FREQUENCY-RELEVANT RESIDUAL LEARNING FOR MULTI-MODAL IMAGE DENOISING

*Xiongwei Liu, Zehua Sheng, and Hui-Liang Shen**

College of Information Science and Electronic Engineering, Zhejiang University, China

ABSTRACT

Recently, multi-modal image processing has shown its great potential in boosting denoising performance in terms of both accuracy and visual quality. However, current studies generally face the challenge of achieving a good balance between noise removal and detail preservation. In this work, we introduce patch-wise frequency decomposition into convolutional neural networks and propose a novel multi-modal image denoising (MID) algorithm. Integrating the frequency-domain information of images from different modalities, our network first predicts a frequency-relevant residual and then regresses the denoised result using a learnable reconstruction kernel. Benefiting from the distinctive properties of noise and true signals as well as the correlation between multi-modal images in the frequency domain, the proposed algorithm can effectively remove noise and simultaneously reconstruct fine details. Extensive experiments demonstrate the superiority and generalizability of our algorithm over state-of-the-art competing algorithms on various MID tasks, including near-infrared guided RGB image denoising, flash guided no-flash image denoising, and RGB guided depth image denoising. Code is available at `https://github.com/liuxw11/FRL`.

***Index Terms*—** Multi-modal image denoising, residual learning, frequency decomposition, spatial reconstruction

1. INTRODUCTION

Due to the inevitable noise corruption in modern camera systems, image denoising is a long-standing topic in computer vision. Recently, the rapid development of deep learning theory significantly enhances the denoising performance [1, 2, 3], in terms of both accuracy and efficiency. But it's still difficult to restore fine structures and eliminate artifacts based only on the noisy input, especially at high noise levels.

To address this problem, one popular trend is to denoise the noisy target image with the help of its well-aligned guidance image [4, 5, 6, 7, 8, 9]. Generally, the guidance image is captured from the same scene with a high signal-to-noise ratio but usually have a different modality. Common multi-modal image denosing (MID) tasks include RGB guided depth image denoising, near-infrared (NIR) guided RGB image denoising, flash guided no-flash image denoising, *etc*.

Among previous studies on multi-modal image restoration, CUNet [6] splits the common and unique features between different modalities for both restoration and fusion tasks. DKN [7] regresses the filtering results by learning spatially variant kernels for each pixel. However, these methods tend to over-smooth weak details. Guided filtering [10] based algorithms such as SVLRM [8] and UMGF [9] aim to explore a linear representation model of the guidance images to estimate the high-quality versions of target images. Generally, these algorithms can reconstruct sharp edges and clear details but suffer from texture copying artifacts.

In this work, we introduce frequency decomposition into convolutional neural networks and propose a novel framework for MID tasks. Using patch-wise 2D discrete cosine transform (2D-DCT), we decompose the input image pair into two frequency tensors. Based on this, our network predicts a multi-channel frequency-relevant residual, modeling the desired information for both noise removal and detail enhancement. The final denoised images are reconstructed by fusing the predicted residuals and the input target frequency tensors with a learnable convolutional kernel. This decomposition-reconstruction architecture enables the network to more effectively distinguish between noise and true signals based on their distinctive properties in the frequency domain. It also helps reconstruct clear edges and details due to the frequency correlation between the input pairs. Hence, the proposed algorithm well balances noise removal and detail preservation.

The contributions of this work are as follows: (1) We propose a frequency-relevant neural network for MID tasks, which can remove noise while preserving fine details. (2) We introduce a flexible and effective reconstruction module that fuses the predicted residual features and the frequency tensors with a learnable convolution kernel. (3) Our algorithm achieves state-of-the-art performance on various MID tasks.

2. PROPOSED ALGORITHM

2.1. Problem Formulation

Let $\mathbf{Y} \in \mathbb{R}^{H \times W}$ denote a noisy image corrupted by additive Gaussian noise. The corresponding clean image and the noise

*Corresponding author.

component are denoted as $\mathbf{X}$, $\mathbf{N} \in \mathbb{R}^{H \times W}$, respectively. For pixel r, the noise model can be formulated as

$$\mathbf{Y}(r) = \mathbf{X}(r) + \mathbf{N}(r), \tag{1}$$

where $\mathbf{N}(r) \sim \mathcal{N}(0, \sigma^2)$ follows a Gaussian distribution.

