https://doi.org/10.2352/EI.2025.37.10.IPAS-227

Multi-Scale Feature Matching for Image Denoising using Residual Swin Transformers

Muqudas Rafiq, Ahsan Jalil, Khurram Usman, Muhammad Abdullah, Bilal Zafar; 10xEngineers Inc.; Irvine, CA

Abstract

Image denoising is a crucial task in image processing, aiming to enhance image quality by effectively eliminating noise while preserving essential structural and textural details. In this paper, we introduce a novel denoising algorithm that integrates residual Swin transformer blocks (RSTB) with the concept of the classical non-local means (NLM) filtering. The proposed solution is aimed at striking a balance between performance and computation complexity and is structured into three main components: (1) Feature extraction utilizing a multi-scale approach to capture diverse image features using RSTB, (2) Multi-scale feature matching inspired by NLM that computes pixel similarity through learned embeddings enabling accurate noise reduction even in high-noise scenarios, and (3) Residual detail enhancement using the swin transformer block that recovers high-frequency details lost during denoising. Our extensive experiments demonstrate that the proposed model with 743k parameters achieves the best or competitive performance amongst the state-of-the-art models with comparable number of parameters. This makes the proposed solution a preferred option for applications prioritizing detail preservation with limited compute resources. Furthermore, the proposed solution is flexible enough to adapt to other image restoration problems like deblurring and super-resolution.

Introduction

Image denoising is an essential step in many image processing and computer vision applications, as noise can significantly degrade the quality of visual data and affect the performance of downstream tasks such as segmentation, recognition, and classification [1, 2]. The goal of denoising is to remove unwanted noise from an image while preserving important structures and details. Traditionally, various filtering methods such as Gaussian, median, and non-local means (NLM) filters have been widely used for image denoising [1]. With the rise of deep learning, various convolutional neural networks (CNNs) [3, 4] and Transformer-based architectures [6, 8, 9] have outperformed the classical methods both in terms of noise removal and detail preservation.

In this paper, we propose a novel feature-matching-based denoising algorithm that combines the strengths of CNNs and transformers. Inspired by the NLM [1] filtering approach, our model leverages the concept of similarity-based feature matching. Unlike traditional NLM that operates on raw pixel intensities, our model utilizes learned feature embeddings from Swin Transformer blocks [9] which significantly improve noise reduction capabilities by capturing both local and global contexts effectively. This combination of NLM with multi-scale transformer-based feature extraction sets our work apart from prior multi-scale methods that typically rely on purely convolutional architectures, thus providing greater robustness to diverse noise patterns and better preservation of structural details. As demonstrated by our experiments, the proposed approach finds a balance between computational complexity and denoising performance achieving the best results among models with comparable parameters. The proposed solution performs particularly well in scenarios with a lot of high frequency content and repetitive structures. The main contributions of the proposed solution are:

- We introduce a multi-scale feature extraction framework based on the use of residual Swin transformer blocks (RSTB) [9] which can capture diverse local and global features across the input image.
- We propose a multi-scale feature matching block (MS-FMB) based on the classical NLM filtering [1] that improves the accuracy of pixel similarity estimation by utilizing feature vectors and enables robust denoising performance even at higher noise levels.
- We present a residual detail enhancement module based on Swin transformer blocks that helps recover lost fine details, resulting in enhanced image quality for scenarios with high frequency content.
- Our extensive experiments on benchmark datasets show that the proposed method achieves better or competitive performance in terms of the peak signal-to-noise ratio (PSNR) in comparison to other state-of-the-art (SOTA) models of comparable size. Furthermore, the proposed solution qualitatively performs better in scenarios with repetitive structures.

In the rest of the paper, we first provide an overview of prior work on denoising followed by a detailed description of the proposed model architecture and methodology. We then present detailed quantitative and qualitative experimental results highlighting various aspects of the proposed solution.

Related Work

Image denoising has been extensively studied over the past few decades, leading to a broad range of approaches, from classical filtering techniques to modern deep learning-based methods. Traditional denoising algorithms, such as Gaussian smoothing, median filtering, and bilateral filtering, were initially employed to remove noise by averaging pixel intensities while preserving edges [11, 12]. Although effective for low noise levels, these methods often fail to preserve finer image details, especially under high noise conditions.

