Super Resolution Image Denoiser Network: Toward Real-world Image Noise Removal

P. Mohit Harsh
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India
mohitharshxiib17@gmail.com

V. Teja Vardhan
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India
abhivardhan17@gmail.com

V. Vishnu Vardhan
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India
vivekvishnu768@gmail.com

J. Manoj Naidu
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India
manojnaidujogi@gmail.com

Dr. V. Padmanabha Reddy
Electronics and Communication Engineering
Institute of Aeronautical Engineering
Hyderabad, India
v.padmanabhareddy@iare.ac.in

***Abstract*—Image denoising techniques have been greatly improved in recent years by deep learning, with approaches like transformers and GANs pushing the limits of performance. These cutting-edge models are usually resource-intensive, need a lot of processing power, big datasets, and a long training period, which slow down real-time inference performance. In order to overcome these obstacles, we present the Super-Resolution Image Denoiser Network (SRIDNet), a small and incredibly powerful model that uses a fraction of the data and processing power to provide high quality results. SRIDNet maintains competitive performance in image denoising while drastically reducing the model size. Across a range of datasets, our model demonstrated the fastest inference times while maintaining high quality output when compared to other models, achieving avg. PSNR of 39.50 dB and 34.23 dB on SIDD and Urban100 datasets respectively.**

***Index Terms*—Convolution Neural Network (CNN), Adversarial Neural Network (GAN), SRIDNet, PSNR, Transformers.**

I. Introduction

In many computer vision and image processing applications, image denoising is an essential preprocessing step that restores distorted images by lowering noise while maintaining crucial structural details. Traditional denoising methods like wavelet transformations and Gaussian filtering are frequently constrained by their incapacity to adjust to the complex and varied noise patterns found in real-world photographs. The subject of image denoising has seen a revolution in recent years due to the development of deep learning, which offers strong data-driven solutions that can learn complex noise distributions and generate aesthetically pleasing outcomes.

The ability of Convolutional Neural Networks (CNNs) [1] to capture local spatial hierarchies through numerous layers of convolutional filters has made CNNs the dominant paradigm for picture denoising. By learning from massive datasets of noisy and clean image pairs, these models have achieved state-of-the-art outcomes, outperforming classical approaches by a wide margin. More complex designs like Transformer models [14] and Generative Adversarial Networks (GANs) [16] have also been proposed recently which further enhance denoising performance by capturing global dependencies and producing more realistic results.

These state-of-the-art models have significant computational costs, high memory needs, and a demand for enormous training datasets, despite their remarkable performance in controlled situations. Their applicability in real-world settings is limited by these constraints, especially in settings with limited resources and limited access to data and hardware.

To overcome these drawbacks, we propose the Super Resolution Image Denoiser Network (SRIDNet), a unique method that balances computational efficiency and denoising performance. Because SRIDNet is made to produce competitive outcomes with much fewer hardware requirements and a simpler model, it can be used in real-world scenarios. SRIDNet offers a reliable noise reduction solution without compromising performance or resource efficiency by combining super-resolution techniques with an effective denoising process. The architecture, training process, and performance evaluation of SRIDNet are presented in this study, demonstrating the network's benefits in terms of accuracy and efficiency over a range of datasets and noise levels.

II. Literature Survey

Deep convolutional neural networks (CNNs) have made substantial progress in the last ten years in high-level vision tasks such as object segmentation, motion analysis, and visual recognition. CNNs have been used more recently for low-level vision tasks like compression artifact reduction, image denoising, and super-resolution (SR). In these tasks, CNNs are trained to map low-quality images to high-quality outputs, usually with the goal of minimizing artifacts or removing noise. While ”deeper is better” is a widely accepted principle in high-level vision tasks—evidenced by networks like VGG, GoogleNet, and ResNet achieving substantial breakthroughs—this principle has not shown as much impact in low-level vision tasks. Despite the use of networks with 20 to 30 layers, such as DnCNN [1] and RED-Net, the performance gains in low-level tasks have been modest compared to earlier methods. This is because low-level vision tasks rely more on pixel-level features, where depth is less crucial. Instead, statistical priors, like non-local similarity or pixel distribution patterns (e.g., Gaussian noise), play a key role in enhancing the accuracy of these tasks, offering a more effective solution to image degradation issues.