Traditional frequency-domain single image denoising algorithms basically divide $\mathbf{Y}$ into multiple overlapped patches on which denoising is processed independently [11, 12]. Since the frequency coefficients of clean patches can achieve good sparsity while those of noise cannot, it's easier to split them in the frequency domain by threshold shrinkage or filtering.

Inspired by this, we apply frequency decomposition to the multi-modal image denoising tasks. Given the noisy target image $\mathbf{Y}$ and its clean guidance image $\mathbf{G}$, we extract two patches $\mathbf{y}$, $\mathbf{g} \in \mathbb{R}^{s \times s}$ from this image pair at the same position, respectively. Despite their modality difference, $\mathbf{y}$ and $\mathbf{g}$ are expected to share many similar features. In other words, based on the sparse coding theory [13], $\mathbf{y}$ and $\mathbf{g}$ can be linearly represented using the same set of atoms that record typical patterns if their contents are correlated. Since the frequency transformation is also linear, this linear representation model still holds in the frequency domain. Hence, the frequency characteristics of the guidance image can help to distinguish whether the frequency coefficients of the noisy target image are contributed by noise or weak details.

In this work, we propose to learn an implicit denoising function $\mathcal{F}(\cdot)$ that takes both the target and guidance frequency tensors as input using a convolutional neural network. Incorporated with frequency decomposition, the clean target image $\mathbf{X}$ can be estimated by

$$\hat{\mathbf{X}} = \mathcal{K}(\mathcal{T}(\mathbf{Y}) + \mathcal{F}(\mathcal{T}(\mathbf{Y}), \mathcal{T}(\mathbf{G}))), \tag{2}$$

where $\mathcal{T}(\cdot)$ is a patch-wise frequency decomposition function, $\mathcal{K}(\cdot)$ is the spatial reconstruction function. $\mathcal{F}(\cdot, \cdot)$ outputs the residual for denoising.

2.2. Network Architecture

Based on the denoising model in (2), our network is composed of three components, *i.e.*, the frequency decomposition module $\mathcal{T}(\cdot)$, the residual learning module $\mathcal{F}(\cdot, \cdot)$, and the spatial reconstruction module $\mathcal{K}(\cdot)$. The network architecture is shown in Fig. 1.

First of all, the input images $\mathbf{Y}$ and $\mathbf{G}$ are transformed into two frequency tensors by $\mathcal{T}(\cdot)$. Specifically, $\mathcal{T}(\cdot)$ divides the input image into overlapped $s \times s$ patches which are then transformed into the frequency domain using 2D-DCT, producing frequency tensor with shape $H \times W \times s^2$. It's implemented as a convolutional layer with fixed weights initialized by DCT kernels. In this work, to achieve a good balance between receptive field and computational cost, we set $s = 7$.

In the residual learning module, two frequency tensors $\mathcal{T}(\mathbf{Y})$ and $\mathcal{T}(\mathbf{G})$ are first fed into two identical encoders

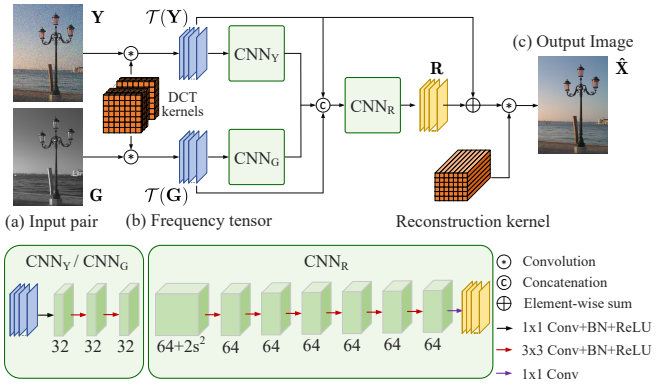


Fig. 1. Architecture of the proposed network.

$\mathbf{CNN_Y}$ and $\mathbf{CNN_G}$, respectively. The output feature maps are usually more expressive than the input frequency tensors, but the perception of fine details from high frequencies may be lost during feature aggregation. Therefore, we concatenate the input frequency tensors and the output feature maps together and feed them into $\mathbf{CNN_R}$ for regressing the frequency relevant residual $\mathbf{R}$, where $\mathbf{R} = \mathcal{F}(\mathcal{T}(\mathbf{Y}), \mathcal{T}(\mathbf{G}))$.

