Pureformer: Transformer-Based Image Denoising

Arnim Gautam[1], Aditi Pawar[1], Aishwarya Joshi[1], Satya Narayan Tazi[1], Sachin Chaudhary[2]
Praful Hambadre[3], Akshay Dudhane[4], Santosh Kumar Vipparthi[5], Subrahmanyam Murala[6],
[1] Government Engineering College Ajmer [2] UPES Dehradun [3]Indian Institute of Technology Mandi
[4]MBZUAI [5]Indian Institute of Technology Ropar [6]Trinity College Dublin

Abstract

Image denoising is a crucial task in computer vision with applications in real-world smartphones image processing, remote sensing, and photography. Traditional convolution neural networks (CNNs) often struggle to reduce noise while preserving fine details due to their limited receptive fields. Transformer-based approaches, such as Restormer, improve long-range feature modeling, while PromptIR enhances local feature refinement. However, existing methods still face challenges in effectively integrating multi-scale features for robust noise reduction. We propose Pureformer, a Transformer-based encoder-decoder architecture specifically designed for image de-noising. The model employs a four-level encoder-decoder structure, where each stage utilizes Multi-Dconv Head Transposed Attention (MDTA) and Gated-Dconv Feed-Forward Network (GDFN) to extract and refine multi-scale features. We proposed a feature enhancer block in the latent space expands the receptive field using a spatial filter bank, improving feature fusion and texture restoration. Skip connections between the encoder and decoder help retain spatial information, ensuring high-fidelity reconstruction. Pureformer is evaluated on the NTIRE 2025 Image Denoising Challenge dataset, achieving a test PSNR of 29.64 dB and SSIM of 0.8601. We also validated our Pureformer on existing benchmark datasets BSD68 and Urban100 datasets. The results demonstrate that Pureformer surpasses existing methods in both noise reduction and detail preservation, making it a strong choice for real-world image denoising. Access our codes and models from https://github.com/Chapstick53/NTIRE2025_cipher_vision.

1. Introduction

Image denoising is a fundamental problem in computer vision, essential for enhancing image quality in applications such as medical imaging [31, 59], remote sensing [14, 40], image super-resolution [3, 19, 52], depth estimation [26–28], precision agriculture [2] and autonomous systems [5, 48, 49]. Noise, particularly Gaussian noise, is often introduced during image acquisition due to sensor limitations [25, 43], environmental interference [29], or transmission artifacts [36, 54]. Traditional denoising methods, including wavelet-based filtering [6], non-local means [4], and block-matching approaches such as BM3D [14], rely on handcrafted priors and often struggle to generalize across diverse noise conditions. Developing robust denoising models capable of restoring fine image details while preserving structural information remains a critical challenge [22, 55, 59].

Deep learning-based denoising methods have demonstrated significant progress over traditional approaches, leveraging large-scale datasets and high-capacity neural networks. Convolutional Neural Networks (CNNs) such as DnCNN [59] and FFDNet [61] have been widely adopted due to their ability to learn complex noise distributions. However, CNN-based methods rely on localized receptive fields, limiting their ability to model long-range dependencies and multi-scale contextual information, often resulting in over-smoothed outputs.

Transformer-based architectures have recently emerged as powerful alternatives, addressing these limitations by leveraging self-attention mechanisms. Vision Transformers (ViTs) [16] introduced global feature modeling but suffer from high computational costs. SwinIR [36] mitigates this issue through hierarchical attention, improving efficiency while maintaining long-range feature capture. Restormer [57] significantly improved image restoration by introducing a channel-wise self-attention mechanism, which processes feature interactions along the channel dimension rather than spatially. This design enables efficient high-resolution processing while maintaining competitive performance. However, Restormer's fixed local window-based approach limits its ability to fully exploit spatial correlations, particularly for fine-grained textures and high-noise scenarios. Moreover, its hierarchical structure, while effective, lacks a dedicated feature refinement module in the latent space, which is crucial for handling severe noise levels. On the other hand, PromptIR [21] has introduced the im-

plicit prompting technique for all-in-one image restoration task, in which they apply implicit prompting at the decoder levels. While in encoder-decoder style architecture, encoder features merges with the decoder features via skip connections. keeping this in mind, due to the application of the implicit prompting only on the decoder side imbalances the harmony within encoder and decoder features. This affects the performance of the model in sever degradations.

To address these challenges, this work proposes Pureformer, a Transformer-based encoder-decoder model designed for image denoising at severe noise level ($\sigma = 50$). We design a four-level hierarchical encoder-decoder structure, where each level comprises Multi-Dconv Head Transposed Attention (MDTA) and Gated-Dconv Feed-Forward Network (GDFN) blocks to extract the robust features. We propose a feature enhancer block in the latent space which is composed of a multi-scale feature extraction through spatial filter bank followed by the series of transformer blocks to merge the multi-scale features and to improve the feature correlation. The propsed approach is depicted in Figure 1.