A. Existing Work

In CBDNet [2], an asymmetric loss function was employed to enhance the model’s ability to generalize to real-world noise scenarios, while also facilitating convenient interactive denoising. The model demonstrated PSNR scores of 30.78 on the SIDD dataset [5] and 38 on the DND dataset. Processing a 512x512 image takes approximately 0.4 seconds.

In their work on RIDNet [3], the authors introduce a CNN-based denoising model specifically designed for both synthetic noise and real-world noisy images. The model’s architecture consists of four Enhanced Attention Mechanism (EAM) blocks, where most convolutional layers have a kernel size of 3x3, except for the final layer in the enhanced residual block and feature attention units, which utilize a 1x1 kernel. The model processes a 512x512 image in approximately 0.2 seconds during evaluation and achieved PSNR scores of 31.38, 39.23, and 38.71 on the BSD68 [4], DnD [6], and SIDD [5] datasets, respectively.

The NBNet [7] architecture is built upon a modified UNet framework. NBNet incorporates four encoder and decoder stages, where feature maps are downsampled using strided convolutions in the encoder and upsampled using deconvolutions in the decoder. It achieved PSNR scores of 29.16, 39.62 and 39.75 on BSD68 [4], DnD [6] and SIDD [5] datasets respectively.

DANet [2] is the latest architecture introduced for real-world image denoising tasks. Its core concept revolves around unfolding in-camera processing pipelines or learning the noise distribution through a generative adversarial network (GAN). This architecture achieved PSNR scores of 39.25 and 39.79 on the SIDD and DND benchmark datasets, respectively.

B. Evaluation Metrics

In image processing, Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), and Peak Signal-to-Noise Ratio (PSNR) are frequently used metrics to evaluate the quality of images, especially when determining how similar a processed image is to its reference image.

1) Mean Squared Error (MSE): MSE represented by equation (1) is a pixel-based metric that calculates the average of the squared differences between corresponding pixels of two images—typically, an original (reference) image and a processed (distorted or denoised) image.

$$MSE = \frac{1}{m \times n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i,j) - K(i,j))^2 \qquad (1)$$

Where:

- The value of the pixel at location (i, j) in the original image is represented by $I(i, j)$.
- The pixel value at location (i, j) in the processed image is denoted by $K(i, j)$.
- The image dimensions are m and n.

However, MSE doesn’t take human visual perception into account and is sensitive to large differences, even if small changes are hard to perceive.

2) Peak Signal-to-Noise Ratio (PSNR): PSNR represented by equation (2) is a more perceptually aligned metric that expresses the ratio between the maximum possible pixel value of the image and the error (noise) introduced by the processing, measured via MSE.

$$PSNR = 10 \times \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \qquad (2)$$

Where:

- MAX_I is the maximum possible pixel value (255 for an 8-bit image).
- MSE is the mean squared error between the two images.

Interpretation:

- Higher PSNR indicates that the processed image is of better quality and closer to the original.
- A PSNR above 30 dB is generally considered acceptable for most image processing tasks, though in applications like medical imaging, higher values are preferred.

PSNR is easy to compute and gives a high-level indication of image quality, but it still does not correlate well with human visual perception.

C. Drawbacks of existing methods

1) High Computational Complexity and Resource Demand: State-of-the-art denoising models typically consist of millions of trainable parameters, leading to significant computational demands. While these models perform exceptionally well in controlled laboratory environments, their need for vast datasets and computational resources makes them impractical for many real-world applications.

2) Inefficiency in Image Processing Speed: Current state-of-the-art models require 0.2 to 0.4 seconds to denoise a single 512x512 image. This processing time limits their usability in scenarios where real-time or high-speed image processing is essential.

III. Proposed Architecture

To mitigate the aforementioned limitations, we propose the Super-Resolution Image Denoiser Network (SRIDNet), a lightweight architecture designed for high speed denoising of images while maintaining the state of the art results. This Convolutional Neural Network (CNN) architecture is split into two primary blocks: the Super-Resolution Block and the Denoiser Block, with each serving a distinct purpose while contributing to the overall objective of producing high-quality, denoised images.

A. Input Layer

Input Shape: $(112, 112, 3)$
The model accepts an input image of size 112×112 with three color channels (RGB) as shown in Fig 1. This compact image resolution is chosen to reduce computational complexity and memory requirements while ensuring efficiency in model training and inference.