Finally, we reconstruct the denoised target image $\hat{\mathbf{X}}$ from $\mathcal{T}(\mathbf{Y}) + \mathbf{R}$ with a single convolutional kernel. The kernel's weights are initialized by the inverse 2D-DCT (2D-IDCT) kernels, for it's difficult to learn the relationship between the frequency and the spatial domain from scratch. During training process, the kernel is learnable to integrate information from both shallow frequency tensors and deep feature maps.

2.3. Reconstruction From Frequency-Relevant Residual

In this section, we discuss the reconstruction process and the effectiveness of residual learning in the frequency domain. Here, it's natural to adopt standard 2D-IDCT as $\mathcal{K}(\cdot)$ to transform the denoised frequency tensor to the spatial domain. However, by rewriting (2) as

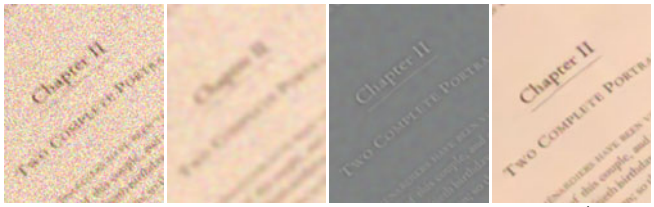
$$\hat{\mathbf{X}} = \mathcal{K}(\mathcal{T}(\mathbf{Y})) + \mathcal{K}(\mathcal{F}(\mathcal{T}(\mathbf{Y}), \mathcal{T}(\mathbf{G}))), \tag{3}$$

we see that $\mathcal{K}(\mathcal{T}(\mathbf{Y}))$ is an identical mapping and has no contribution to noise removal in this case. In addition, the learned residual is expected to be the frequency coefficients of noise for each channel, where the flexibility of convolutional neural networks in representation cannot be fully utilized.

Considering this, we allow the reconstruction kernel to be learnable in the training stage. As visualized in Fig. 2, the reconstruction module estimates a base component of the denoised image from the noisy frequency tensor $\mathcal{T}(\mathbf{Y})$. Both high-frequency noise and details are smoothed. Correspondingly, the contributions of frequency-relevant residual mainly lie in two aspects, *i.e.*, the enhancement of high-frequency details and the removal of low-frequency noise. Thanks to the explicit frequency decomposition of the image pair, it's

Algorithms	RGB/NIR		Flash/No-Flash		RGB/Depth	
	$\sigma = 25$	$\sigma = 50$	$\sigma = 25$	$\sigma = 50$	$\sigma = 25$	$\sigma = 50$
DKN [7]	30.77 / 0.8063	28.23 / 0.7194	33.67 / 0.8692	30.95 / 0.7983	40.32 / 0.9664	36.80 / 0.9400
CUNet [6]	31.40 / 0.8414	28.59 / 0.7707	34.38 / 0.9000	30.32 / 0.8336	40.90 / 0.9759	37.88 / 0.9648
UMGF [9]	31.13 / 0.8303	26.96 / 0.7552	32.81 / 0.8739	31.60 / 0.8602	40.47 / 0.9728	37.40 / 0.9625
SVLRM [8]	31.25 / 0.8469	29.12 / 0.8015	34.09 / 0.8911	32.02 / 0.8626	40.61 / 0.9739	36.85 / 0.9560
Ours	**32.87 / 0.8760**	**30.22 / 0.8230**	**35.22 / 0.9081**	**32.79 / 0.8780**	**41.93 / 0.9790**	**38.24 / 0.9654**

Table 1. PSNR (dB) / SSIM values of RGB/NIR, RGB/depth and flash/no-flash denoising results at noise levels $\sigma = 25, 50$.



(a) $\mathbf{Y}$ (b) $\mathcal{K}(\mathcal{T}(\mathbf{Y}))$ (c) $\mathcal{K}(\mathbf{R})$ (d) $\hat{\mathbf{X}}$

Fig. 2. An example of flash/no-flash denoising result. (a) Input noisy target image. (b) Reconstruction from the target frequency tensor. (c) Reconstruction from the learned residual, normalized for visualization. (d) Output image.

straightforward to conduct these two tasks simultaneously by referring to different frequency components of the guidance image. Therefore, our algorithm can achieve excellent visual quality in terms of noise removal and artifact alleviation especially around flat areas, while restoring sharp edges and fine details at the same time.

