

Multiple Degradation and Reconstruction Network for Single Image Denoising via Knowledge Distillation

Juncheng Li^{1,2}, Hanhui Yang¹, Qiaosi Yi², Faming Fang², Guangwei Gao³, Tieyong Zeng^{1*}, Guixu Zhang²

¹The Chinese University of Hong Kong ²East China Normal University

³Nanjing University of Posts and Telecommunications

{cvjunchengli, qiaosiyijoyies, csggao}@gmail.com

{hhYang, zeng@math.cuhk.edu.hk, {fmfang, gxzhang}@cs.ecnu.edu.cn}

Abstract

Single image denoising (SID) has achieved significant breakthroughs with the development of deep learning. However, the proposed methods are often accompanied by plenty of parameters, which greatly limits their application scenarios. Different from previous works that blindly increase the depth of the network, we explore the degradation mechanism of the noisy image and propose a lightweight Multiple Degradation and Reconstruction Network (MDRN) to progressively remove noise. Meanwhile, we propose two novel Heterogeneous Knowledge Distillation Strategies (HMDS) to enable MDRN to learn richer and more accurate features from heterogeneous models, which make it possible to reconstruct higher-quality denoised images under extreme conditions. Extensive experiments show that our MDRN achieves favorable performance against other SID models with fewer parameters. Meanwhile, plenty of ablation studies demonstrate that the introduced HMDS can improve the performance of tiny models or the model under high noise levels, which is extremely useful for related applications.

1. Introduction

Single image denoising (SID) aims to reconstruct a clean image from the noisy one, which has been widely used as an initial step for many high-level computer vision tasks. This is because the quality of the denoised images will significantly influence the accuracy of these downstream tasks. However, it still is a challenging task since many important features in the original image has be seriously occluded or corrupted by noise.

During the past decades, the Convolutional Neural Network (CNN) has achieved remarkable achievements in image processing [5, 21, 32], which also greatly promotes the development of SID. Recently, many CNN-based models

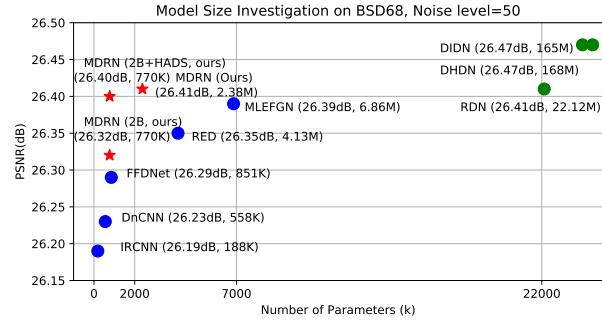


Figure 1. Model size investigation. Red stars denote our models.

have been proposed for SID, especially for the Additive White Gaussian Noise (AWGN). Different from traditional methods, these methods [7, 13, 18, 22, 25, 31, 34, 38] usually learn the mapping between noisy and clear images by building a well-designed CNN. For example, Zhang et al. [38] proposed the first end-to-end trainable CNN model (DnCNN) for Gaussian denoising, which took the advantage of batch normalization and residual learning to recover clean images. Meanwhile, Zhang et al. [40] also proposed a flexible FFDNet, which took the noisy image and estimated noise level map as inputs, thus a single model could deal with noise on different levels. Liu et al. [18] presented a multi-level wavelet CNN (WCNN) for better trade-off between the receptive field size and the computational efficiency. Fang et al. [7] introduced image edge prior into CNN and proposed a Multilevel Edge Features Guided Network (MLEFGN) for SID. On the other hand, some large models have also been proposed and constantly refreshed the best results. For instance, Zhang et al. [43] proposed a Residual Dense Network (RDN) for image restoration with residual dense block. Park et al. [22] studied a Densely Connected Hierarchical Network (DHDN) by using a modified U-net architecture; Although the above models achieve promising results, they still do not handle high-noise images

*Corresponding author

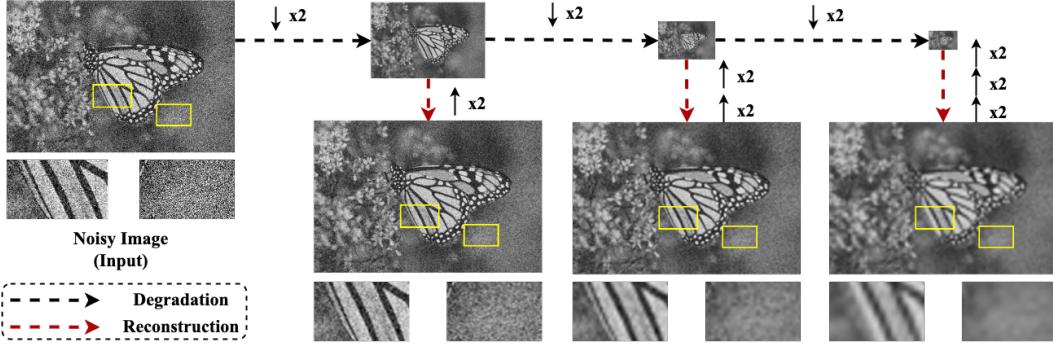


Figure 2. An example of image denoising ($\sigma = 50$) by using multiple degradation and reconstruction operations (via Bicubic).

well. Meanwhile, these models are often accompanied by a large number of parameters (Figure 1), which greatly limits their application in real scenarios.

