterloo Exploration Database (WED) [49], DIV2K [50], and Flick2K [51] were utilized. The denoising effect was evaluated using objective metrics such as the Peak Signal-to-Noise Ratio (PSNR) and SSIM, along with subjective human visual assessments. All experiments were conducted on a graphics workstation equipped with an Intel i7-11700H CPU, which we bought online, an NVIDIA RTX3090 graphics card, and 32 GB of memory, the manufacturer is Inter, located in Santa Clara, CA, USA.

4.2. Ablation Experiments

The benchmark reference network model defaults to the following hyperparameter settings: 60 channels, six RTCBs in the feature extraction section, each containing six Transformer layers, and inclusion of both the sampling and refinement modules. This study tested various configurations of channel numbers, RTCB numbers, and BTL numbers on the Set12 dataset to evaluate their impact on the denoising effect and corresponding computational cost of the FRformer network. The results in Tables 1–3 indicate that increasing the number of channels, RTCBs, and BTLs leads to improved denoising effects, but also results in higher parameter counts and computational demands. In order to optimize the denoising effect while minimizing the resource requirements of the network, this study adopted a configuration with 60 channels, four RTCBs, and six BTLs. This configuration achieved a PSNR of 34.84 while maintaining relatively few network parameters and a low computational load.

To evaluate the influence of the sampling and refinement modules on the denoising capabilities, parameter load, and computational efficiency of the network, experiments were performed with and without these modules. The experiments utilized identical datasets and configurations as previously described. The results are listed in Table 4. According to the data from Table 4, the sampling module significantly reduces the computational

complexity, while the refinement module enhances the denoising effectiveness. Consequently, it was decided to integrate both the sampling and refinement modules into the final model configuration.

Table 1. Denoising effect and corresponding computational cost under different network channel sizes.

Network Channels	24	36	48	60	72	84	96
PSNR (dB)	33.37	34.49	34.74	34.93	35.00	35.15	35.21
Parameters (millions)	0.38	0.80	1.38	2.11	3.00	4.04	5.24
Floating-point Operations (billions)	1.55	3.28	5.64	8.64	12.27	16.55	21.45

Table 2. Denoising effect and corresponding computational cost under different RTCB configurations.

RTCB Numbers	1	2	4	6	8	10	12
PSNR (dB)	34.12	34.44	34.84	34.93	34.97	35.03	35.08
Parameters (millions)	0.75	1.03	1.57	2.11	2.65	3.19	3.74
Floating-point Operations (billions)	3.09	4.20	6.42	8.64	10.86	13.08	15.30

Table 3. Denoising effect and corresponding computational cost under different BLT configurations.

BLT Numbers	1	2	4	6	8	10	12
PSNR (dB)	34.07	34.27	34.56	34.84	34.89	34.95	35.00
Parameters (millions)	0.77	0.93	1.25	1.57	1.89	2.20	2.52
Floating-point Operations (billions)	3.16	3.81	5.12	6.42	7.72	9.03	10.33

Table 4. Impact of sampling and refinement modules on the denoising effect and corresponding computational cost.

Case No	Sampling Module	Refinement Module	PSNR (dB)	Parameters (Millions)	Floating-Point Operations (Billions)
1	✓	✓	34.84	1.57	6.42
2	✓		34.33	1.12	4.58
3		✓	34.96	1.57	25.64

To evaluate the influence of the pre-training phase of MPFDIP on subsequent unsupervised DIP training, the relationship between PSNR values and the number of iterations was documented under two conditions: with the pre-trained model loaded and with model parameters randomly initialized. The specific values are presented in Table 5. The experiments assessed PSNR results under the condition of artificially added Gaussian noise with a noise level of 25 on the Set12 dataset. By comparing PSNR values of network output images at various iteration steps in both scenarios, it becomes evident that the masked pre-training strategy of the MPFDIP model significantly accelerates the network's convergence speed. With only 400 iterations, its denoising performance substantially surpasses that of the randomly initialized parameter approach at 2000 iterations, enhancing the model's execution efficiency by 80%, while also achieving better denoising results. Here, we conducted a detailed analysis, using the starfish and sailboat images from the Set12 dataset as illustrative examples, to demonstrate this principle more concretely. For the starfish image with a resolution of 256 $\times$ 256, our iterative process spanning over two thousand steps consumed approximately 261.12 s. Remarkably, we achieved a maximum PSNR of 30.02 dB within these steps without utilizing a pre-trained model. However, an even higher PSNR of 30.26 dB was attained within a mere 400 steps, accompanied by a processing duration of about 52.22 s. On the other hand, for the sailboat image with a resolution of 512 $\times$ 512, iterating over two thousand steps demanded roughly 837.84 s. Specifically, the maximum PSNR attained within these steps, in the absence of a pre-trained model, was 29.30 dB. Surpassing this benchmark, a PSNR of 29.81 dB was achieved within just 400 steps, requiring a processing time of approximately 167.57 s. The experiments conducted above