3. EXPERIMENTS

3.1. Experimental Settings

To evaluate the performance of our algorithm, we conduct experiments on three different multi-modal image denoising tasks, *i.e.*, NIR guided RGB image denoising, flash guided no-flash image denoising, and RGB guided depth image denoising.

Datasets. For RGB/NIR and flash/no-flash denoising tasks, the experiments are conducted on the RGB-NIR Scene Dataset [14] and the Flash and Ambient Illuminations Dataset [15], respectively. Both datasets are randomly divided into two subsets for training (70%) and evaluation (30%). The NYU v2 Dataset [16] is chosen for RGB/depth image denoising. Following [7], the first 1000 image pairs are used for training while the remaining 449 ones are used for evaluation.

Training details. In the training process, we randomly crop the input images into 128×128 patch pairs. The target images are added with white Gaussian noise. Our network is trained using the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1 \times 10^{-4}$. We adopt the ℓ_1 loss function defined as $\mathcal{L}(\mathbf{X}, \hat{\mathbf{X}}) = ||\mathbf{X} - \hat{\mathbf{X}}||_1$, and train the network for about 4×10^4 iterations with batch size 24. The learning rate is initially set to 1×10^{-4} and decayed to 5×10^{-5} when half of the iterations are completed.

	RGB/NIR	Flash/No-Flash	RGB/Depth
w/o frequency decomposition	29.77 / 0.8120	31.98 / 0.8605	37.12 / 0.9541
w/o learnable reconstruction	30.00 / 0.8162	32.17 / 0.8627	37.66 / 0.9606
full architecture	**30.22 / 0.8230**	**32.79 / 0.8780**	**38.24 / 0.9654**

Table 2. Ablation study results of PSNR (dB)/ SSIM for RGB/NIR, RGB/depth, flash/no-flash denoising at $\sigma = 50$.

3.2. Evaluation and Comparison

We quantitatively evaluate our algorithm on three denoising tasks under two different noise levels, *i.e.*, $\sigma = 25,\ 50$. In addition, we also compare it to the state-of-the-art multi-modal image restoration algorithms including CUNet [6], DKN [7], SVLRM [8] and UMGF [9]. Two metrics, PSNR and SSIM, are used to assess the denoising performance. Table 1 lists the average PSNR and SSIM values obtained by different algorithms on the testing sets, demonstrating that our algorithm achieves the highest accuracy on all three MID tasks.

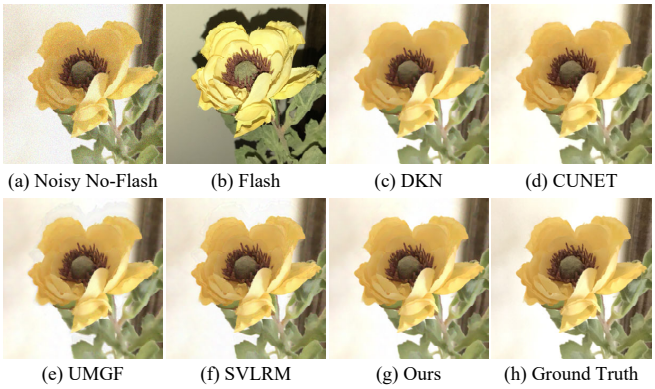
For RGB/NIR denoising tasks, as Fig. 3 shows, our algorithm restores fine details and flat regions without artifacts at the same time. Particularly, around texture-rich areas that are mixed with flat contents and complicated structures, as marked by red boxes in Fig. 3, the competing methods either produce over-smoothed results or cannot remove noise completely. In comparison, ours achieves a good balance.

For flash/no-flash denoising, our algorithm can effectively handle structural inconsistencies typically caused by the shadows in the flash images. In comparison, as Fig. 4 shows, texture copying artifacts are noticeable in the denoised results obtained by SVLRM and UMGF. For RGB/depth denoising, Fig. 5 shows that we produce more distinguishable edges than other competing methods.