Considering the popularity of smart mobile devices, it is extremely important to build a lightweight and efficient SID model. As shown in Figure 2, we find that some noise in the noisy image can be eliminated by using the simple down- and up-sampling operations. However, we also notice that the downsampling operation also results in the loss of useful information when removing noise. Meanwhile, the bigger the downsampling rate, the more noise will be removed, but the reconstructed image will become more blurred. To solve the aforementioned problem, we propose a Multiple Degradation and Reconstruction Network (MDRN) to progressively remove image noise inspired by Unet++ ([45]). Meanwhile, Multi-scale Aggregation Block (MSAB) and Multi-scale Aggregation Group (MSAG) are specially designed for multi-scale feature extraction. Meanwhile, we propose two novel Heterogeneous Knowledge Distillation Strategies (HKDS) to enable MDRN to learn richer and more accurate features from the teacher model, thereby further improving the model performance. In other words, the model uses the knowledge provided by external heterogeneous models to realize the automatic learning of deep feature priors. In summary, the contributions of this paper are as follows:

(1). We propose a lightweight and efficient Multiple Degradation and Reconstruction Network (MDRN) for image denoising, which can progressively remove image noise via multiple down- and up-sampling operations.

(2). We propose a Multi-Scale Aggregation Group (MSAG) for feature extraction. MSAB is the basic components of MDRN, which can extract rich multi-scale features with few parameters.

(3). We propose two Heterogeneous Knowledge Distillation Strategies (HKDS) for SID. With the help of HKDS, MDRN can learn richer and more accurate features from the teacher model, thereby improving the denoising ability of the lightweight model under extreme conditions.

2. Related Work

2.1. Image Denoising

The degradation mode of the noisy image is usually formulated as $y = x + v$, where x is the clean image, v is the additive white Gaussian noise (AWGN), and y is noisy image polluted by AWGN. Recently, many excellent methods have been proposed for clear images estimation, including spatial filtering methods [2,9], model-based methods [6,27], and deep learning-based methods [7,13,18,25,31,37,38,40,43]. Among them, deep learning-based methods have become the mainstream of SID, which aim to solve the problem via learning the mapping between the noisy and clean images. For example, Zhang et al. [38] proposed a DnCNN for the Gaussian noise removal, which achieved competitive results by took advantage of batch normalization and residual learning. Yu et al. [34] proposed a Deep Iterative Down-up Network (DIDN) for image denoising, which achieved promising results. However, directly minimizing the loss function is difficult to learn accurate mapping. To solve this problem, Fang et al. [7] introduced the edge priors to guide image denoising and proposed a Multi-level Edge Features Guided Network (MLEFGN). However, the edge prior still is an artificially designed image prior and it is only one of many image priors. This certain extent limits its flexibility and versatility so that it can only improve the local effects in the image. Different from it, we aim to explore a method that allows the model to automatically learn the deep priors according to the need of the image, and to build a more efficient and lightweight SID model.

2.2. Knowledge Distillation

Knowledge distillation (KD) was first proposed by Hinton et al. [10], also known as the teacher-student framework. The goal of the KD mechanism is to transfer the knowledge from a complex model (Teacher) to a simple one (Student), which is often used to compress model size. Recently, many different KD strategies have been proposed, including output transfer [10,44], feature transfer [23,35], and relation

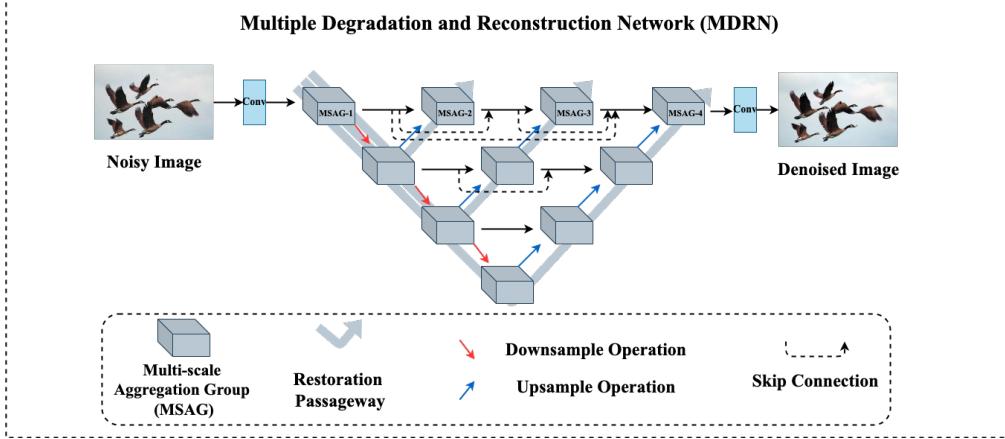


Figure 3. The architecture of Multiple Degradation and Reconstruction Network (MDRN). It is worth noting that convolutional and deconvolutional layers are introduced to replace the original down and up-sampling operations performed by Bicubic.

transfer [33]. All these mechanisms have been widely used in the image classification task and achieved remarkable results. Meanwhile, some researchers are also trying to introduce the KD mechanism into the image restoration tasks. For example, Hong et al. [11] proposed a KD-based method for image dehazing with the heterogeneous task. Specifically, the teacher model is an off-the-shelf autoencoder network, used to learn the reconstruction process from clear to clear images and transfer the learned knowledge to the dehazing model during the training process. Learning knowledge from clear images is an interesting idea. However, our investigation shows that the reconstruction of clear to clear images using a flat network such as U-net [24], is a relatively simple task that is difficult to ensure all layers or modules in the teacher model are fully trained since a flat network with skip connection could directly transmit all original information to the last layer. Therefore, it will make the teacher model fail to guide the student model efficiently. In this paper, we aim to explore more interpretable knowledge distillation strategies for SID to further improve the denoising ability of the tiny model, especially under high noise levels.