have convincingly demonstrated that the adoption of pre-training techniques can ensure a significant improvement in denoising performance with minimal time costs.

Table 5. PSNR values for different iteration steps with and without the pre-trained model.

Iteration Steps	With the Pre-Trained Model	Without the Pre-Trained Model
0	20.45	6.79
200	29.38	21.36
400	30.81	26.96
600	31.03	27.98
800	31.14	28.59
1000	31.19	29.08
1200	31.23	29.46
1400	31.25	29.75
1600	31.20	29.98
1800	31.18	30.16
2000	31.17	30.30

Finally, to assess the impact of different combinations of target images on the performance of the MPFDIP model, we selected four representative denoising models from mainstream methods. We tested their denoising effects under various combinations of target images, as illustrated in Table 6. Each number in Table 6 represents a different combination of denoising algorithms, with the baseline row indicating the PSNR results when each denoising model operates independently. According to Table 6, when only two denoising algorithms are used and their performance significantly differs, the PSNR values decrease more compared to the baseline (combinations 1, 2, and 3). However, when three or more denoising models are used to generate target images, the variations in PSNR values relative to the baseline are more stable. After conducting comparative experiments with all combinations, based on the PSNR results of the output images, we selected combination number 10 to generate preprocessed images and used them as target images for the unsupervised training phase of the DIP model. Specifically, we employed DRUNet, SwinIR, and Restormer to process the noisy images and obtain preprocessed images, which, along with the noisy images, served as the target images.

Table 6. The impact of different combinations of denoising methods on model performance.

Combination No.	FFDNet	DRUNet	SwinIR	Restormer	PSNR
1	✓	✓			30.85
2	✓		✓		30.84
3	✓			✓	30.83
4		✓	✓		31.05
5		✓		✓	31.06
6			✓	✓	31.07
7	✓	✓	✓		30.95
8	✓	✓		✓	30.99
9	✓		✓	✓	31.01
10		✓	✓	✓	**31.09**
11	✓	✓	✓	✓	31.03
Baseline	30.46	30.94	31.02	31.04	

4.3. Quantitative Results

Initially, denoising effects were compared on the Set12 dataset for grayscale synthetic Gaussian noise images at noise levels of 15, 25, and 50, as depicted in Table 7. The proposed MPFDIP method exhibits superior performance in both PSNR and SSIM compared to other methods. Importantly, the MPFDIP model significantly outperforms the MT denoising model, which also employs a masking strategy, in both PSNR and SSIM metrics. This

underscores the crucial role of the unsupervised training aspect of the DIP model in enhancing the denoising performance.

Table 7. Comparison of denoising effects on the SET12 dataset for synthesized Gaussian noise images.

Method	PSNR (dB)				SSIM			
	15	25	50	Average	15	25	50	Average
BM3D	32.36	29.96	26.7	29.67	0.9499	0.9228	0.8664	0.9130
DnCNN	32.67	30.35	27.18	30.07	0.9526	0.9286	0.8769	0.9194
FFDNet	32.77	30.46	27.35	30.19	0.9542	0.9308	0.8821	0.9224
DRUNet	33.25	30.94	27.90	30.70	0.9577	0.9363	0.8939	0.9293
SwinIR	33.36	31.01	27.91	30.76	0.9584	0.9368	0.8940	0.9297
Restormer	33.35	31.04	28.01	30.80	0.9609	0.9376	0.8958	0.9314
MT	29.24	27.06	19.00	25.10	0.8496	0.7504	0.3389	0.6463
DIP	31.35	28.92	25.56	28.61	0.9354	0.9007	0.8279	0.8880
MPFDIP	**33.68**	**31.19**	**28.11**	**30.99**	**0.9625**	**0.9388**	**0.8969**	**0.9327**