Compared to Restormer, the proposed model explicitly incorporates spatial feature aggregation, improving its ability to reconstruct fine textures in highly degraded images. Compared to PromptIR, Pureformer maintains a deeper hierarchical structure, ensuring robustness against high-noise scenarios. The main contribution of the paper are as follows:

- Introduction of Pureformer, a new Transformer-based encoder-decoder architecture designed to reduce noise from given severely noisy image.
- We propose a feature enhancer block in the latent space that extracts multi-scale features using a spatial filter bank. This is followed by a series of transformer blocks that enhance feature correlations and refine the latent space representations.
- Comparative evaluation on NTIRE 2025 Image Denoising Challenge dataset and existing benchmark datasets with state-of-the-art methods, demonstrating improved denoising performance in high-noise settings.

Extensive benchmarking on the NTIRE 2025 Image Denoising Challenge dataset [12], achieving PSNR: 29.65 dB, SSIM: 0.8601 on test data. The paper is structured as follows: Section 2 reviews related work, Section 2 details the proposed approach, Section 4 presents experimental results and comparisons, Section 5 presents results and discussion and Section 6 concludes with key findings and future directions.

2. Related Work

Image denoising is a critical task in computer vision, aiming to recover high-quality images from noisy inputs. Traditional approaches relied on handcrafted priors such as wavelet thresholding [6], total variation minimization [46],

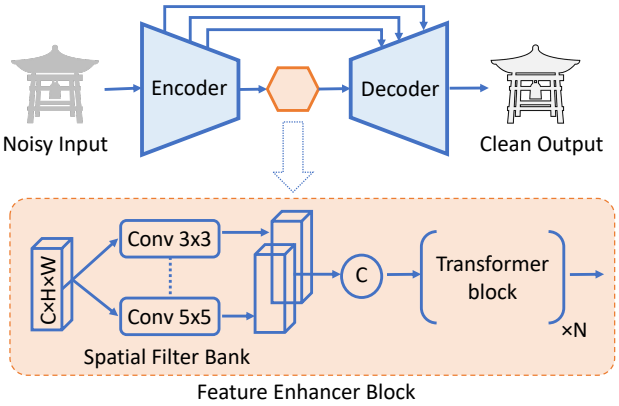

Figure 1. Proposed Pureformer encoder-decoder architecture for image denoising. The input noisy image is processed through a multi-level encoder, a feature enhancer block, and a multi-level decoder. Each encoder and decoder level employs xN transformer blocks [57], consisting of Multi-Dconv Head Transposed Attention (MDTA) and Gated-Dconv Feed-Forward Network (GDFN) blocks (shown in Figure 2). The feature enhancer block, placed in the latent space, expands the receptive field using a spatial filter bank. The multi-scale features are then concatenated and refined through xN transformer blocks to enhance feature correlation and merge multi-scale information effectively.

and non-local means filtering [4]. One of the most widely used classical methods is CBM3D (Collaborative Filtering for Color Images) [13], which leverages non-local filtering and block-matching techniques to remove noise while preserving texture. These methods effectively removed noise in structured regions but struggled with preserving fine textures and details. Patch-based methods like BM3D [14] and WNNM [23] improved performance by modeling non-local self-similarity in images. Bayesian formulations [40], sparse coding [54], and Markov random fields [45] further enhanced denoising by learning image priors. However, these methods lacked adaptability to diverse and complex noise patterns. The advent of deep learning introduced data-driven models that could learn intricate noise distributions and significantly outperform traditional techniques in various applications, including medical imaging [31], remote sensing [40], and autonomous systems [5].

Convolutional Neural Networks (CNNs) have been extensively employed for image denoising due to their ability to capture local structures within images. Early deep learning-based methods, such as TNRD [11] and RED-Net [41], demonstrated the effectiveness of residual learning and end-to-end training for denoising. DnCNN [59] introduced batch normalization and deeper architectures, significantly improving performance over traditional methods like BM3D [14]. FFDNet [61] further extended this by incorporating a tunable noise level map, allowing flexibility in handling varying noise intensities. Other CNN-based ap-