3. Proposed Method

In this paper, we propose a lightweight and efficient Multiple Degradation and Reconstruction Network (MDRN) for SID. Meanwhile, we present a Heterogeneous Architecture Distillation Strategy (HADS) and a Heterogeneous Mode Distillation Strategy (HMDS) to automatically learn the required deep prior knowledge from heterogeneous models, thereby further improving the model performance.

3.1. MDRN

As shown in Figure 3, the core part of the Multiple Degradation and Reconstruction Network (MDRN) is an in-

verted pyramid, which consists of a series of Multi-scale Aggregation Groups (MSAGs). Meanwhile, we can clearly observe that the model contains multiple degradation and reconstruction operations, which are used to remove noise and restore high-frequency detail, respectively. It is worth noting that convolutional and deconvolutional layers are introduced to replace the original down and up-sampling operations performed by Bicubic. This method can automatically learn how to remove useless noises while preserving useful information. In addition, after each downsampling operation, we will restore it immediately, thus we can obtain rich features reconstructed after different scales of down-sampling operations. Furthermore, we introduce the dense connection strategy at each level of the model to improve the utilization of hierarchical features, thus further improving the model performance.

3.1.1 Multi-scale Aggregation Group

The Multi-scale Aggregation Group (MSAG) is the most basic and important module of MDRN. According to Figure 4, we can observe that MSAG is composed of a 1×1 fusion layer and N MSABs, which is a simple but efficient feature extraction module. Meanwhile, we also apply a long skip connection for global residual learning like [41]. This strategy can increase the information flow and solve the gradient disappearance problem.

Multi-scale Aggregation Block. In order to extract rich features, we propose a Multi-scale Aggregation Block (MSAB). MSAB is an efficient feature extraction module that can extract rich multi-scale features with few parameters, which is inspired by MSRB [17] and MDCB [16], and the number of parameters of the module is reduced by using the channel splitting mechanism and dilated convolutional layers. As shown in Figure 4, after a 1×1 convolutional

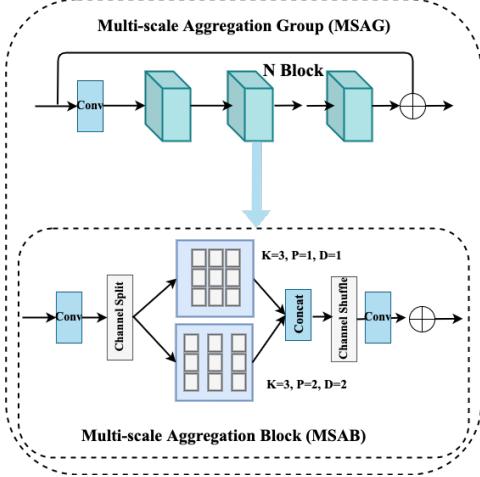


Figure 4. The architecture of MSAG. K , P , and D represent the kernel size, padding, and dilation rate, respectively.

layer, we use the channel splitting operation to divide the features into two groups. Then, we apply the convolutional layer with different dilation rates on different groups to extract image features with different scales. After that, all extracted features are concatenated, and the channel shuffle operation is applied to overcome the side effects brought by the channel splitting. Finally, a 1×1 convolutional layer is used for feature fusion and the residual learning strategy is introduced to further improve the information flow.

Hierarchical Dense Connection. MDRN is a multiple degradation and reconstruction model, thus can produce rich degraded and restored features. However, if these features are independent of each other, it is not beneficial for the final denoised image reconstruction. To solve this problem, we introduce dense connections at each level to fully utilize hierarchical features. Specifically, the current MSAG receives the feature maps of all preceding MSAGs x_0^j, x_1^j, \dots , and x_{i-1}^j in the same level as inputs

$$x_i^j = H_i^j([x_0^j, x_1^j, \dots, x_{i-1}^j]), \quad (1)$$

where $[x_0^j, x_1^j, \dots, x_{i-1}^j]$ refers to the concatenation operation and $H_i^j(\cdot)$ denotes the i^{th} MSAG in the j^{th} level.

3.2. Heterogeneous Knowledge Distillation

Although MDRN is an effective SID model, it is still difficult to reconstruct high-quality denoised images under extreme conditions, such as tiny models or severe noises. Therefore, we propose two novel Heterogeneous Knowledge Distillation Strategies (HKDS) to enable MDRN to automatically learn the deep feature priors from the teacher model. Different from previous works [11, 15, 30] that learned the knowledge from different models or tasks, we explore more efficient knowledge distillation strategies

from the model and the task itself.

Heterogeneous Architecture Distillation (HADS).

HADS aims to transfer the knowledge from a large model to the small one. As shown in Figure 5 (Top), the framework contains two models, a large teacher model and a student model. **Both of these two models have the same backbone, but the number of MSAB in each MSAG of these two models is different.** In other words, the purple module has more MSAB than the gray one. Except for this, all settings of these two models are consistent, including the noise level. Under this setting, each MSAG in the teacher model can extract richer and more accurate features than the student. We hope that the features extracted by MSAG in the student model can be as similar as possible to the features extracted by MSAG in the teacher model. To achieve this, we introduce Knowledge Distillation Loss (KDL) between the reconstructed outputs (MSAG-2, MSAG-3, and MSAG-4) of these two models to enable the student model can learn more accurate features from the teacher. With the help of the HADS, the student model can converge faster and the model performance can be further improved, which benefits for lightweight and accurate model construction.

Heterogeneous Mode Distillation (HMDS).