Secondly, to evaluate the denoising effect of the MPFDIP method on simulated mixed noise images, a denoising comparison experiment was conducted on the BSD68 dataset with Gaussian–Poisson mixed noise at three different noise level intensities. These tests are represented by identifiers 1, 2, and 3, as shown in Table 8. Identifier 1 corresponds to adding Gaussian noise with a standard deviation of 15 and Poisson noise with intensity 1, identifier 2 corresponds to Gaussian noise with a standard deviation of 25 and Poisson noise with intensity 2, and identifier 3 corresponds to Gaussian noise with a standard deviation of 50 and Poisson noise with intensity 3. The MPFDIP denoising model demonstrates superior PSNR and SSIM values compared to other algorithms when dealing with various levels of Gaussian–Poisson mixed noise. The denoising effect is significantly improved compared to the preprocessing algorithms used, and it is also better than the original DIP denoising model. This indicates that masked pre-training can enhance the generalization capability of the denoising model, making it more robust.

Table 8. Comparison of denoising effects on the BSD68 dataset for mixed Gaussian–Poisson noisy images.

Method	PSNR (dB)				SSIM			
	1	2	3	Average	1	2	3	Average
BM3D	29.45	25.89	23.43	26.26	0.9133	0.8307	0.7738	0.8393
DnCNN	29.02	24.61	23.00	25.54	0.8115	0.6343	0.6111	0.6856
FFDNet	28.86	24.30	22.47	25.21	0.7878	0.5771	0.4986	0.6212
DRUNet	28.83	24.26	22.72	25.27	0.7978	0.6111	0.5781	0.6623
SwinIR	28.82	24.27	22.63	25.24	0.7966	0.6052	0.5726	0.6581
Restormer	28.71	24.05	21.73	24.83	0.8856	0.7320	0.6685	0.7620
MT	27.97	23.90	17.83	23.23	0.8125	0.6083	0.3206	0.5805
DIP	27.58	25.80	23.05	25.48	0.8655	0.8157	0.7236	0.8016
MPFDIP	**30.65**	**27.39**	**23.85**	**27.30**	**0.9310**	**0.8748**	**0.7840**	**0.8633**

Additionally, to assess the denoising performance of the MPFDIP on real noisy images, tests were conducted on the Nam, PolyU, and LDCT real noisy datasets. The results are shown in Table 9. The MPFDIP denoising model leads other comparative algorithms in terms of PSNR values. Although the SSIM values on the Nam and PolyU real noise imagy datasets are slightly lower than those achieved by the BM3D method, the SSIM values for denoising low-dose CT images are higher than BM3D. Overall, MPFDIP still leads in terms of the average SSIM value compared to other comparative algorithms.

Table 9. Denoising effect across Nam, PolyU, and LDCT datasets.

Method	PSNR (dB)				SSIM			
	Nam	PolyU	LDCT	Average	Nam	PolyU	LDCT	Average
BM3D	41.19	38.65	31.73	37.19	**0.9944**	**0.9880**	0.9007	0.9610
DnCNN	36.90	35.15	31.32	34.46	0.9128	0.8817	0.7428	0.8458
FFDNet	41.88	38.47	31.64	37.33	0.9717	0.9557	0.7508	0.8927
DRUNet	43.08	39.50	31.52	38.03	0.9899	0.9744	0.7497	0.9047
SwinIR	42.94	39.37	31.50	37.94	0.9892	0.9732	0.7526	0.9050
Restormer	41.55	38.87	30.59	37.00	0.9900	0.9827	0.8859	0.9529
MT	35.38	35.34	31.15	33.96	0.9620	0.9544	0.7630	0.8931
DIP	40.41	37.90	31.42	36.58	0.9890	0.9831	0.8960	0.9560
MPFDIP	**43.29**	**39.83**	**32.02**	**38.38**	0.9942	0.9878	**0.9098**	**0.9639**

4.4. Qualitative Comparison

To visually analyze the denoising effects of the MPMTDIP model, we first applied various denoising methods to the starfish image from the Set12 dataset, which was subjected to Gaussian noise with a noise level of 50. The overall denoised image and the visual effects of specific localized areas are shown in Figure 6. From the enlarged sub-images, it is evident that for areas with color uniformity, supervised deep-learning-based denoising models like FFDNet, DRUNet, and SwinIR tend to produce overly smoothed results. The performance of the unsupervised DIP algorithm is also suboptimal. The MT model, which uses training models with Gaussian noise levels of 15, includes a masking module to enhance network generalization. However, as it remains a supervised denoising model, its performance on noise images with a level of 50 is poor. In contrast, the MPMTDIP model, which integrates both supervised and unsupervised learning strategies, shows significant advantages, with restoration results that are closest to the original image and exhibit the best noise reduction and detail preservation capabilities.