A significant breakthrough in denoising came with the introduction of NLM filtering, which computes the similarity between patches in an image and averages them based on this similarity [1, 2]. The success of NLM lies in its ability to exploit the self-similarity in natural images, though it tends to be computationally expensive and struggles with complex textures. Following NLM, Block-Matching and 3D Filtering (BM3D) [18] and Weighted Nuclear Norm Minimization (WNNM) [19] advanced denoising by utilizing patch-based collaborative filtering in a 3D transform domain, providing improved results, especially in textured regions. These methods laid the foundation for subsequent multi-scale and patch-based feature extraction techniques.

With the advent of deep learning, CNNs have been widely adopted for image restoration tasks, including denoising. One of the earliest CNN-based denoising models demonstrated that CNNs could outperform traditional filters by learning data-driven features [13]. This was significantly improved by the introduction of DnCNN [3], a deep residual learning-based model that facilitated the learning process by focusing on residual noise. The success of DnCNN led to more advanced models like MemNet [14] and FFDNet [15], which further optimized multi-layer CNNs to handle varying noise levels. The encoder-decoder-based U-Net architecture [5], originally developed for biomedical segmentation, has also been effectively adapted for image denoising due to its ability to capture both local and global features through its symmetrical structure. Extensions of this concept, RED-Net [16] and the Residual Dense Network [17], utilized similar encoder-decoder designs with residual connections. These architectures leverage skip connections to preserve fine-grained details, effectively enhancing denoising performance.

The idea of leveraging self-similarity, as utilized in NLM and BM3D, was later incorporated into deep learning frameworks. Techniques like the non-local recurrent network [22] and non-local blocks [23] applied non-local relationships directly in the feature space, improving robustness under high noise levels. Moreover, multi-scale processing became a prominent direction, with early methods like BM3D and WNNM inspiring deep learning models that utilize multi-scale feature extraction. Variational Denoising Network [20] and Real Image Denoising Network [21] incorporated these ideas into their deep architectures, improving their ability to capture both fine and coarse features, thus achieving better denoising results.

More recently, transformers and attention mechanisms, originally developed for natural language processing tasks [8], have been successfully adapted for image denoising. Vision Transformers demonstrated that attention-based models could rival CNNs for tasks such as image classification [6]. Building on this foundation, SwinIR [9], a transformer-based architecture that effectively captures both local and global contextual information through hierarchical attention mechanisms demonstrated SOTA denoising performance. DeamNet [7] employed deep feature extraction along with attention mechanisms to focus on more relevant areas of an image for noise removal. Although effective, these models often require considerable computational resources making their deployment difficult in resource constrained applications.

Our work builds upon these advancements by proposing a hybrid approach that integrates multi-scale feature matching and residual detail enhancement with a transformer-based architecture.

Proposed Network Model

In this section, we present a detailed description of the proposed model illustrated in Fig. 1 followed by the training details. As depicted in Fig. 1, the proposed model architecture consists of three main components: 1) Feature Extraction, 2) Multi-scale Feature Matching, and 3) Residual Detail Enhancement. Each of these components is described in detail below.

Feature Extraction Block

In the first part of the model, we perform shallow feature extraction independently for each input image channel using RSTBs [9] to learn feature embeddings at multiple scales. Swin transformers utilize hierarchical feature extraction and local window-based attention, making them efficient in capturing both local and global features, effectively distinguishing signal from noise. With x, y denoting the pixel positions and c the image channels, the noisy grayscale/color image $f(x,y,c)$ is given as input to the RSTB layers.

The RSTB performs convolutions at multiple scales, allowing the model to learn feature vectors that encode both structural and textural details in the image. The Swin Transformer layers (STL) leverage an attention mechanism that further enhances the ability to differentiate between meaningful features and noise. The output of the feature extraction block, a multi-channel feature map $M(x,y,z)$, representing the different learned features extracted from the noisy image corresponding to each pixel along the z dimension is given by

$$M(x,y,z) = U(f(x,y)). \tag{1}$$

Multi-scale Feature Matching Block

In the second part of the proposed model, we utilize the learned feature vectors $M(x,y,z)$ to determine the similarity between pixels directly in the spatial domain. Specifically, for each pixel i in the noisy image, we calculate its similarity with other pixels j within a local neighborhood window of size $w \times w$. This approach is inspired by the classical NLM filtering method [1], which performs pixel-wise matching using raw pixel values. The proposed method, however, uses the learned feature vectors for matching, which significantly improves the accuracy of pixel similarity estimation and robustness to noise.