B. Super-Resolution Block

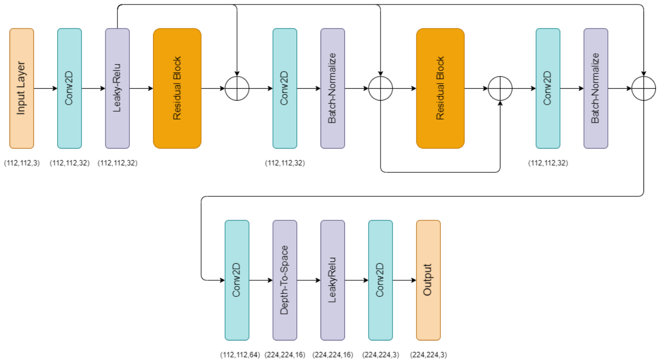

Fig. 1. Super-Resolution Block

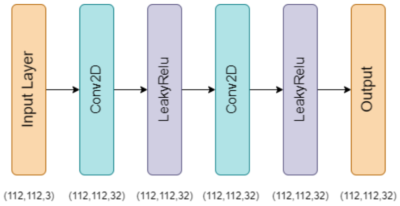

Fig. 2. Residual Block

Output Shape: $(224, 224, 3)$
Trainable Parameters: 110,515

The Super-Resolution Block is responsible for upscaling the input image from (112×112) to (224×224) by leveraging a series of *convolutional layers* and residual connections. This block effectively doubles the spatial resolution of the image and forms the encoder part of the architecture as shown in Fig 2.

- *Initial Convolution Layer:* A *Conv2D* layer with 32 filters and a kernel size of 3×3 is applied to the input, followed by a *LeakyReLU activation*. This step extracts low-level features from the image.
- Residual Blocks: Two *Residual Blocks* are employed. Each block consists of two convolutional layers with a kernel size of 3×3 and 32 filters, followed by *Batch Normalization* and *LeakyReLU* activations. These blocks capture intricate features while addressing the vanishing gradient problem. Skip connections are added to improve feature flow and gradient propagation, enhancing model training.
- *Final Convolution and Upsampling:* The final convolutional layer in the Super-Resolution Block has 64 filters, followed by a *Depth-to-Space operation* (also known as Pixel Shuffling), which rearranges the tensor to increase its spatial resolution to 224×224. This technique provides an efficient way to upscale the image.
- *Output:* A *Conv2D* layer with 3 filters reconstructs the RGB image, which now has double the spatial dimensions (224×224) compared to the input.

C. Denoiser Block

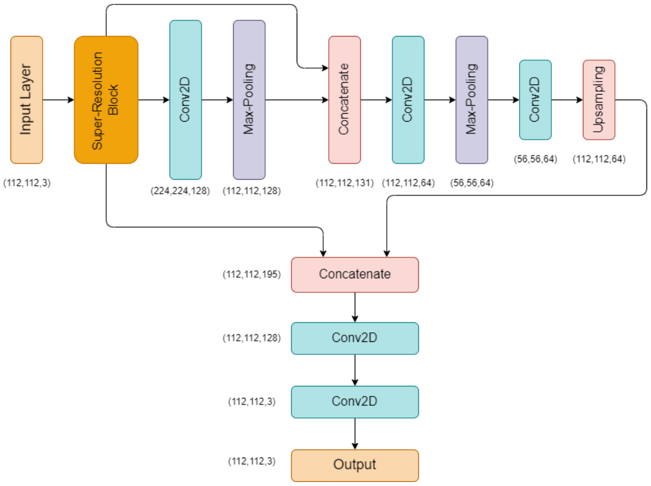

Fig. 3. Model Architecture

Output Shape: $(112, 112, 3)$
Trainable Parameters: 344,259

The Denoiser Block performs the task of noise removal by processing the upscaled image from the Super-Resolution Block. This component can be considered a *decoder block*, reducing the resolution back to (112×112) while preserving important visual features through skip connections and convolution operations as shown in Fig 3.

- *Convolution Layers:* The first convolution layer in the Denoiser Block has 128 filters, followed by a *MaxPool-*

ing2D layer that downscales the image to (112×112) for processing.