To further analyze the visual effects of the MPFDIP denoising model on images with Gaussian–Poisson mixed noise, various denoising methods were applied to the test004 image from the BSD68 dataset, labeled with mixed noise intensity number 2. The enlarged sub-images shown in Figure 7 reveal that in the duckbill portion of the test004 image, residual noise remains noticeable in the outputs of DnCNN, FFDNet, DRUNet, SwinIR, and Restormer models. This is attributed to the fact that supervised denoising models often experience a decline in performance when dealing with images with noise distributions different from those encountered during training. Additionally, the unsupervised DIP denoising model exhibits poorer results due to severe damage in the target images. Although the MT model, based on a masking strategy, incorporates a masking module to enhance network generalization, it is inherently a supervised model; thus, its denoising performance on Gaussian–Poisson mixed noise images is still significantly impacted. In contrast, compared to other denoising methods, MPFDIP demonstrates superior noise reduction capabilities for mixed noise, better preserves image details, and produces images with a closer resemblance to the original image. Its ability to remove more noise from images surpasses that of supervised deep models.

Additionally, to validate the visual effects of the MPFDIP denoising model under real noise conditions, various denoising methods were applied to noisy images from the low-dose CT dataset. The results are illustrated in Figure 8. Observations of the detailed sections of the denoised images reveal that the results from DnCNN and FFDNet exhibit lower clarity. The outcomes from DRUNet and SwinIR are relatively smoother, while the results from MT and DIP appear more blurred. In contrast, the outcomes from the MPFDIP model closely resemble those of normal-dose CT images, effectively reducing the interference caused by noise and providing a higher image clarity. This once again underscores its strong generalization capabilities.