With $m(z)_i$ denoting the learned feature embedding of the center pixel under consideration i and $m(z)_j$ representing the embedding of a neighboring pixel j within the window, the pixel similarity metric between pixels i and j is computed based on the Euclidean distance between their feature embeddings given by

$$D_{i,j} = ||m(z)_i - m(z)_j||^2. \tag{2}$$

After calculating the Euclidean distances between the center and neighboring pixels using their feature vectors, a normalized inverse exponential scaling is applied to convert the calculated distances into normalized weights as

$$w_{i,j} = \frac{exp(-D_{i,j})}{\sum_j exp(-D_{i,j})}. \tag{3}$$

These weights are then used to compute the updated value of the center pixel using a weighted average of all the pixels in the $w \times w$ window around the center pixel. This process is repeated for multiple window sizes $w = [5,7,9,25]$ as shown in Fig. 1. The various steps of the MS-FMB are summarized in Algorithm 1.

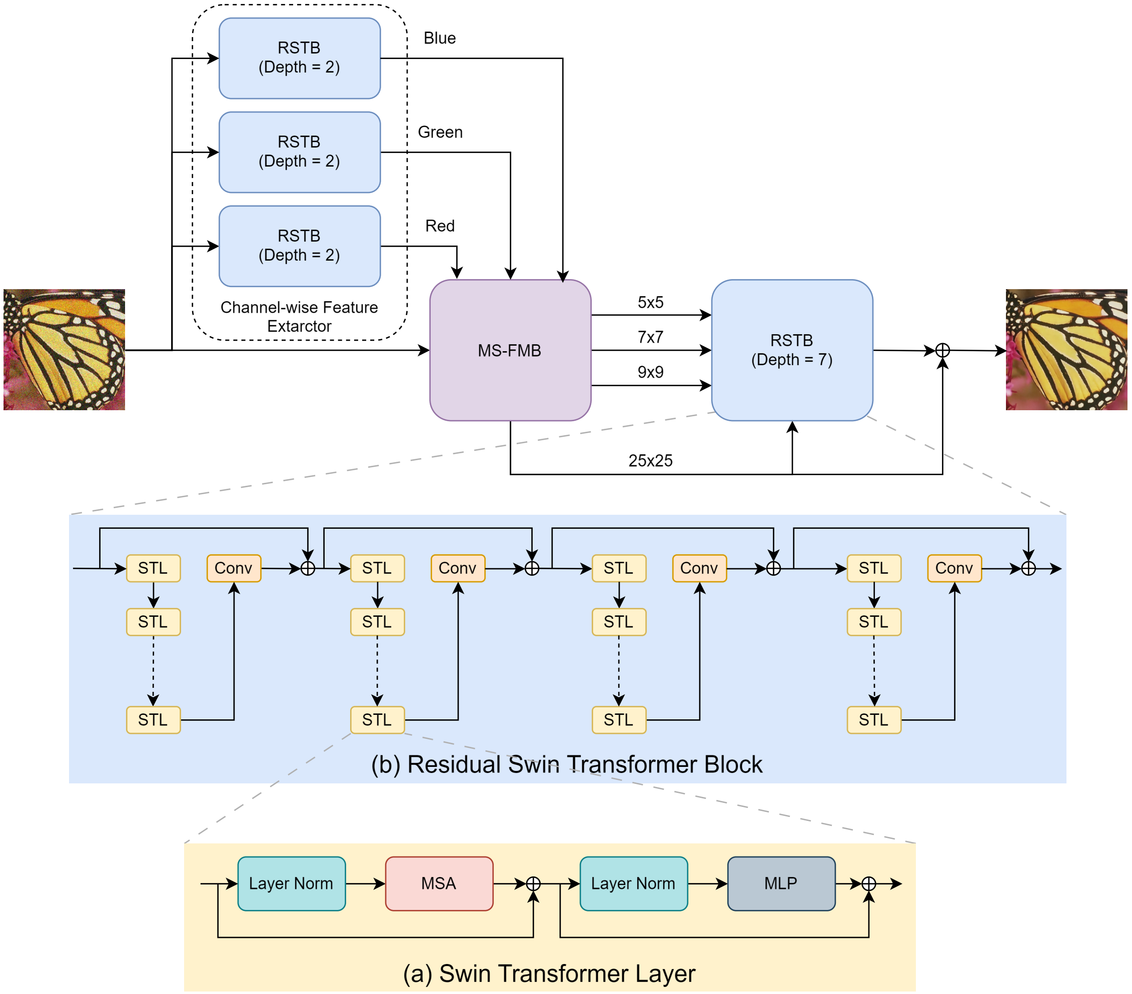

Figure 1: Proposed image denoising model architecture. The model consists of a channel-wise feature extractor using RSTBs, followed by a MS-FMB with multiple window sizes. A deeper RSTB is used to refine the output and recover fine details. Subfigures (a) and (b) show the STL and the RSTB respectively.

Algorithm 1 Multi-scale Feature Matching Block

```
Require: Noisy image $f(x,y)$, RSTB feature extractor $U$
Ensure: Estimated true image $u(x,y)$
1: Learn feature embeddings using RSTB feature extractor: $M(x,y,z) = U(f(x,y))$
2: for each pixel $i$ in the feature map do
3:     for each pixel $j$ in the neighborhood window of size $w \times w$ centered at $i$ do
4:         Compute the mean squared distance between feature vectors: $D_{i,j} = ||m(z)_i - m(z)_j||^2$
5:         Compute the weight for pixel $j$: $w_{i,j} = \frac{exp(-D_{i,j})}{\sum_j exp(-D_{i,j})}$
6:     end for
7:     Compute the weighted average of pixel values using weights $w_{i,j}$: $u(i) = \sum_j w_{i,j} f(j)$
8: end for
9: return Denoised image $u(x,y)$
```

Residual Detail Enhancement Block