- *Concatenation with Input:* A skip connection is established where the original input image is concatenated with the downsampled version of the upscaled image, producing an intermediate feature map of size $(112 \times 112 \times 131)$. This fusion ensures the model retains essential features from the initial input.
- *Downsampling and Upsampling:*
  - The image is further downsampled using another convolutional layer with 64 filters, followed by another *MaxPooling2D* operation that reduces the spatial size to (56×56).
  - An *UpSampling2D* layer increases the resolution back to (112×112).
- *Concatenation with Intermediate Features:* Skip connections are introduced again by concatenating the upsampled feature map with previous intermediate layers, producing a feature map of size $(112 \times 112 \times 195)$.
- *Final Convolutions:* The concatenated feature map is passed through a convolutional layer with 128 filters and finally reduced to 3 channels (RGB) using a *Conv2D* layer with 3 filters. The output is the denoised image, reconstructed to its original resolution of (112×112).

D. Working Principle

- The architecture leverages the *Super-Resolution Block* to enhance the resolution of the input image from (112×112) to (224×224). This block employs a combination of *convolutional layers*, *residual connections*, and *depth-to-space transformation* to ensure that the upscaled image retains fine details without introducing significant artifacts.
- Following this, the *Denoiser Block* takes over, utilizing downsampling and upsampling techniques, combined with skip connections, to effectively remove noise from the upscaled image. The final result is a clean, denoised image at the original resolution of (112×112).
- *Skip connections* play a crucial role in both blocks by preserving critical features from the earlier stages and preventing information loss during downsampling and upsampling.

E. Summary of Parameters

- *Total Parameters:* 454,902
- *Trainable Parameters:* 454,774
- *Non-trainable Parameters:* 128

This architecture is highly optimized for the task of image denoising, leveraging the efficiency of the *Super-Resolution Block* to work with smaller input patches, significantly reducing the amount of data and time required for training and inference. Furthermore, the *Denoiser Block* ensures that the model can effectively remove noise while retaining critical image features, providing high-quality denoised images as output.

IV. Implementation

A. Data Preprocessing

1) SIDD [5]: In our study, we utilized images from the SIDD [5] dataset, resizing them to dimensions of (2576, 1456, 3). From each image, we extracted patches with a resolution of (112, 112, 3). For training, we randomly selected 7,000 patches, while 3,000 patches were allocated for validation. The final 10 images from the dataset were reserved for testing, which, after patching, yielded 2,990 test patches.

2) Urban100 [9]: In our study, we utilized the Urban100 [9] dataset to extract a total of 80 images for the training set. These images were cropped to a size of $560 \times 560 \times 3$ pixels. Subsequently, we generated patches of size $112 \times 112 \times 3$ from each image, resulting in a total of 2,000 patches designated for training with a validation split of 0.1. For the testing phase, we selected 20 remaining images from the dataset, which were similarly cropped to $560 \times 560 \times 3$ pixels. Following the patch generation process, we obtained 500 patches of size $112 \times 112 \times 3$ for testing purposes.

B. Model Training

The Super-Resolution Image Denoiser Network (SRIDNet) was trained using the TensorFlow and Keras frameworks on image patches of size (112, 112, 3). The training process was carried out in three stages, each with progressively reduced learning rates to enhance model performance and stability. The model was first trained using a learning rate of 0.001 across 20 epochs. After then, there were 10 further epochs with a further decreased learning rate of 0.00001, and 20 more epochs at a lowered learning rate of 0.0001. Peak Signal-to-Noise Ratio (PSNR) was utilized as the assessment metric, and the Adam optimizer was utilized to minimize the mean squared error (MSE) loss function. The batch size was set to 16 for all training stages. On the Urban100 dataset, the model required approximately 45 seconds per epoch, while on the SIDD dataset, it required 130 seconds per epoch. These timings reflect the computational complexity and dataset size differences between the two datasets.

V. Results and Comparisons

The proposed Super-Resolution Image Denoiser Network (SRIDNet) was evaluated on the Urban100 dataset and compared against state-of-the-art denoising models, including DnCNN [1], FFDNet [12], DRUNet [13], SwinIR [14] and Restomer [17] across different noise levels ($\sigma = 15, 25, 50$). The evaluation metric is Peak Signal-to-Noise Ratio (PSNR) as presented in table 1. The results indicate that SRIDNet achieved an average PSNR of 34.23 dB across the test images, with a maximum PSNR of 35.82 dB. Without the Super-Resolution block, SRIDNet acts just as a Deep CNN network giving the same results as DnCNN. Fig 4 and Fig 5 show the denoising results of SRIDnet on Urban100 dataset with added gaussian noise.