Although the deep model can learn more accurate features, it is still difficult to deal with severe noise (e.g., $\sigma = 70$) since the input image has been seriously damaged. In this case, changing the size of the model itself is difficult to effectively improve the model performance. To address this issue, we propose a Heterogeneous Mode Distillation Strategy (HMDS). As shown in Figure 5 (Bottom), the framework also contains two models. It is worth noting that these two models are exactly the same in model structure and size. The only difference is that the noise levels handled by these two models are different. In other words, HMDS aims to transfer the knowledge from a denoising model designed for slight noise to a model designed for severe noise. For example, the denoising model designed for low noise levels (e.g., $\sigma = 30, 50$) is used as the teacher model to guide the model designed for high noise levels (e.g., $\sigma = 70$). This is because low noise level images are easier to restore than high noise level images. Therefore, the student model can learn from the teacher model more accurate high-frequency details which are absent from its own input image. Meanwhile, the model can recover more accurate texture details even under high noise levels. The same as the HADS, we also introduce knowledge distillation loss between these two models.

3.3. Loss Function

Reconstruction Loss. Following previous works, we also adopt L1 loss to measure the difference between the denoised image and the corresponding clean image

$$\mathcal{L}_{RL} = \|F_{SID}(I_{noisy}) - I_{clear}\|_1, \quad (2)$$

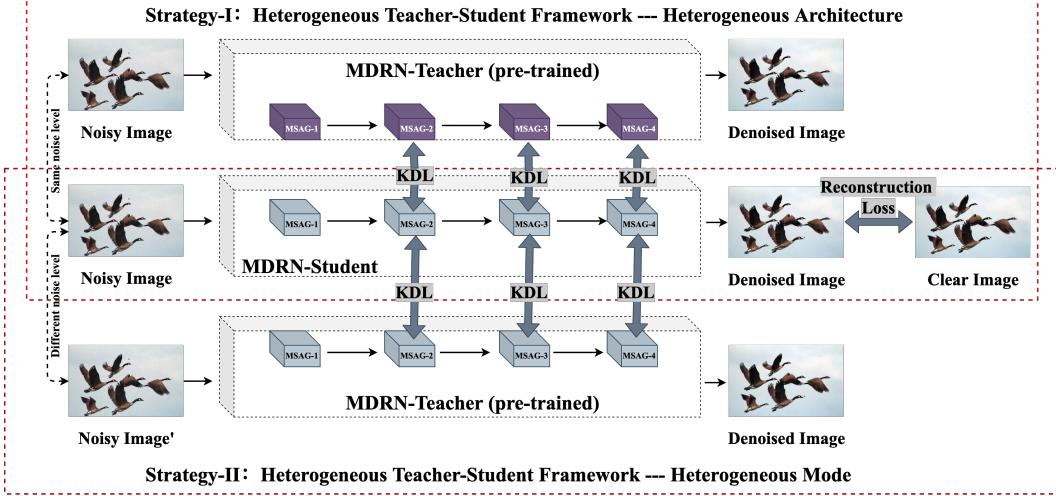


Figure 5. Schematic diagram of two heterogeneous knowledge distillation strategies. It is worth noting that only the top 4 MSAGs of MDRN are visualized in this figure for a clearer illustration .

where I_{noisy} and I_{clear} represent the noisy input and corresponding clear image, respectively. $F_{SID}(\cdot)$ denotes the proposed MDRN and $F_{SID}(I_{noisy})$ is the denoised result.

Knowledge Distillation Loss. In this paper, we propose two heterogeneous knowledge distillation strategies. In order to ensure that these two strategies can truly realize knowledge transfer, we propose a Knowledge Distillation Loss (KDL). KDL is essentially a feature matching function, which can be defined as

$$\mathcal{L}_{KDL} = \sum_{i=2,3,4} \|S_G^i(I_{noisy}) - T_G^i(I'_{noisy})\|_1, \quad (3)$$

where $S_G^i(I_{noisy})$ and $T_G^i(I_{noisy})$ denote the output features of the i^{th} MSAG in the student and teacher models, respectively. Meanwhile, the noise level of the input noisy image $I_{noisy} = I'_{noisy}$ and $I_{noisy} \neq I'_{noisy}$ when using the HADS and HMDS, respectively. Therefore, the total loss is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{RL} + \mathcal{L}_{KDL}. \quad (4)$$

4. Experiment

4.1. Datasets

Following previous works [7, 34, 38, 40, 42], we also choose the AWGN as our research object due to its extensiveness and practicality. During training, we use DIV2K [1] as our training dataset since it is a high-quality image restoration dataset. As for the test datasets, we use Set12 [36], BSD68 [20], and Urban100 [12] for grayscale image denoising and choose Kodak24 [8], CBSD68 [20], and Urban100 [12] for the color image denoising. Meanwhile, in order to further verify the effectiveness and robustness of MDRN, we utilize RN6 [14] and RN15 [14] to test the ability of MDRN for real noise removing.

4.2. Implementation Details

Model Setting. In the final version of MDRN, the number of MSAB in each MSAG is set to 8 and the input and output channels of each MSAB or MSAG are set to 64. In addition, the entire model contains three degradation processes, and the kernel size and stride are set to 2 both in the convolutional and deconvolutional layers. Meanwhile, we also introduce the self-ensemble mechanism [28] to further improve the model performance, which is denoted as MDRN+. In addition, we provide a tiny version of MDRN, which contains 2 MSABs in MSAG, named MDRN (2B).