In the final part of the model, a deeper RSTB is employed to further refine the output, capturing finer details of the image. This residual enhancement step leverages the hierarchical attention mechanism of Swin Transformers to enhance the feature representation, ensuring that essential image details are preserved while noise is effectively removed in the previous step (weighted averaging at multiple window sizes). This combination of residual connections and deep feature extraction aids in the overall robustness and quality of the denoised image by adding back the high frequency details in the denoised image.

Training Details

We use the Adam back-propagation algorithm to train the entire model. It is important to note that although the trainable weights are only in the RSTB part of the model, back-propagation is done for the entire architecture. This is done by treating the entire model as a computational graph to calculate derivatives for the gradient descent algorithm. This enables the model to learn noise-specific embeddings that help provide the most information for pixel matching, specifically in the later part of the model.

Dataset	σ	BM3D	DnCNN	FFDNet	SwinIR	Ours
Urban100	15	32.35	32.64	32.40	**33.70**	32.62
	25	29.70	29.95	29.90	**31.30**	29.96
	50	25.95	26.26	26.50	**27.98**	26.32
BSD68	15	31.07	31.73	31.63	**31.97**	31.63
	25	28.57	29.23	29.18	**29.50**	29.19
	50	25.60	26.23	26.29	**26.58**	26.14

Table 1: PSNR comparison for grayscale image denoising across various datasets and noise levels.

Dataset	σ	BM3D	DnCNN	FFDNet	SwinIR	Ours
Urban100	15	33.93	32.98	33.83	**35.13**	34.75
	25	31.36	30.81	31.40	**32.90**	31.51
	50	27.93	27.59	28.05	**29.82**	28.21
CBSD68	15	33.52	33.90	33.87	**34.42**	33.75
	25	30.71	31.24	31.21	**31.78**	31.07
	50	27.38	27.95	27.96	**28.56**	27.88
McMaster	15	34.06	33.45	34.66	**35.61**	34.67
	25	31.66	31.52	32.35	**33.20**	32.33
	50	28.51	28.62	29.18	**30.21**	29.18

Table 2: PSNR comparison for color image denoising across various datasets and noise levels.

The learning rate and betas used for training were 10^{-3} and $(0.9, 0.999)$ respectively. We used the DIV2K [27] and Flickr [28] dataset for training the model with mean-squared error (MSE) as the objective function loss metric.