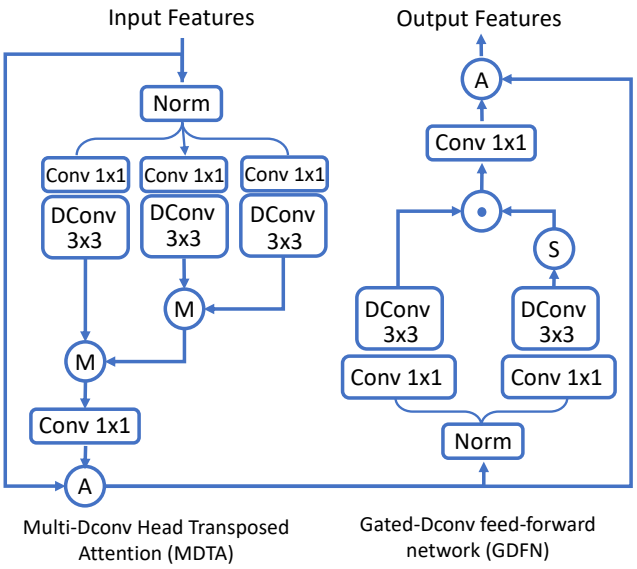

Figure 2. Transformer Block details: **Left side:** Multi-Dconv Head Transposed Attention (MDTA), which enables spatially enriched query-key feature interaction across channels instead of the spatial dimension. **Right side:** Gated-Dconv Feed-Forward Network (GDFN), which facilitates controlled feature transformation to ensure the propagation of useful information. Encircled M, A and S denotes matrix multiplication, addition and sigmoid operation respectively.

proaches, including IRCNN [60], MemNet [47], and NLRN [38], introduced iterative refinement and non-local feature aggregation to enhance denoising robustness. While these models effectively suppress noise, they remain constrained by their local receptive fields, limiting their ability to model long-range dependencies essential for capturing global context. Recent hybrid models like ADNet [51] and RIDNet [1] have attempted to mitigate this issue by integrating attention mechanisms, but their reliance on convolutions still restricts global feature modeling. To improve feature extraction and texture preservation, several CNN-based models have been proposed. BRDNet (Block Residual Denoising Network) [49] introduced dense residual connections within block-wise structures to facilitate multi-scale feature aggregation. More recently, AirNet (All-in-One Restoration Network) [32] incorporated self-attention modules within a CNN framework, bridging the gap between CNNs and Transformers. Unlike traditional CNN models, AirNet dynamically adjusts its receptive field based on the noise characteristics, allowing it to handle varying noise intensities more effectively. However, despite these improvements, CNN-based models still struggle with capturing long-range dependencies, limiting their ability to fully restore complex noise-corrupted images [7, 17, 18].

To overcome the limitations of CNNs, Transformer-based models have been explored for image restoration tasks [20]. The self-attention mechanism in Transformers enables global context modeling, which is beneficial for capturing long-range dependencies. Vision Transformers (ViTs) [16] introduced this concept to computer vision, demonstrating their ability to process spatial correlations more effectively than CNNs. IPT [8] extended this by leveraging large-scale pretraining on multiple image restoration tasks, achieving state-of-the-art (SOTA) performance in denoising and super-resolution. SwinIR [36] improved upon standard Transformers by incorporating a hierarchical structure and window-based self-attention, leading to better computational efficiency. Other works such as Deformable Transformer [62] and LeWin Transformer [33] further optimized Transformer architectures for image restoration by reducing computational overhead while maintaining strong global feature extraction. Additionally, Uformer [53] introduced a UNet-inspired hierarchical Transformer architecture tailored for image denoising, striking a balance between efficiency and performance. Despite these advancements, many Transformer-based models still struggle with high computational costs and inefficient processing of fine-grained details in high-noise images.

Restormer [57] addressed the computational challenges associated with high-resolution image restoration by introducing a transposed attention. Unlike conventional spatial self-attention, which operates across pixel locations, Restormer models feature interactions along the channel dimension, significantly reducing computational complexity while preserving global dependencies. This approach demonstrated superior efficiency in image denoising, deblurring, and deraining. However, despite its effectiveness, Restormer's fixed local window-based approach limits its ability to fully exploit spatial correlations, particularly in high-noise scenarios. Swin2SR [37] refined SwinIR by integrating residual Swin Transformer blocks, further enhancing denoising performance. Other Transformer-based methods, such as NAFNet [9] and HAT [10], introduced lightweight attention modules and hierarchical aggregation strategies to improve efficiency in real-time denoising applications. Meanwhile, AIDT [58] adopted an adaptive iterative framework to refine noise suppression progressively. While these advancements have improved image denoising, existing models still face challenges in retaining fine details and efficiently handling severe noise conditions.

Heterogeneous Window Transformer (HWformer) [50] addressed computational efficiency concerns by introducing heterogeneous global windows that shift both horizontally and vertically. This design effectively balances long- and short-range feature interactions, reducing computational overhead while preserving spatial coherence. EfficientFormer [34] further explored the hybridization of Transformers and CNNs, achieving a balance between com-

putational efficiency and denoising performance. Other hybrid models, such as RestNet [39] and DAT [15], employed dynamic attention mechanisms to optimize spatial feature extraction, resulting in improved image clarity with minimal artifacts. Meanwhile, MADFormer [24] introduced a memory-augmented Transformer architecture that progressively refines noisy inputs through multiple iterations, effectively suppressing noise while preserving fine textures. These approaches highlight the ongoing research efforts in optimizing Transformer-based architectures for image denoising, yet challenges remain in handling extreme noise levels while maintaining computational efficiency.