(a) Noisy RGB (b) NIR (c) DKN (d) CUNET

(e) UMGF (f) SVLRM (g) Ours (h) Ground Truth

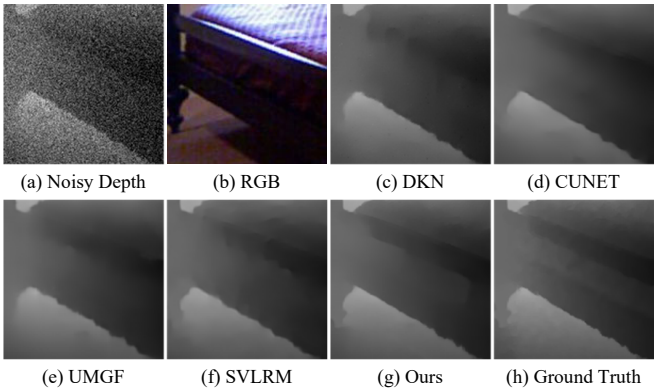
Fig. 3. Denoising results of RGB/NIR image pairs at noise level $\sigma = 25$.



(a) Noisy No-Flash (b) Flash (c) DKN (d) CUNET

(e) UMGF (f) SVLRM (g) Ours (h) Ground Truth

Fig. 4. Denoising results of flash/no-flash pairs at $\sigma = 25$.



(a) Noisy Depth (b) RGB (c) DKN (d) CUNET

(e) UMGF (f) SVLRM (g) Ours (h) Ground Truth

Fig. 5. Denoising results of RGB/depth pairs at $\sigma = 25$.

3.3. Ablation Studies

We conduct ablation studies for the proposed network architecture, which mainly focuses on the effectiveness of the frequency decomposition and the learnable reconstruction module. By replacing the DCT kernels with normal convolutional kernels and modifying the skip connection, the model is converted to a spatial residual learning network (without frequency decomposition). By replacing the reconstruction module with standard 2D-IDCT, the model only regresses frequency coefficients of the noise (without learnable reconstruction). Though the model sizes are kept the same for the two situations, we observe a significant performance drop as Table 2 shows. Hence, both of them contribute to the performance improvement as analyzed in Section 2.

4. CONCLUSIONS

In this work, we propose a novel multi-modal image denoising network based on residual learning in the frequency domain. The introduction of frequency decomposition helps the network to simultaneously accomplish noise removal and structure restoration by referring to different frequency components of the guidance image. Further, to regress the final denoised results, we adopt a learnable reconstruction block to fuse the predicted residual features and the frequency tensors. Experimental results show that, compared to the state-of-the-art multi-modal image restoration approaches, our algorithm obtains the highest denoising accuracy on various MID tasks. It also achieves the best visual quality in terms of both detail preservation and artifacts alleviation.

5. REFERENCES

[1] 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*, vol. 26, no. 7, pp. 3142–3155, 2017.

[2] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018.

[3] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Toward convolutional blind denoising of real photographs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1712–1722.

[4] Pingfan Song and Miguel RD Rodrigues, "Multimodal image denoising based on coupled dictionary learning," in *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2018, pp. 515–519.

[5] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Joint image filtering with deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1909–1923, 2019.

[6] Xin Deng and Pier Luigi Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[7] Beomjun Kim, Jean Ponce, and Bumsub Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 579–600, 2021.

[8] Jiangxin Dong, Jinshan Pan, Jimmy Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang, "Learning spatially variant linear representation models for joint filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 01, pp. 1–1, 2021.

[9] Zenglin Shi, Yunlu Chen, Efstratios Gavves, Pascal Mettes, and Cees G. M. Snoek, "Unsharp mask guided filtering," *IEEE Transactions on Image Processing*, vol. 30, pp. 7472–7485, 2021.

[10] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[11] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[12] Guoshen Yu and Guillermo Sapiro, "DCT image denoising: a simple and effective image denoising algorithm," *Image Processing On Line*, vol. 1, pp. 292–296, 2011.

[13] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[14] Matthew Brown and Sabine Süsstrunk, "Multi-spectral SIFT for scene category recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 177–184.

[15] Yagiz Aksoy, Changil Kim, Petr Kellnhofer, Sylvain Paris, Mohamed Elgharib, Marc Pollefeys, and Wojciech Matusik, "A dataset of flash and ambient illumination pairs from the crowd," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 634–649.

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the European Conference on Computer Vision*. Springer, 2012, pp. 746–760.