Experimental Results and Discussion

In this section, we compare the performance of our proposed model against several state-of-the-art denoising algorithms: BM3D [18], DnCNN [3], FFDNet [15], and SwinIR [9]. We use PSNR as our evaluation metric. Additionally, we compare the computational efficiency of the models in terms of the parameter count. Our results demonstrate that the proposed model achieves a competitive balance between denoising performance and model size, which is particularly important when deploying in resource-constrained environments.

Quantitative Results

Table 1 and Table 2 show the PSNR values across different datasets (CBSD68 [24], Urban100 [26], McMaster [25]) and noise levels (15, 25 and 50) for grayscale and color image denoising respectively. It can be observed that SwinIR achieves the best performance (highlighted in black) amongst all the algorithms being compared. This is expected given the drastic difference (≈15x) in size between SwinIR and the rest of the algorithms being compared. This performance difference is increased at lower noise levels. Going a step further, the second and third best performing algorithms are highlighted in blue and red respectively in Table 1 and Table 2. It can be seen that the proposed solution achieves the best performance after SwinIR in majority of the cases. Furthermore, the performance difference is minimal for the cases where one of the other algorithms outperforms the proposed solution.

Qualitative Results

In addition to the quantitative results, we also provide qualitative comparisons in Figure 2. It can be observed that the proposed model preserves fine details and edges better than DnCNN and FFDNet, particularly in high-frequency regions. For the grayscale results, the proposed solution actually preserves details even better than SwinIR as highlighted by the zoomed in portion. For the colored results, SwinIR produces slightly better visual quality though the difference is subtle, making the proposed model an attractive solution for use in practical applications with limited resources.

Performance and Efficiency Trade-off

While SwinIR outperforms all other models in terms of PSNR and visual quality, it has significantly more parameters (11.5M). In contrast, our model provides competitive denoising performance at a fraction of the parameter count (743K). This balance between performance and computational efficiency makes our model more suitable for scenarios where memory and processing power are limited. DnCNN (558K) and FFDNet (490K) have slightly fewer parameters, however, as demonstrated by the quantitative and qualitative results, the proposed solution performs better in terms of PSNR and also better preserves fine details making it suitable for real-world applications that require both high performance and low computational cost.

Robustness and Flexibility

Our model's architectural innovations, such as multi-scale feature matching and residual Swin Transformer blocks, offer significant benefits. The multi-scale feature matching mechanism allows our model to capture both coarse and fine details, essential for handling real-world noise distributions that are often non-uniform. Additionally, the Swin Transformer blocks enable our model to preserve high-frequency details, such as edges and textures, even under heavy noise. The hierarchical architecture of the Swin Transformer blocks adds further flexibility of our model, making it adaptable to a wide range of image restoration tasks, such as super-resolution, inpainting, and deblurring. This flexibility is an advantage over CNN based models like DnCNN and FFDNet, which do not generalize as effectively to these tasks.