PromptIR [21] introduced a prompt-based learning paradigm for universal image restoration. Unlike conventional methods that require separate models for each type of image degradation, PromptIR leverages degradation-specific prompts to dynamically adapt its restoration network. This approach allows the model to generalize across multiple degradation scenarios without requiring explicit knowledge of the type or severity of corruption. Other works, such as AirFormer [35], explored the integration of adaptive prompts with self-attention mechanisms to further refine denoising performance. GLUformer [55] proposed a Global-Local Window Transformer block within a layered encoder-decoder structure, effectively capturing both local and long-range dependencies. Meanwhile, DenSformer [56] incorporated dense residual connections within Transformer layers, reinforcing local and global information integration to improve stability and denoising performance. Similarly, SUNet [22] utilized the Swin Transformer within a UNet-based model, efficiently merging hierarchical feature learning with long-range context modeling to enhance image restoration in high-noise settings.

Transformer-based image denoising models face challenges in balancing computational efficiency with feature retention. While methods like Restormer and PromptIR improve efficiency and adaptability, they struggle with preserving fine textures and high-frequency details in severe noise conditions. Building on existing advancements, we propose Pureformer, a Transformer-based encoder-decoder architecture that overcomes limitations in current models. By integrating a four-level hierarchical structure with MDTA and GDFN blocks, Pureformer enhances long-range dependencies and local feature interactions. Additionally, a feature enhancer block in the latent space expands the receptive field and improves feature fusion, leading to superior noise suppression and improved high-noise image restoration.

3. Proposed Approach

Overall Pipeline: The proposed Pureformer model follows an encoder-decoder architecture to effectively remove noise from images. The input noisy image I $\in$ C$\times$ H$\times$ W is first passed through a multi-level encoder, which extracts hierarchical feature representations $C_0 \in f \times H \times W$. Each encoder level consists of $\times$N transformer blocks, which include Multi-Dconv Head Transposed Attention (MDTA) and Gated-Dconv Feed-Forward Network (GDFN) modules to extract long-range dependencies. At latent space, the extracted features are then passed through our feature enhancer block. Our feature enhancer block expands the receptive field using a spatial filter bank, improving the model's ability to capture contextual information. These extracted multi-scale features are then concatenated along channel dimension and refined through $\times$N transformer blocks to enhance feature correlation and merge multi-scale information effectively. The processed features are then fed into a multi-level decoder, which reconstructs the denoised image using learned representations. Each decoder level, similar to the encoder, employs $\times$N transformer blocks to refine feature maps and preserve important structural details. Finally, the refined features are decoded into a clean, high-quality image, restoring fine textures while suppressing noise efficiently.

3.1. Efficient Transformer Block

To extract long range dependancisies, we utilise existing efficient transformer block [57]. It consists of two modules, multi-Dconv head transposed attention (MDTA) which performs (spatially enriched) query-key feature interaction across channels rather the spatial dimension, and Gated-Dconv feed-forward network (GDFN) that performs controlled feature transformation as shown in the Figure 2. Here, the GDFN regulates the flow of information across the hierarchical levels in the pipeline, ensuring that each level captures fine details that complement the others. In contrast to MDTA, which primarily enhances features with contextual information.

3.2. Feature Enhancer Block

Let $C_L \in f \times H \times W$ represents the input features to the latent space. Our feature exnhancer block processes the L_0 trough the spatial filter bank S_f where f represents the number of convolution layers. This filter bank has multiple convolution layers differing filter/kernel sizes. Further, it concatinates the extracted multi-scale features along channel dimension and process through the $\times N$ transformer block. The operation of our feature enhancer block is defined as below,

$$L_{S_f} = \tau[C_L^1, C_L^2, ..., C_L^f] \tag{1}$$

where, C_L^1 denotes the extracted features through the 1^{st} convolution layer of k$\times$k size filter, $[\cdot]$ represents the channel-wise concatination operation, τ represents the transformer block (as shown in the Figure 1).

4. Experimental Setup

To demonstrate the effectiveness of the proposed Pureformer, we evaluate it on NTIRE 2025 image denoising challenge on test set having noise level σ = 50. We also evaluate it on existing benchmark datasets BSD68 [42] and Urban100 [21] with noise level σ = 50.

4.1. Training Details

Our training strategy is carefully designed to balance efficiency and performance. We follow the competition guidelines and utilize the provided training set (DIV2K (1000) and LSDIR (86,991)) to train our model for the image denoising task. We consider patch-based training with a patch size of 128×128 followed by augmentations. The training process uses the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of $1e^{-4}$, which is gradually reduced to $1e^{-6}$ using a cosine annealing schedule that includes a linear warmup for 15 epochs. The batch size is set to 4, consisting of $4 \times 3 \times 128 \times 128$ patches, and training is conducted on $2\times$ A100 GPUs. Data augmentation techniques such as random cropping, flips, 90° rotations, and mixup are applied to improve generalization. We use L1 Loss to optimize the parameters.