Training Setting. During training, we randomly choose 16 noisy patches as inputs, the learning rate is initialized as 10^{-4} and halved every 100 epochs. In addition, the MDRN is implemented with the PyTorch framework, updated with the Adam optimizer, and a total of 500 epochs. It is worth noting that when the heterogeneous knowledge distillation strategy is applied to MDRN to achieve joint training, the teacher model is pre-trained and the weights are fixed throughout the process without updating.

4.3. Comparison with Classic SID Methods

In order to verify the effectiveness of the proposed MDRN, we compare it with more than 10 classic SID methods, including BM3D [4], RED30 [19], TNRD [3], IRCNN [39], DnCNN [38], MemNet [26], FFDNet [40], MLEFGN [7], MFENANN [29], and DRNe [37]. All aforementioned methods are the most widely used denoiser and all of them achieved the SOTA results at the time.

Results on Gray-Scale Images: In Table 1, we provide the PSNR results of these methods on grayscale images. According to the table, we can clearly observe that our MDRN+ and MDRN achieve the best and the second-

Method	Set12			BSD68			Urban100			
	Noise Level	$\sigma=15$	$\sigma=25$	$\sigma=50$	$\sigma=15$	$\sigma=25$	$\sigma=50$	$\sigma=15$	$\sigma=25$	$\sigma=50$
BM3D	32.37	29.97	26.72	31.08	28.57	25.62	32.34	29.70	25.94	
RED30	32.83	30.48	27.34	31.72	29.26	26.35	32.75	30.21	26.64	
TNRD	32.50	30.06	26.81	31.42	28.92	25.97	31.98	29.29	25.71	
IRCNN	32.77	30.38	27.14	31.63	29.15	26.19	32.49	29.82	26.14	
DnCNN	32.86	30.43	27.18	31.73	29.23	26.23	32.68	29.97	26.28	
FFDNet	32.75	30.43	27.32	31.63	29.19	26.29	32.42	29.92	26.52	
MLEFGN	33.04	30.66	27.54	31.81	29.34	26.39	33.21	30.64	27.22	
MFENANN	32.95	30.63	27.55	31.73	29.29	26.38	-	-	-	
DRNet	33.01	30.64	27.46	31.81	29.35	26.39	-	-	-	
MDRN (Ours)	33.06	30.67	27.56	31.83	29.36	26.41	33.22	30.67	27.24	
MDRN+ (Ours)	33.10	30.71	27.60	31.86	29.39	26.44	33.31	30.78	27.31	

Table 1. PSNR comparison with classic SID methods on grayscale images with noise levels $\sigma = 15, 25$, and 50 .

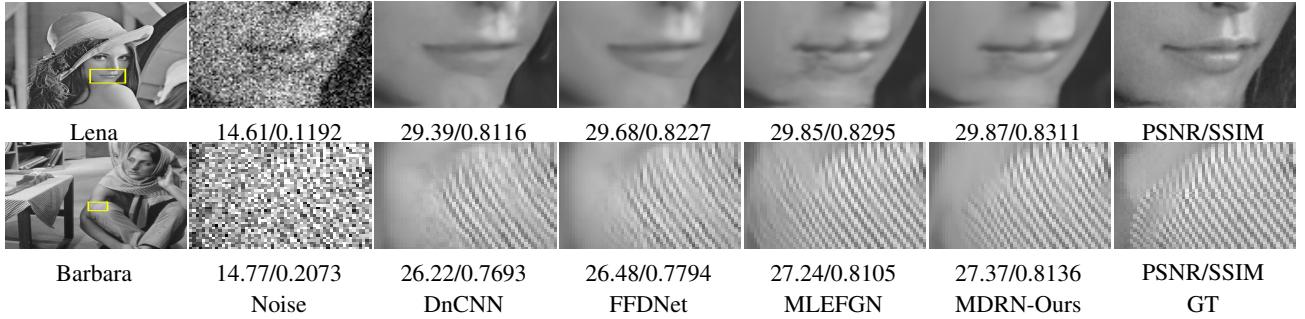


Figure 6. Visual comparison with DnCNN, FFDNet, and MLEFGN on grayscale images ($\sigma = 50$). Zoom in to view details.

best results, respectively. In Figure 6, we provide the visual comparison with DnCNN, FFDNet, and MLEFGN on grayscale images. Among them, DnCNN and FFDNet achieve competitive results and still are the most widely used denoisers. MLEFGN is a well-designed SID model, which achieved the best results under the same level parameters at the time. According to the figure, we can observe that images reconstructed by these methods still contain a lot of noises and artifacts. Moreover, denoised images become blurred since high-frequency details have been severely destroyed. On the contrary, MDRN can reconstruct high-quality denoised images with more texture details.

Results on Color Images: To further verify the validity of the model, we also provide the PSNR results of these methods on color images in Table 2. According to the table, we can observe that MDRN still achieves the best results on color images. Meanwhile, we also provide the denoised color images in Figure 7. Obviously, MDRN can reconstruct high-quality denoised images with better image edges, even compared with MLEFGN. All these results fully verified the effectiveness of the proposed MDRN.

4.4. Results with HKDS

Although our proposed MDRN has achieved excellent performance, it still does not work well under extreme conditions (tiny version models or severe noise levels). In order to improve the denoising ability of the tiny model, we pro-

pose the Heterogeneous Architecture Distillation Strategy (HADS). Meanwhile, to improve the ability of the model to process images with high noise levels, we propose the Heterogeneous Mode Distillation Strategy (HMDS). In this part, we show the model performance with these strategies.