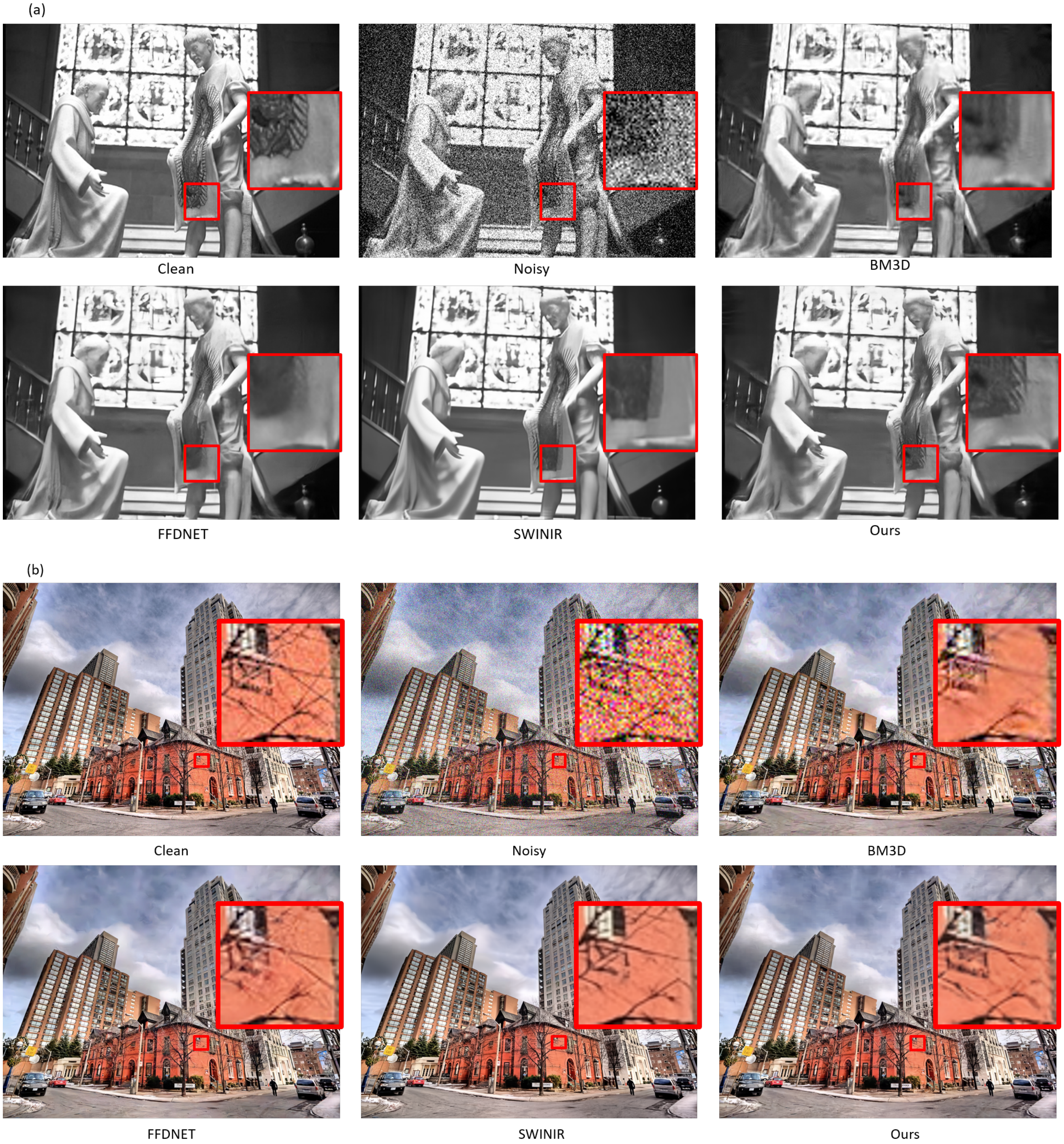

Figure 2: Qualitative comparison of denoised (a) grayscale and (b) color images across different algorithms at noise level σ=50. For the grayscale images, it can be observed that our model preserves fine details within the zoomed portion even better than SwinIR with 15x fewer parameters. For the colored images, SwinIR visually looks somewhat better than the proposed solution.

Future Work

Our ongoing work in this direction will leverage the flexibility of the multi-scale feature matching and transformer-based architecture of the proposed model to extend it to other image processing tasks, such as super-resolution, inpainting, and deblurring. Another important aspect of the denoising problem is to consider realistic noise modeling and noise adaptation techniques to improve the model's robustness to real-world noise. We also plan to explore optimization strategies, such as model pruning and quantization, to further reduce the computational footprint for real-time applications. Lastly, incorporating self-supervised learning methods will enable the model to generalize better with less reliance on labeled data, enhancing its practical applicability across diverse tasks.

Conclusion

In this paper, we proposed a novel image denoising model inspired by NLM that integrates multi-scale feature matching with residual Swin transformer blocks, offering a balance between performance and computational efficiency. Our model performs better than other SOTA denoising models with comparable number of parameters particularly in settings with a high degree of texture or repetitive structures. While the performance of the proposed solution is inferior to significantly larger networks like SwinIR and DeamNet in terms of PSNR, it seems to better preserve high frequency structure in the denoised images as demonstrated in our qualitative results. Additionally, our architecture shows potential for extension to other image restoration tasks such as super-resolution, inpainting, and deblurring. Future work will focus on optimizing the model for real-world noise, improving its adaptability, and its application to diverse image processing tasks.

Acknowledgements

We would like to thank *Timofey Uvarov* from *Fourier Imaging Lab* for his insightful comments and suggestions, which significantly improved the quality of this paper.

References

[1] A. Buades, B. Coll, and J.-M. Morel, ”A non-local algorithm for image denoising,” in Proc. of IEEE CVPR'05, 2005, pp. 60-65.