5. Results and Discussion

In this section, we present a comprehensive evaluation of our Pureformer model against state-of-the-art (SOTA) denoising methods.

The performance is assessed on the NTIRE2025 Image Denoising Challenge dataset as well as widely recognized benchmarks such as BSD68[42] and Urban100 [21]. The effectiveness of Pureformer is measured using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), which are standard metrics for image quality assessment.

5.1. NTIRE 2025 Image Denoising Challenge

The NTIRE2025 Image Denosing Challenge has development and test phase. The test set consists of 200 images and having Gaussian noise with $\sigma = 50$). The final ranking is determined based on PSNR. Table Table 1 summarizes the challenge results for Test phase, where our Pureformer achieves a PSNR of 29.64 dB and SSIM of 0.8601, securing the 8^{th} position among all participating teams. We also show the visual results comparison between our Pureformer and existing PromptIR approaches which shows that our Pureformer restores the input noisy images by reducing the noise level at greater extent as shown in Figure 3.

In Figure 3, a visual comparison of image denoising performance across different methods is presented. The Figure 3 displays close-up views of the input noisy image, ground truth, results from an existing method, and the proposed Pureformer. Each denoised output is annotated with its corresponding PSNR value, reflecting the quality of restoration. Among the compared methods, Pureformer achieves the highest PSNR, indicating superior capability in recovering fine image details and structural integrity. Visually, the output from Pureformer is noticeably closer to the ground truth, exhibiting fewer artifacts and better-preserved textures, thereby validating its effectiveness in challenging high-noise scenarios.

Table 1. NTIRE2025 Image Denoising Challenge Results on a Test set ($\sigma = 50$).

Team	PSNR (dB)	SSIM	Ranking
SRC-B	31.20	0.8884	1
SNUCV	29.95	0.8676	2
BuptMM	29.89	0.8664	3
HMiDenoise	29.84	0.8653	4
Pixel Purifiers	29.83	0.8652	5
Alwaysu	29.80	0.8642	6
Tcler Denoising	29.78	0.8632	7
cipher_vision (Ours)	**29.64**	**0.8601**	**8**

5.2. Zero Shot Evaluation on Existing Datasets

Here, we evaluate our Pureformer on existing benchmark datasets BSD68 [42] and Urban100 [30] having Gaussian noise $\sigma = 50$. We carry zero-shot evaluation on these datasets and compare it with the existing conventional and transformer-based methods. The results are shown in Table 2 which depicts that our Pureformer clearly outperforms the existing conventional and most recent approaches on both BSD68 and Urban100 datasets.

Table 2. Quantitative image denoising results on BSD68 and Urban100 datasets with ($\sigma = 50$).

Methods	BSD68		Urban100	
	PSNR	SSIM	PSNR	SSIM
BM3D [13]	27.36	0.763	27.93	0.840
DnCNN [59]	27.92	0.789	27.59	0.833
IRCNN [60]	27.88	0.790	27.70	0.840
BRDNet [49]	28.16	0.794	28.56	0.858
AirNet [32]	28.23	0.806	28.88	0.871
Restormer [57]	28.41	0.810	29.31	0.878
PromptIR [21]	28.49	0.813	29.39	0.881
Pureformer	**28.68**	**0.819**	**29.78**	**0.890**

The superior performance of Pureformer is attributed to: *Hierarchical Feature Learning*: The multi-scale encoder-decoder structure enables better noise suppression and detail retention. *Feature Enhancer Block*: This block refines