Results with HADS. The core idea of HADS is to use a large model as Teacher to guide the tiny model to learn more accurate features, thus further improving the quality of the reconstructed images. In Tables 3 and 4, we show the PSNR results of the FFDNet and the tiny version of MDRN (2B) with and without HADS on grayscale and color images, respectively. Compared to FFDNet, MDRN (2B) can achieve better results with fewer parameters. However, we also found that this improvement is not significant. To further improve the model performance, we used the pre-trained final MDRN ($N = 8$) as the Teacher and introduced HADS during training. According to these results, we can clearly observe that with the help of HADS, the performance of MDRN (2B) can be further improved. Therefore, the new results are significantly improved compared to FFDNet. In Figure 8, we also provide the visual comparison of MDRN (2B) with and without HADS on color images. Obviously, with the help of HADS, MDRN can reconstruct cleaner images with more accurate edges. All the above experiments show that the proposed HADS is effective, which can further improve the lightweight model performance.

Results with HMDS. High noise level image denoising

Method	Kodak24			CBSD68			Urban100		
	Noise Level	$\sigma=30$	$\sigma=50$	$\sigma=70$	$\sigma=30$	$\sigma=50$	$\sigma=70$	$\sigma=30$	$\sigma=50$
CBM3D	30.89	28.63	27.27	29.73	27.38	26.00	30.36	27.94	26.31
RED30	29.71	27.62	26.36	28.46	26.35	25.08	29.02	26.40	24.74
TNRD	28.83	27.17	24.94	27.64	25.96	23.83	27.40	25.52	22.63
IRCNN	31.24	28.93	N/A	30.22	27.86	N/A	30.28	27.69	N/A
DnCNN	31.39	29.16	27.64	30.40	28.01	26.56	30.28	28.16	26.17
MemNet	29.67	27.65	26.40	28.39	26.33	25.08	28.93	26.53	24.93
FFDNet	31.39	29.10	27.68	30.31	27.96	26.53	30.53	28.05	26.39
MLEFGN	31.67	29.38	27.94	30.56	28.21	26.75	31.32	28.92	27.28
MDRN (Ours)	31.68	29.40	27.96	30.57	28.23	26.77	31.35	28.96	27.32
MDRN+ (Ours)	31.73	29.44	28.01	30.61	28.27	26.82	31.41	29.00	27.37

Table 2. PSNR comparison with classic SID methods on **color** images with noise levels $\sigma = 30, 50$, and 70 . Best results are highlighted.

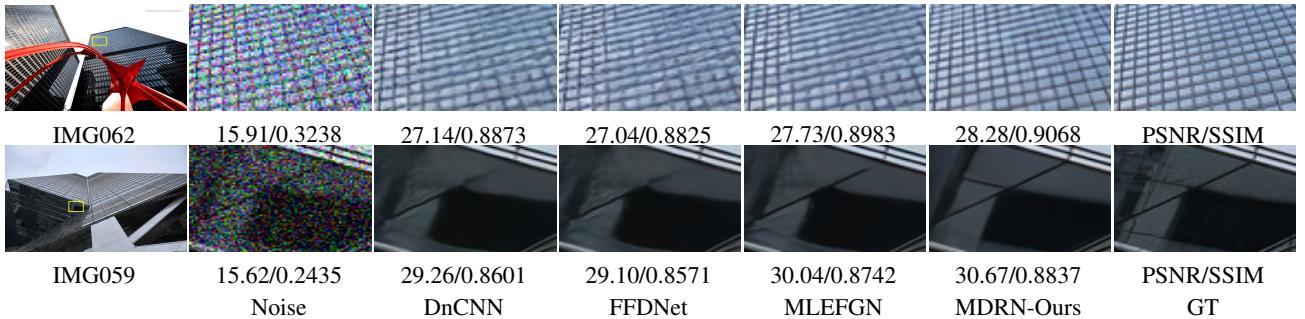


Figure 7. Visual comparison with DnCNN, FFDNet, and MLEFGN on color images ($\sigma = 50$). Zoom in to view details.

Method	Para.	Set12		BSD68		Urban100	
Noise Level	-	$\sigma=15$	$\sigma=50$	$\sigma=15$	$\sigma=50$	$\sigma=15$	$\sigma=50$
FFDNet	851K	32.75	27.32	31.63	26.29	32.42	26.52
MDRN (2B)	770K	32.87	27.35	31.72	26.31	32.82	26.74
MDRN (2B) + HADS	770K	32.96	27.43	31.78	26.40	32.98	26.84

Table 3. PSNR results of MDRN with and without the HADS on grayscale images. Best results are highlighted.

Method	Para.	Kodak24		CBSD68		Urban100	
Noise Level	-	$\sigma=30$	$\sigma=70$	$\sigma=30$	$\sigma=70$	$\sigma=30$	$\sigma=70$
FFDNet	851K	31.39	27.68	30.31	26.53	30.53	26.39
MDRN (2B)	770K	31.40	27.69	30.38	26.59	30.66	26.57
MDRN (2B) + HADS	770K	31.47	27.73	30.42	26.63	30.78	26.64

Table 4. PSNR results of MDRN with and without the HADS on color images. Best results are highlighted.

is still a challenging task since noise will severely damage the image. Recently, the most common method to solve this problem is to build larger CNN models. Different from previous works, we propose HMDS to enable the model to learn more accurate high-frequency details from the Teacher. Specifically, according to the core idea of the strategy, we use the model trained under $\sigma = 25$ or $\sigma = 50$ as the teacher model to guide the model trained for $\sigma = 50$ or $\sigma = 70$. In Tables 5, we show the PSNR results of MDRN with and without HMDS. According to the results, we can observe that (1). compared to MLEFGN [7], MDRN can achieve better results with fewer parameters; (2). with