[2] A. Buades, B. Coll, and J.-M. Morel, ”A review of image denoising algorithms, with a new one,” Multiscale Modeling & Simulation, vol. 4, no. 2, pp. 490-530, 2005.

[3] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, ”Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising,” IEEE Trans. on Image Processing, vol. 26, pp. 3142-3155, 2017.

[4] C. Ledig, et al., ”Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. of IEEE CVPR, 2017, pp. 4681-4690.

[5] O. Ronneberger, P. Fischer, and T. Brox, ”U-Net: Convolutional networks for biomedical image segmentation,” International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2015, pp. 234-241.

[6] A. Dosovitskiy, et al., ”An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint:2010.11929, 2020.

[7] Z. Ren, Y. Zhang, P. Yi, L. Xu, W. Wang, Y. Luo, H. Sun, and M. Shan, ”DeamNet: Deep image denoising with adaptive activation function and attention mechanism,” IEEE Trans. on Image Processing, vol. 30, pp. 8221-8236, 2021.

[8] A. Vaswani, et al., ”Attention is all you need,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[9] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, ”SwinIR: Image restoration using Swin Transformer,” arXiv preprint arXiv:2108.10257, 2021, doi: 10.1109/iccvw54120.2021.00210.

[10] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep residual learning for image recognition,” in Proc. of IEEE CVPR, 2016, pp. 770-778.

[11] C. Tomasi and R. Manduchi, ”Bilateral filtering for gray and color images,” Sixth International Conference on Computer Vision, 1998.

[12] P. Perona and J. Malik, ”Scale-space and edge detection using anisotropic diffusion,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 12, no. 7, pp. 629-639, 1990.

[13] V. Jain and H. Seung, ”Natural image denoising with convolutional networks,” Advances in Neural Information Processing Systems, 2009.

[14] Y. Tai, J. Yang, X. Liu, and C. Xu, ”MemNet: A persistent memory network for image restoration,” in Proc. of IEEE ICCV, 2017, pp. 4539-4547.

[15] K. Zhang, W. Zuo, and L. Zhang, ”FFDNet: Toward a fast and flexible solution for CNN-based image denoising,” IEEE Trans. on Image Processing, vol. 27, no. 9, pp. 4608-4622, 2018.

[16] X. Mao, C. Shen, and Y. B. Yang, ”Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” Advances in Neural Information Processing Systems, vol. 29, pp. 2802-2810, 2016.

[17] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, ”Residual dense network for image super-resolution,” in Proc. of IEEE CVPR, 2018, pp. 2472-2481.

[18] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, ”Image denoising by sparse 3-D transform-domain collaborative filtering,” IEEE Trans. on Image Processing, vol. 16, no. 8, pp. 2080-2095, 2007.

[19] S. Gu, L. Zhang, W. Zuo, and X. Feng, ”Weighted nuclear norm minimization with application to image denoising,” in Proc. of IEEE CVPR, 2014, pp. 2862-2869.

[20] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, ”Variational denoising network: Toward blind noise modeling and removal,” Advances in Neural Information Processing Systems, 2019.

[21] S. Anwar and N. Barnes, ”Real image denoising with feature attention,” in Proc. of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3155-3164.

[22] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, ”Multi-level wavelet-CNN for image restoration,” in Proc. of IEEE CVPRW, 2018, pp. 773-782.

[23] X. Wang, R. Girshick, A. Gupta, and K. He, ”Non-local neural networks,” in Proc. of IEEE CVPR, 2018, pp. 7794-7803.

[24] D. Martin, C. Fowlkes, D. Tal, and J. Malik, ”A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in Proc. of IEEE ICCV, 2001, vol. 2, pp. 416-423.

[25] L. Zhang, R. Zhang, W. Shi, and H. Li, ”Color Image Denoising Using Patch-Based K-SVD and NLM,” in Proc. of ICIP, 2011.

[26] J.-B. Huang, A. Singh, and N. Ahuja, ”Single image super-resolution from transformed self-exemplars,” in Proc. of IEEE CVPR, pp. 5197–5206, 2015.

[27] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, ”NTIRE 2017 challenge on single image super-resolution: Methods and results,” in Proc. of IEEE CVPRW, pp. 1110–1121, 2017.

[28] Flickr, ”Flickr Image Dataset,” `https://www.flickr.com`, 2004.