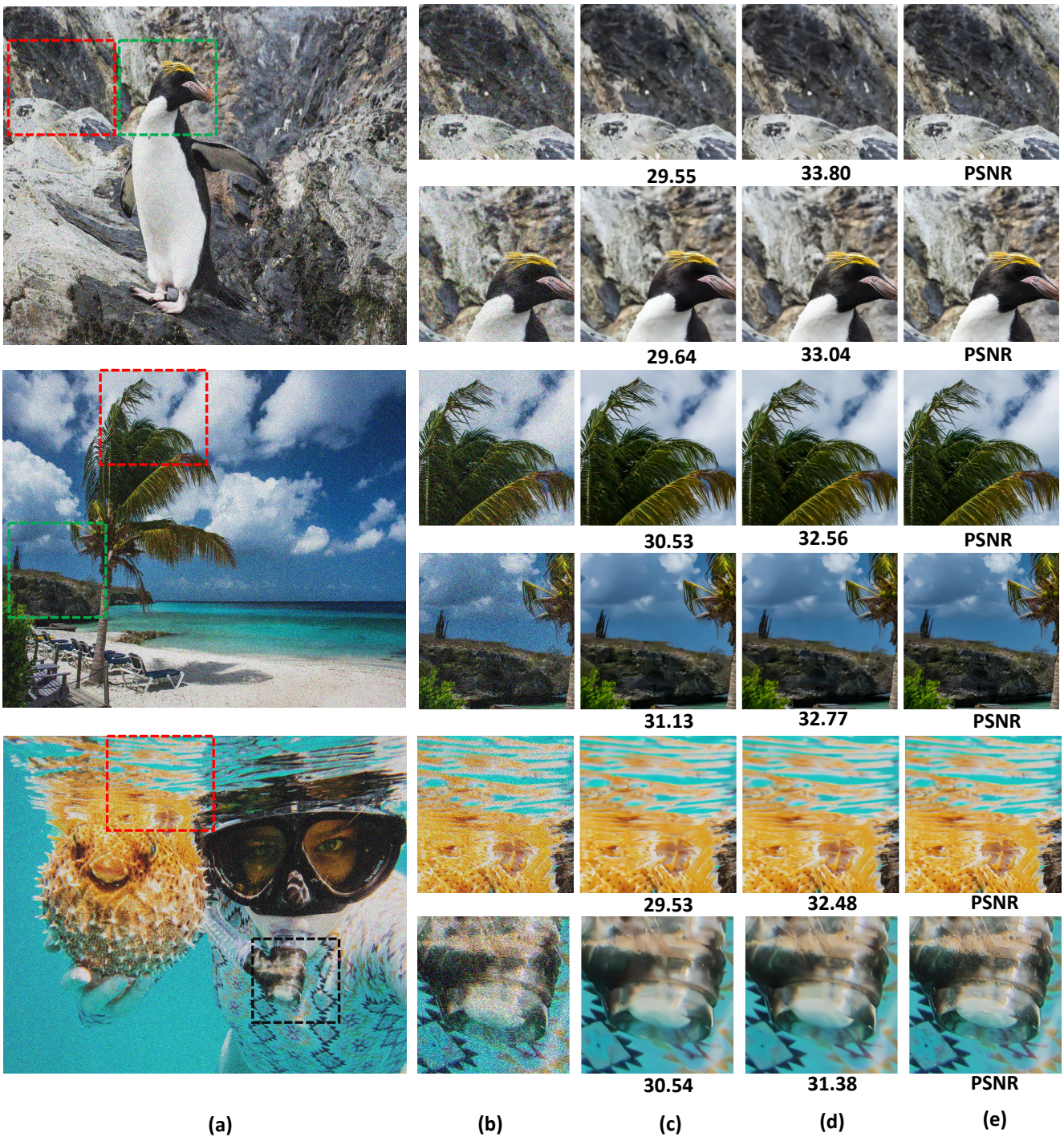

Figure 3. Visual results comparison of proposed Pureformer on NTIRE 2025 Image Denoising Challenge validation dataset. (a) Input noisy image, (b) Noisy Patch image, (c) Results of the PromptIR [21], (d) Results of the proposed Pureformer, and (e) Ground truth image.

representations by expanding the receptive field, improving local and global feature interactions. *Efficient Attention Mechanisms*: The MDTA and GDFN modules capture long-range dependencies while ensuring computational efficiency.

5.3. Computational Complexity Analysis

Here, We have compared the computational cost of our proposed Pureformer with Restormer [57] and PromptIR [21] in terms of Parameters and FLOPs. PromptIR is the most resource-intensive, with 35.59M parameters and 158.14 GFLOPs, making it ideal for high-performance setups. Restormer offers a balanced trade-off with 26.13M parameters and 140.99 GFLOPs. In contrast, Pureformer is the most lightweight and efficient model, requiring only 11.76M parameters and 64.31 GFLOPs, making it well-suited for real-time and edge-device applications.

In our proposed Pureformer architecture, we deliberately reduce the base channel dimension (dim) from 48 (as used in Restormer) to 32. This decision is driven by our goal to develop a lightweight and computationally efficient image restoration model without significantly compromising performance. The number of parameters in convolutional and attention layers scales approximately quadratically with respect to the channel dimension. Hence, using dim=32 reduces the overall computational load and memory footprint substantially, making the model more suitable for real-time or resource-constrained deployment.

Table 3. Comparison of model complexity and computational cost for image denoising

Methods	Parameters (M)	FLOPs (G)
Restormer [57]	26.13	140.99
PromptIR [21]	35.59	158.14
Pureformer	**11.76**	**64.31**

5.4. Discussion

The results demonstrate that Pureformer surpasses both classical CNN-based denoising methods and state-of-the-art Transformer-based models like Restormer [57] and PromptIR [44]. While inspired by Restormer's architectural design, Pureformer introduces a feature enhancer block that strengthens spatial-channel feature correlations. This block extracts multi-scale spatial features using a spatial filter bank, followed by Transformer layers that refine feature relationships through transposed attention on the channel dimension. By enhancing feature coherence, Pureformer significantly boosts image denoising performance, effectively addressing limitations in existing models.