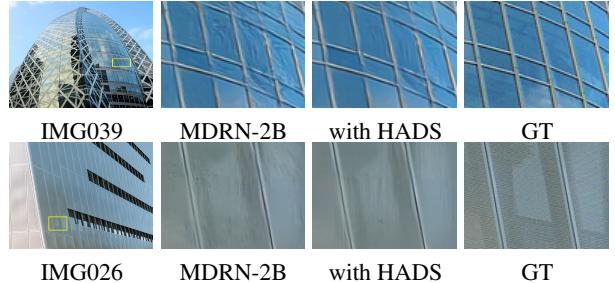


Figure 8. Visual comparison of MDRN (2B) with and without HADS on color images ($\sigma = 70$). Zoom in to view details.

the help of HMDS, the performance of our MDRN can be further improved. Similarly, we also provide the visual comparison of MDRN with and without HMDS on color images in Figure 9. Obviously, when the image suffers from severe noise interference, it is extremely difficult to restore clean images. Fortunately, with the help of HMDS, MDRN can reconstruct more clean and accurate denoised images. All aforementioned experiments fully demonstrate the effectiveness of the proposed HMDS.

4.5. Study on High Noise Level

High noise level image denoising is still a challenging task since noise will severely damage images. In this part, we provide more denoising results of high-noise images in Figure 10. According to the results, we can observe that

Method	Parm.	Set12	BSD68	Urban100
Noise Level	-	$\sigma=50$	$\sigma=50$	$\sigma=50$
MLEFGN	6.86M	27.54	26.39	27.22
MDRN	2.38M	27.56	26.41	27.24
MDRN + HMDS (25)	2.38M	27.63	26.46	27.38
Method	Parm.	Kodak24	CBSD68	Urban100
Noise Level	-	$\sigma=70$	$\sigma=70$	$\sigma=70$
MLEFGN	6.86M	27.94	26.75	27.28
MDRN	2.38M	27.96	26.77	27.32
MDRN + HMDS (50)	2.38M	28.00	26.81	27.45

Table 5. PSNR results of MDRN with and without HMDS on grayscale and color images, respectively.

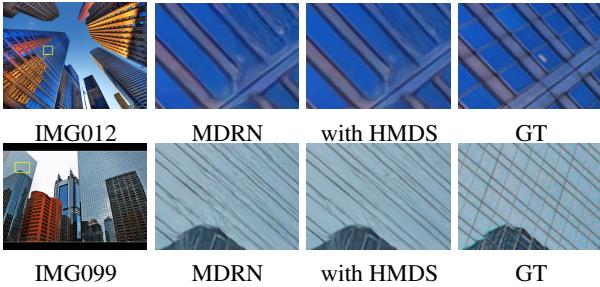


Figure 9. Visual comparison of MDRN with and without HMDS (50) on color images ($\sigma = 70$). Zoom in to view details.

MLEFGN is still difficult to reconstruct high-quality denoised images under the high noise level even though it is an excellent SID model. On the contrary, our MDRN can reconstruct clearer and more accurate denoised images, even the model parameters of MDRN is only one-third of MLEFGN. This is because MDRN can extract rich multi-scale image features, thus it can reconstruct more accurate denoised images.

4.6. Study on Model Size

Recently, some outstanding SID models have been proposed, such as RDN [43], RNAN [42], DIDN [34], and DHDN [22]. However, it cannot be ignored that the parameters of these models are ten or even hundred times that of our MDRN. In this paper, we provide a lightweight MDRN, which achieves a good balance between the size and performance of the model. However, we should notice that the model size of MDRN can be easily changed by changing the number of MSAB (N) in MSAG. In Figure 11, we provide the model performance with different N . According to the figure, we can find (1). as N increases, the model performance can be further improved; (2). when $N = 20$, MDRN exceeds RNAN with fewer parameters; (3). when $N = 40$, MDRN exceeds RDN with only half of the parameters. Meanwhile, MDRN achieves close performance to DIDN with only one-fifteenth of the parameters. This fully demonstrates the effectiveness and potential of MDRN.

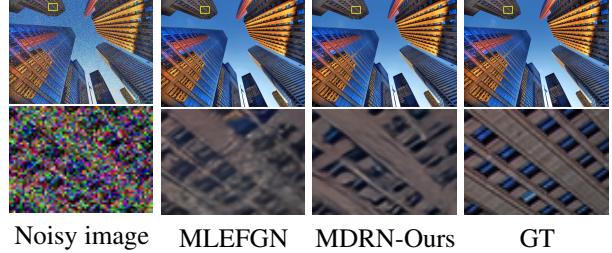


Figure 10. Visual comparison of the denoising results of MDRN and MLEFGN on high-noise images (noise level: $\sigma = 70$).

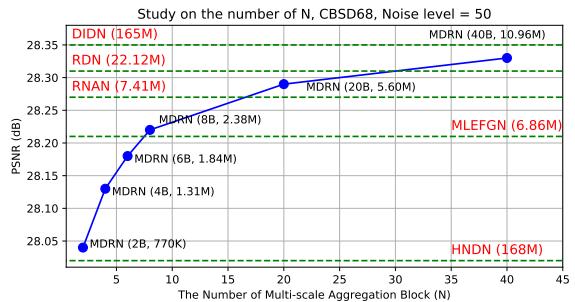


Figure 11. Study on the number (N) of MSAB in each MSAG.