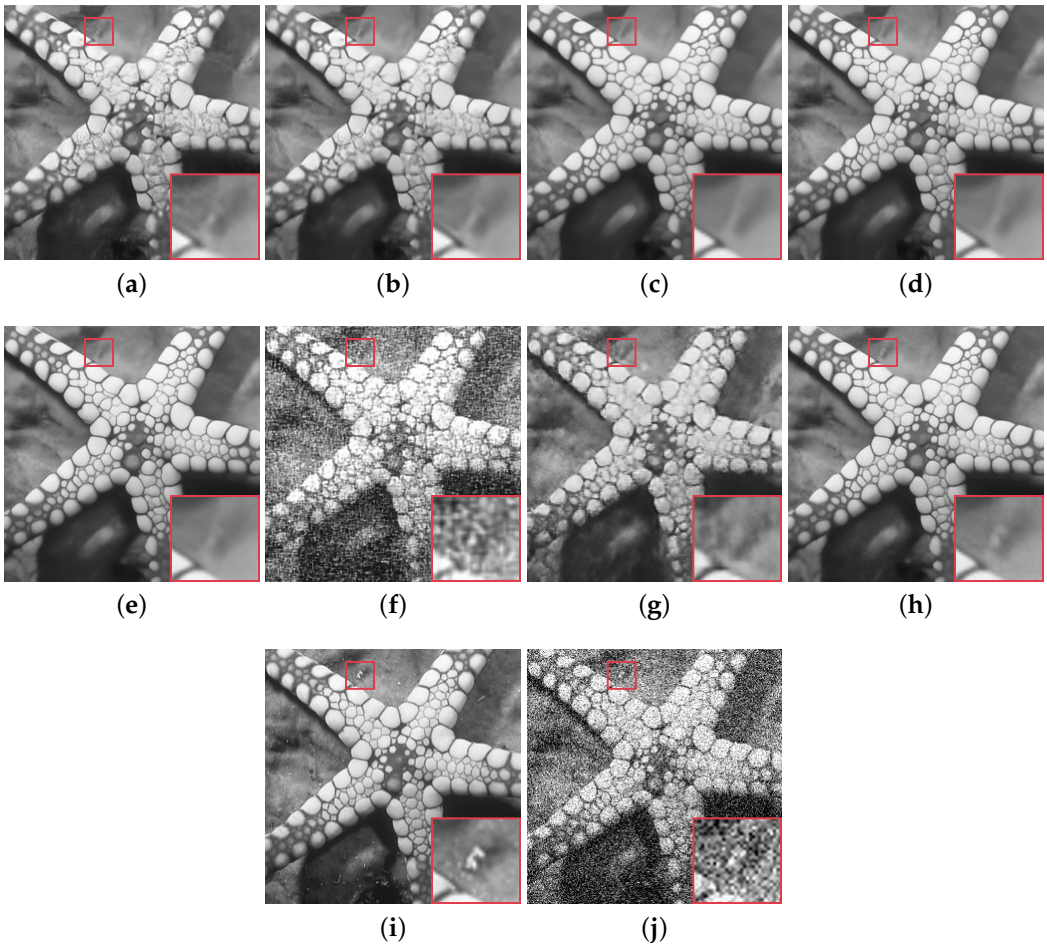

Figure 6. Comparison of denoising effects of various methods on Gaussian noise. (**a**) PSNR = 25.60 dB, DnCNN [14]; SSIM = 0.7714; (**b**) PSNR = 25.72 dB; SSIM = 0.7769, FFDNet [15]; (**c**) PSNR = 26.49 dB; SSIM = 0.8013, DRUNet [16]; (**d**) PSNR = 26.55 dB; SSIM = 0.8030, SwinIR [18]; (**e**) PSNR = 26.67 dB; SSIM = 0.8827, Restormer [19]; (**f**) PSNR = 19.17 dB; SSIM = 0.4487, MT [47]; (**g**) PSNR = 24.35 dB; SSIM = 0.8217, DIP [29]; (**h**) PSNR = 27.98 dB; SSIM = 0.9061, MPFDIP; (**i**) reference; (**j**) PSNR = 14.90 dB; SSIM = 0.3340, noisy.

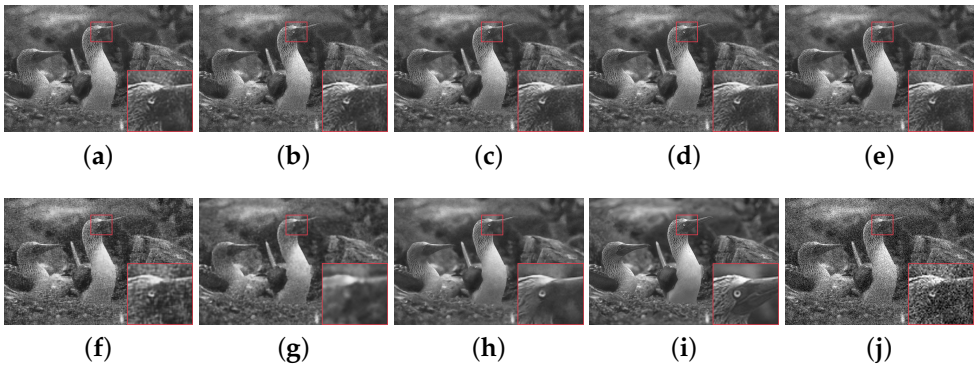

Figure 7. Comparison of denoising effects of various methods on Gaussian–Poisson mixed noise (**a**) PSNR = 25.90 dB; SSIM = 0.5978, DnCNN [14]; (**b**) PSNR = 25.53 dB; SSIM = 0.5647, FFDNet [15]; (**c**) PSNR = 25.47 dB; SSIM = 0.5608, DRUNet [16]; (**d**) PSNR = 25.51 dB; SSIM = 0.5645, SwinIR [18]; (**e**) PSNR = 25.36; SSIM = 0.7294, Restormer [19]; (**f**) PSNR = 24.23 dB; SSIM = 0.5143, MT [47]; (**g**) PSNR = 27.86 dB; SSIM = 0.8542, DIP [29]; (**h**) PSNR = 30.19 dB; SSIM = 0.9014, MPFDIP; (**i**) reference; (**j**) PSNR = 18.34 dB; SSIM = 0.3530, noisy.