6. Conclusion

In this paper, we introduced Pureformer, a Transformer-based encoder-decoder architecture for image denoising. Unlike conventional CNN-based methods, our model employs a four-level hierarchical structure with Transformer blocks, including Multi-Dconv Head Transposed Attention (MDTA) and Gated-Dconv Feed-Forward Network (GDFN), to effectively capture both local and global dependencies. The proposed feature enhancer block in the latent space further improves multi-scale feature fusion, enhancing denoising performance. We evaluated Pureformer on the NTIRE 2025 Image Denoising Challenge testset, demonstrating state-of-the-art performance with a PSNR of 29.64 dB and SSIM of 0.8601. We also evaluate and compare our Pureformer with other existing methods on existing benchmark datasets BSD68 and Urban100. Our method effectively suppresses noise while preserving fine image details, achieving competitive results compared to existing approaches. Future work will explore improvements in computational efficiency and generalization to real-world noise distributions. Additionally, integrating adaptive attention mechanisms and self-supervised learning techniques could further enhance the robustness of our model. Our work highlights the potential of Transformer-based architectures for image denosing, paving the way for further advancements in low-level vision tasks.

References

[1] Saeed Anwar and Nick Barnes. Real image denoising: A new benchmark dataset and a baseline model. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 12019–12028, 2019. 3

[2] Sandesh Bhagat, Manesh Kokare, Vineet Haswani, Praful Hambarde, Trupti Taori, PH Ghante, and DK Patil. Advancing real-time plant disease detection: A lightweight deep learning approach and novel dataset for pigeon pea crop. *Smart Agricultural Technology*, 7:100408, 2024. 1

[3] Goutam Bhat, Martin Danelljan, Radu Timofte, Yizhen Cao, Yuntian Cao, Meiya Chen, Xihao Chen, Shen Cheng, Akshay Dudhane, Haoqiang Fan, et al. Ntire 2022 burst super-resolution challenge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1041–1061, 2022. 1

[4] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 60–65, 2005. 1, 2

[5] Aayush Chakrabarty, Tobias Fischer, Paul Fieguth, and Krzysztof Czarnecki. Learning sensor multiplicity for robust visual perception in autonomous systems. *IEEE Robotics and Automation Letters*, 4(4):3402–3409, 2019. 1, 2

[6] S. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9):1532–1546, 2000. 1, 2

[7] Sachin Chaudhary and Subrahmanyam Murala. Tsnet: deep network for human action recognition in hazy videos. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3981–3986. IEEE, 2018. 3

[8] Hanting Chen, Yunhe Wang, Zhenhua Liu, Wei Zhang,

Chunjing Xu, and Chang Xu. Pre-trained image processing transformer. *CVPR*, 2021. 3

[9] Lequan Chen, Yihao Liu, Shanghua Gao, Yifan Sun, Qiong Chen, and Liang Lin. Simple baselines for image restoration. *ECCV*, 2022. 3

[10] Xi Chen, Zhiwei Xiong, Dong Liu, and Wenjun Zeng. Hat: Hybrid attention transformer for image restoration. *NeurIPS*, 2023. 3

[11] Yunjin Chen, Wei Yu, and Thomas Pock. On learning optimized reaction diffusion processes for effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10):2556–2570, 2017. 2

[12] NTIRE Challenge Committee. Ntire 2025 image denoising challenge dataset. *CVPR NTIRE Workshop*, 2025. 2

[13] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In *2007 IEEE international conference on image processing*, pages I–313. IEEE, 2007. 2, 5

[14] 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. 1, 2

[15] Qinglong Dai, Xinyu Chen, and Chenchen Wang. Dat: Dynamic attention transformer for image restoration. *NeurIPS*, 2023. 4

[16] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3

[17] Akshay Dudhane, Kuldeep M Biradar, Prashant W Patil, Praful Hambarde, and Subrahmanyam Murala. Varicolored image de-hazing. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4564–4573, 2020. 3

[18] Akshay Dudhane, Praful Hambarde, Prashant Patil, and Subrahmanyam Murala. Deep underwater image restoration and beyond. *IEEE Signal Processing Letters*, 27:675–679, 2020. 3

[19] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the ieee/cvf Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022. 1

[20] Akshay Dudhane, Omkar Thawakar, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Dynamic pre-training: Towards efficient and scalable all-in-one image restoration. *arXiv preprint arXiv:2404.02154*, 2024. 3

[21] Z. Zhang et al. Promptir: Prompting image restoration with query-based transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 4, 5, 6, 7

[22] Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. SUNet: Swin transformer UNet for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 1, 4

[23] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[24] Yifan Gu, Zhenyu Tang, and Xiaowen Huang. Madformer: Memory-augmented denoising transformer for image restoration. *CVPR*, 2023. 4

[25] Shuhang Guo, Zhen Chen, and Lei Zhang. Toward convolutional blind denoising of real photographs. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[26] Praful Hambarde, Akshay Dudhane, Prashant W Patil, Subrahmanyam Murala, and Abhinav Dhall. Depth estimation from single image and semantic prior. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1441–1445. IEEE, 2020. 1