Moreover, we provide a more intuitive comparison in Figure 1. It is worth noting that the point in the upper left corner of the figure represents a better balance between model size and performance. According to the figure, we can clearly observe that our MDRN achieves the best trade-off between model size and performance.

5. Conclusion

In this paper, we proposed a lightweight and efficient Multiple Degradation and Reconstruction Network (MDRN) for SID, which realizes automatic noise removal and feature restoration through multiple down- and up-sampling operations. Meanwhile, we proposed two novel strategies, namely Heterogeneous Architecture Distillation Strategy (HADS) and Heterogeneous Mode Distillation Strategy (HMDS) to improve the performance of tiny models and the model under high noise levels, respectively. Extensive experiments have fully proved the effectiveness of the proposed model and strategies.

6. Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grant 2021YFE0203700, Grant NSFC/RGC N_CUHK 415/19, Grant ITF MHP/038/20, Grant RGC 14300219, Grant CRF 8730063, in part by the National Natural Science Foundation of China under Grant 61871185, and in part by the Shanghai Rising-Star Program under Grant 21QA1402500.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 5
- [2] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, volume 2, pages 60–65, 2005. 2
- [3] Y. Chen and T Pock. Trainable nonlinear reaction diffusion: a flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1256–1272, 2017. 5
- [4] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 5
- [5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1
- [6] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin Li. Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, 22(4):1620–1630, 2013. 2
- [7] Faming Fang, Juncheng Li, Yiting Yuan, Tieyong Zeng, and Guixu Zhang. Multilevel edge features guided network for image denoising. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 1, 2, 5, 7
- [8] Rich Franzen. Kodak lossless true color image suite. *source: http://r0k.us/graphics/kodak*, 4, 1999. 5
- [9] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):1397–1409, 2013. 2
- [10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2
- [11] Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *CVPR*, pages 3462–3471, 2020. 3, 4
- [12] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 5
- [13] Rushi Lan, Haizhang Zou, Cheng Pang, Yanru Zhong, Zhenbing Liu, and Xiaonan Luo. Image denoising via deep residual convolutional neural networks. *Signal, Image and Video Processing*, 15(1):1–8, 2021. 1, 2
- [14] Marc Lebrun, Miguel Colom, and Jean Michel Morel. The noise clinic: a blind image denoising algorithm. *Image Processing on Line*, 5:1–54, 2015. 5
- [15] Huixia Li, Chenqian Yan, Shaohui Lin, Xiawu Zheng, Baochang Zhang, Fan Yang, and Rongrong Ji. Pams: Quantized super-resolution via parameterized max scale. In *ECCV*, 2020. 4
- [16] Juncheng Li, Faming Fang, Jiaqian Li, Kangfu Mei, and Guixu Zhang. Mdcn: Multi-scale dense cross network for image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2547–2561, 2020. 3
- [17] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, pages 517–532, 2018. 3
- [18] Pengju Liu, Hongzhi Zhang, Wei Lian, and Wangmeng Zuo. Multi-level wavelet convolutional neural networks. *IEEE Access*, 7:74973–74985, 2019. 1, 2
- [19] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NeurIPS*, pages 2802–2810, 2016. 5
- [20] David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*, 2001. 5
- [21] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1
- [22] Bumjun Park, Songhyun Yu, and J. Jeong. Densely connected hierarchical network for image denoising. *CVPRW*, pages 2104–2113, 2019. 1, 8
- [23] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2
- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3
- [25] SMA Sharif, Rizwan Ali Naqvi, and Mithun Biswas. Learning medical image denoising with deep dynamic residual attention network. *Mathematics*, 8(12):2192, 2020. 1, 2
- [26] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, pages 4539–4547, 2017. 5
- [27] Wesley Tansey, Jesse Thomason, and James G Scott. Maximum-variance total variation denoising for interpretable spatial smoothing. pages 2460–2467, 2018. 2
- [28] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, pages 1865–1873, 2016. 5
- [29] Yi Wang, Xiao Song, Guanghong Gong, and Ni Li. A multi-scale feature extraction-based normalized attention neural network for image denoising. *Electronics*, 10(3):319, 2021. 5
- [30] Haiyan Wu, Jing Liu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Knowledge transfer dehazing network for nonhomogeneous dehazing. In *CVPRW*, pages 478–479, 2020. 4
- [31] Dong Yang and Jian Sun. BM3D-Net: a convolutional neural network for transform-domain collaborative filtering. *IEEE Signal Processing Letters*, 25(1):55–59, 2018. 1, 2
- [32] Qiaosi Yi, Juncheng Li, Faming Fang, Aiwen Jiang, and Guixu Zhang. Efficient and accurate multi-scale topological network for single image dehazing. *IEEE Transactions on Multimedia*, 2021. 1
- [33] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, pages 4133–4141, 2017. 3

- [34] Songhyun Yu, Bunjun Park, and J. Jeong. Deep iterative down-up cnn for image denoising. *CVPRW*, pages 2095–2103, 2019. 1, 2, 5, 8
- [35] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2
- [36] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *ICCS*, pages 711–730, 2010. 5
- [37] Jiahong Zhang, Yonggui Zhu, Wenyi Li, Wenlong Fu, and Lihong Cao. Drnet: A deep neural network with multi-layer residual blocks improves image denoising. *IEEE Access*, 9:79936–79946, 2021. 2, 5
- [38] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 1, 2, 5
- [39] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, pages 3929–3938, 2017. 5
- [40] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 1, 2, 5
- [41] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 3
- [42] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *ICLR*, 2019. 5, 8
- [43] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 8
- [44] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 2
- [45] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *zhou2018unet*, pages 3–11. 2018. 2