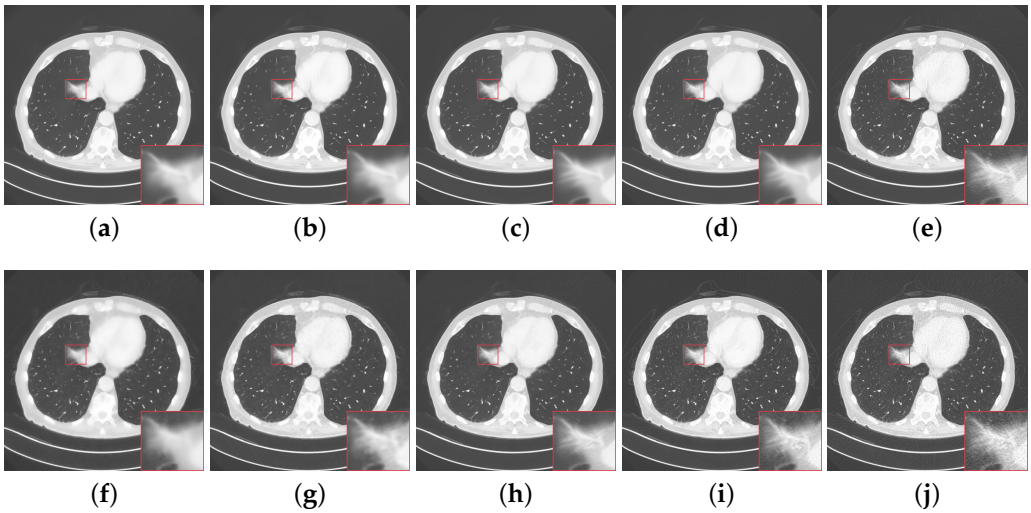

Figure 8. Comparison of denoising effects of various methods on low-dose CT images: (**a**) PSNR = 31.61 dB; SSIM = 0.7506, DnCNN [14]; (**b**) PSNR = 31.88 dB; SSIM = 0.7577, FFDNet [15]; (**c**) PSNR = 31.76 dB; SSIM = 0.7561, DRUNet [16]; (**d**) PSNR = 31.76 dB; SSIM = 0.7593, SwinIR [18]; (**e**) PSNR = 31.10 dB; SSIM = 0.8963, Restormer [19]; (**f**) PSNR = 31.35 dB; SSIM = 0.7720, MT [47]; (**g**) PSNR = 31.74 dB; SSIM = 0.9000, DIP [29]; (**h**) PSNR = 32.62 dB; SSIM = 0.9202, MPFDIP; (**i**) reference (**j**) PSNR = 28.61 dB; SSIM = 0.8288, noisy.

5. Conclusions

The MPFDIP model, a denoising model, has been developed and extensively tested in various synthetic and real noisy environments. The experimental results demonstrate that MPFDIP surpasses both the original DIP model and supervised denoising models. This improvement can be attributed to several factors. Firstly, MPFDIP utilizes network parameters obtained from supervised pre-training to initialize the unsupervised DIP phases, resulting in a significantly accelerated convergence. Secondly, this approach effectively combines the strengths of both supervised and unsupervised learning. It allows the model to learn prior information during the supervised pre-training phase and adapt during the unsupervised online training phase, enabling a superior performance in image denoising tasks across diverse application scenarios. Additionally, refined multi-target image techniques and the adaptive loss function further enhance the denoising performance of MPFDIP. Finally, the two-stage construction of the MPFDIP model allows for future incorporation of higher-performance supervised denoising models as pre-training models, ensuring its extendability and the potential for further improvements in denoising results. Note that the model we propose is designed to achieve an enhanced denoising performance with minimal time costs. Compared to current supervised models, our model still has room for improvement in terms of the execution speed. Therefore, future research will focus on further enhancing the execution speed of our model.

Author Contributions: S.X. and S.J. contributed to the conception of the study, S.J. wrote the main manuscript text, N.X. and C.Z. contributed significantly to analyses and manuscript preparation, and Q.C. and M.X. conducted experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China, grant number 62162043.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be made available on request.

Acknowledgments: The authors would like to acknowledge the reviewers and the AE for their constructive comments and suggestions that helped to improve the paper's quality.