[27] Praful Hambarde, Subrahmanyam Murala, and Abhinav Dhall. Uw-gan: Single-image depth estimation and image enhancement for underwater images. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

[28] Praful Hambarde, Gourav Wadhwa, Santosh Kumar Vipparthi, Subrahmanyam Murala, and Abhinav Dhall. Occlusion boundary prediction and transformer based depth-map refinement from single image. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 1

[29] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Richard T. Tan, and Marc Levoy. Burst photography for high-dynamic-range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, 35(6):192:1–192:12, 2016. 1

[30] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 5

[31] Wesam A. Jifara, Anazida Zain, Mohd A. Naufal, Yufeng Sheng, and Sheroz K. Sheroz. Medical image denoising using pre-trained deep learning models. *International Journal of Engineering and Technology*, 7(3):192–196, 2018. 1, 2

[32] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17452–17462, 2022. 3, 5

[33] Jiaxiang Li, Wenhai Wang, Enze Xie, Tong Lu, Hongsheng Li, and Yuhui Yuan. Efficient window attention transformer. *NeurIPS*, 2022. 3

[34] Yanyu Li, Shuyang Sun, Hong Zhang, Yuchao Lu, Rongrong Wang, Xian-Sheng Chen, and Jian Zheng. Efficientformer: Vision transformers at mobilenet speed. *arXiv preprint arXiv:2206.08890*, 2022. 3

[35] Zhenxing Li, Xiangyu Zhang, and Jian Sun. Airformer: Adaptive image restoration transformer with degradation-aware prompts. *NeurIPS*, 2023. 4

[36] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE Interna-*

tional Conference on Computer Vision (ICCV) Workshops, pages 1833–1844, 2021. 1, 3

[37] Jingyun Liang, Kai Zhang, Luc Van Gool, and Radu Timofte. Swin2sr: Swin transformer for compressed image super-resolution and restoration. *arXiv preprint arXiv:2303.10443*, 2023. 3

[38] Dongwei Liu, Zhaowen Wang, Baoyuan Wang, and Stephen Lin. Non-local recurrent network for image restoration. *NeurIPS*, 2018. 3

[39] Jianbo Liu, Zhaoyang Sun, and Yang Wang. Restnet: Residual network for image restoration with transformers. *ICLR*, 2023. 4

[40] Xiaoxiao Ma, Wei Zhao, Jiaqi Ma, and et al. A bayesian approach to image denoising based on remote sensing data. *Remote Sensing*, 11(19):2248, 2019. 1, 2

[41] X. Mao, C. Shen, and Y. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2802–2810, 2016. 2

[42] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 416–423. IEEE, 2001. 5

[43] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[44] Vaishnav A. Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. Promptir: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 7

[45] Stefan Roth and Michael J. Black. Fields of experts. *International Journal of Computer Vision (IJCV)*, 82(2):205–229, 2009. 2

[46] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992. 2

[47] Ying Tai, Jian Yang, and Xiaoming Liu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4539–4547, 2017. 3

[48] Rahul Tekchandani, Ritik Maheshwari, Praful Hambarde, Satya Narayan Tazi, Santosh Kumar Vipparthi, and Subrahmanyam Murala. Luminate: Linguistic understanding and multi-granularity interaction for video object segmentation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 4028–4034. IEEE, 2024. 1

[49] Chunwei Tian, Yong Xu, and Wangmeng Zuo. Image denoising using deep cnn with batch renormalization. *Neural Networks*, 121:461–473, 2020. 1, 3, 5

[50] Yingying Tian, Jian Zhang, and Yanning Zhang. Heterogeneous window transformer for image denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[51] Ce Wang, Xin Tian, Zhiwei Xiong, and Dong Liu. Attention-based densenet for image denoising. *IEEE Transactions on Image Processing*, 2020. 3

[52] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Ming Cheng, Haoyu Ma, Qiufang Ma, Xiaopeng Sun, et al. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1346–1372, 2023. 1

[53] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Jianzhuang Liu, and Shuhang Gu. Uformer: A unet-based efficient transformer for image restoration. *arXiv preprint arXiv:2201.02110*, 2022. 3

[54] Yuanming Xu, Jianrui Cai, Fan Zhu, and et al. Trilateral weighted sparse coding for multispectral image denoising. *IEEE Transactions on Image Processing*, 27(5):2473–2486, 2018. 1, 2

[55] Jian Xue, Yuchao Dai, and Hongdong Li. GLUformer: Global-local window transformer for efficient image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 4

[56] Jianwei Yao, Hao Tang, Song Bai, and Ling Shao. DenSformer: A dense transformer for image denoising. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 4

[57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1, 2, 3, 4, 5, 7

[58] Haoyu Zhang, Xintao Wang, Ying Shan, and Liang Lin. Adaptive iterative denoising transformer for image restoration. *CVPR*, 2023. 3

[59] 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

[60] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 3, 5

[61] 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

[62] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2021. 3