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

Abbreviations

DIP	Deep Image Prior
FRformer	Fast Restormer
MPFDIP	Masked-Pre-training-Based Fast Deep Image Prior Denoising Model
BM3D	Block-Matching and 3D Filtering
NLM	Non-Local Mean Filter
NCSR	Non-Local Centralized Sparse Representation
WNNM	Weighted Nuclear Norm Minimization
DNNs	Deep Neural Networks
DnCNN	Denoising Convolutional Neural Network
FFDNet	Fast and Flexible Denoising Convolutional Neural Network
DRUNet	Dilated Residual UNet
SNR	Signal-to-Noise Ratio
SwinIR	Image Restoration Using Swin Transformer
MDTA	Multi-Dconv Head Transposed Attention
GDFN	Gated-Dconv Feed-forward Network
EWT	Efficient Wavelet Transformer
NEF	Neighborhood Feature Enhancement
MLP	Multiple-Layer Perceptron
PwConv	Point-wise Convolution
DwConv	Depth-wise Convolution
GELU	Gaussian Error Linear Unit
RTCB	Residual Transformer Convolutional Block
BTL	Basic Transformer Layer
FRM	Feature Refinement Module
MRTCB	Mini-residual Transformer Convolutional Block
MIM	Masked Image Modeling
LAD	Least Absolute Deviation
LSE	Least Squares Error
LDCT	Low-Dose CT Grand Challenge
MT	Masked Training
BSD	Berkeley Degmentation Dataset
WED	Waterloo Exploration Database
PSNR	Peak Signal-to-Noise Ratio

References

1. Chen, Z.; Kaushik, P.; Shuangfei, Z.; Alvin, W.; Zhile, R.; Alex, S.; Alex, C.; Li, F. AutoFocusFormer: Image Segmentation off the Grid. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18227–18236.
2. Jie, Q.; Wu, J.; Pengxiang, Y.; Ming, L.; Ren, Y.; Xuefeng, X.; Yitong, W; Rui, W.; Shilei, W.; Xin, P.; et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19446–19455.
3. Li, X.; Liu, Z.; Wu, J.J. Attentional Full-Relation Network for Few-Shot Image Classification. *Chin. J. Comput.* **2023**, *46*, 371–384.
4. Ahmad, M.; Mazzara, M. SCS-Net: Sharpend cosine similarity based neural network for hyperspectral image classification. *IEEE Geosci. Remote. Sens. Lett.* **2024**, *21*, 1–4.
5. Xiong, W.; Xiong, C.; Gao, Z.R. Channel Attention Embedded Transformer for Image Super-Resolution Reconstruction. *J. Image Graph. China* **2023**, *28*, 3744–3757.
6. Zhou, D.W.; Liu, Z.H.; Liu, Y.K. Image Super-Resolution Algorithm Based on Pixel Contrast Learning. *Acta Autom. Sin.* **2024**, *50*, 181–193.
7. Jin, Y.; Yang, W.; Tan, R.T. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In Proceedings of the European Conference on Computer Vision, Tel-Aviv, Israel, 23–27 October 2022; pp. 404–421.
8. Ying, Z.; Li, G.; Ren, Y.; Wang, R.; Wang, W. A new image contrast enhancement algorithm using exposure fusion framework. In Proceedings of the Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, 22–24 August 2017; pp. 36–46.
9. Dabov, K.; Foi, A.; Katkovnik, V.; Egiazarian, K. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **2007**, *16*, 2080–2095. [CrossRef] [PubMed]

10. Buades, A.; Coll, B.; Morel, J.M. A non-local algorithm for image denoising. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; pp. 60–65.
11. Dong, W.; Zhang, L.; Shi, G.; Li, X. Nonlocally centralized sparse representation for image restoration. *IEEE Trans. Image Process.* **2012**, *22*, 1620–1630. [CrossRef] [PubMed]
12. Gu, S.; Zhang, L.; Zuo, W.; Feng, X. Weighted nuclear norm minimization with application to image denoising. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2862–2869.
13. Sheng, J.; Lv, G.; Wang, Z.; Feng, Q. SRNet: Sparse representation-based network for image denoising. *Digit. Signal Process.* **2022**, *130*, 103702. [CrossRef]
14. Zhang, K.; Zuo, W.; Chen, Y.; Chen, Y.; Zhang, L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **2017**, *26*, 3142–3155. [CrossRef]
15. Zhang, K.; Zuo, W.; Zhang, L. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Trans. Image Process.* **2018**, *27*, 4608–4622. [CrossRef]
16. Zhang, K.; Li, Y.; Zuo, W.; Zhang, L.; Van Gool, L.; Timofte, R. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6360–6376. [CrossRef] [PubMed]
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
18. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image restoration using Swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, USA, 19–25 June 2021; pp. 1833–1844.
19. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M. 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.
20. 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.
21. Li, J.; Cheng, B.; Chen, Y.; Gao, G.; Shi, J.; Zeng, T. EWT: Efficient wavelet-transformer for single image denoising. *arXiv* **2023**, arXiv:2304.06274.
22. Yuan, J.; Zhou, F.; Guo, Z.; Li, X.; Yu, H. HCformer: Hybrid CNN-transformer for LDCT image denoising. *J. Digit. Imaging* **2023**, *36*, 2290–2305. [CrossRef] [PubMed]
23. Brooks, T.; Mildenhall, B.; Xue, T.; Chen, J.; Sharlet, D.; Barron, J.T. Unprocessing images for learned raw denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11036–11045.
24. Lehtinen, J.; Jacob, M.; Jon, H.; Samuli, L.; Tero, K.; Miika, A.; Timo, A. Noise2Noise: Learning image restoration without clean data. *arXiv* **2018**, arXiv:1803.04189.
25. Krull, A.; Buchholz, T.O.; Jug, F. Noise2void-learning denoising from single noisy images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2129–2137.
26. Batson, J.; Royer, L. Noise2Self: Blind Denoising by Self-Supervision. In Proceedings of the International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 9–15 June 2019; pp. 524–533.
27. Huang, T.; Li, S.; Jia, X.; Lu, H.; Liu, J. Neighbor2Neighbor: A self-supervised framework for deep image denoising. *IEEE Trans. Image Process.* **2022**, *31*, 4023–4038. [CrossRef] [PubMed]
28. Quan, Y.; Chen, M.; Pang, T.; Ji, H. Self2self with dropout: Learning self-supervised denoising from single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1890–1898.
29. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep image prior. *Int. J. Comput. Vis.* **2020**, *128*, 1867–1888. [CrossRef]
30. Xu, S.P.; Li, F.; Chen, X.H.; Chen, X.J.; Jiang, S.L. An Image Denoising Model Constructed Using Improved Deep Image Prior. *Acta Electron. Sin.* **2022**, *50*, 1573–1578.
31. Xu, S.P.; Xiao, N.; Luo, J.; Cheng, X.H.; Chen, X.J. Dual-Channel Deep Image Prior Denoising Model. *Acta Electron. Sin.* **2024**, *52*, 58–68.
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
34. Zhang, K.; Li, Y.; Liang, J.; Cao, J.; Zhang, Y.; Tang, H.; Timofte, R.; Gool, L.V. Practical blind image denoising via Swin-Conv-UNet and data synthesis. *Mach. Intell. Res.* **2023**, *20*, 822–836. [CrossRef]
35. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
36. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
37. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.

38. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
39. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
40. Hénaff, O.J.; Srinivas, A.; Fauw, J.D.; Razavi, A.; Doersch, C.; Eslami, S.M.; Oord, A.V. Data-efficient image recognition with contrastive predictive coding. In Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020; pp. 4182–4192.
41. Meinhardt, T.; Moller, M.; Hazirbas, C.; Cremers, D. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1781–1790.
42. Roth, S.; Black, M.J. Fields of experts: A framework for learning image priors. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; pp. 860–867.
43. Roth, S.; Black, M.J. Fields of experts. *Int. J. Comput. Vis.* **2009**, *82*, 205–229. [CrossRef]
44. Nam, S.; Hwang, Y.; Matsushita, Y.; Kim, S.J. A holistic approach to cross-channel image noise modeling and its application to image denoising. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1683–1691.
45. Xu, J.; Li, H.; Liang, Z.; Zhang, D.C.; Zhang, L. Real-world noisy image denoising: A new benchmark. *arXiv* **2018**, arXiv:1804.02603.
46. Yan, R.; Liu, Y.; Liu, Y.; Wang, L.; Zhao, R.; Bai, Y.; Gui, Z. Image denoising for low-dose CT via convolutional dictionary learning and neural network. *IEEE Trans. Comput. Imaging* **2023**, *9*, 83–93. [CrossRef]
47. Chen, H.; Gu, J.; Liu, Y.; Magid, S.A.; Dong, C.; Wang, Q.; Pfister, H.; Zhu, L. Masked image training for generalizable deep image denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1692–1703.
48. Chen, Y.; Pock, T. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1256–1272. [CrossRef] [PubMed]
49. Ma, K.; Duanmu, Z.; Wu, Q.; Wang, Z.; Yong, H.; Li, H.; Zhang, L. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Trans. Image Process.* **2016**, *26*, 1004–1016. [CrossRef] [PubMed]
50. Agustsson, E.; Timofte, R. NTIRE Challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 126–135.
51. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.