The proposed model was evaluated on the SIDD dataset and benchmarked against several state-of-the-art denoising models, including DANet [16], VDN [15], RIDNet [3], and CBDNet

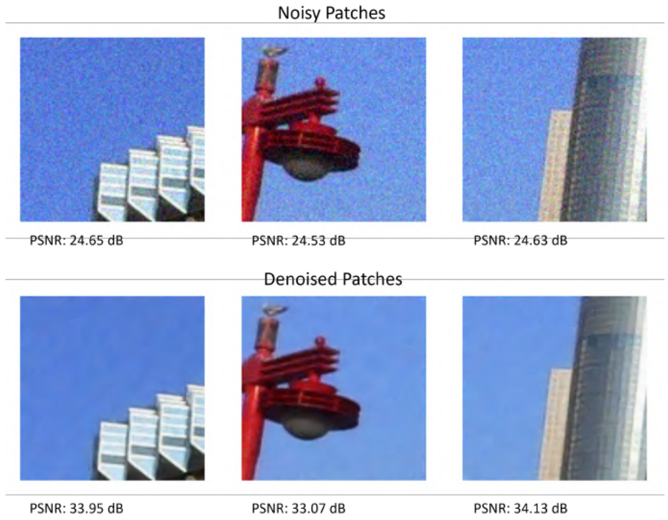

Fig. 4. PSNR results on $112 \times 112 \times 3$ image patches ($\sigma = 15$) from Urban100 dataset

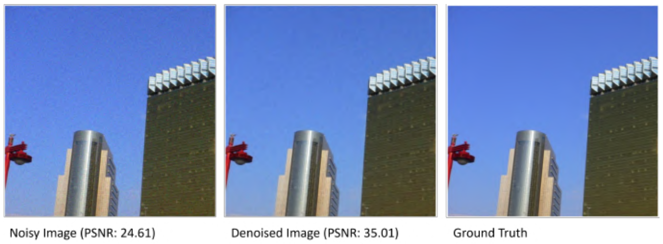

Fig. 5. PSNR results on an entire image ($\sigma = 15$) of size $560 \times 560 \times 3$ from Urban100 dataset

[2]. The model achieved an average PSNR of 39.5 on the SIDD dataset, with an average inference time of 95 ms for images sized 2576×1456, which were divided into 299 patches of size 112×112

Fig 6 and Fig 7 illustrate the denoising results of low-light and bright-light images from the SIDD [5] dataset, respectively, using $(112 \times 112 \times 3)$ image patches.

Table II presents a comparison of denoising performance across different models on the SIDD [5] dataset, showing that our model performed as good as the current state of the art models by achieving a average PSNR of 39.50 dB. While DnCNN [1] and FFDNet [12]—models of similar size—performed well on images with additive Gaussian white noise (AGWN) from Urban100 [9] dataset, they underperformed on real-world photographic images from the SIDD [5] dataset. In contrast, our model demonstrated robust performance across both datasets with greater efficiency. The Super-Resolution block plays a key role in improving the denoising results, as the model otherwise works just as a Deep CNN network and yields the same results as DnCNN.

Table III compares the inference times of the models evaluated on Nvidia RTX 3060 (12 GB) GPU. SRIDNet offers more consistent results on both the Urban100 and SIDD [5] datasets, while maintaining faster response times.

Model	σ	PSNR(dB)
DnCNN	15	32.98
	25	30.81
	50	27.59
FFDNet	15	33.83
	25	31.40
	50	28.05
DRUNet	15	34.81
	25	32.60
	50	29.61
SwinIR	15	35.13
	25	32.90
	50	29.82
Restomer	15	33.67
	25	31.39
	50	28.33
SRIDNet (no SR)	15	32.58
	25	30.78
	50	27.35
SRIDNet	15	**34.23**
	25	**32.01**
	50	**28.67**

TABLE I
URBAN100 IMAGE DENOISING RESULTS AND COMPARISON WITH EXISTING MODELS

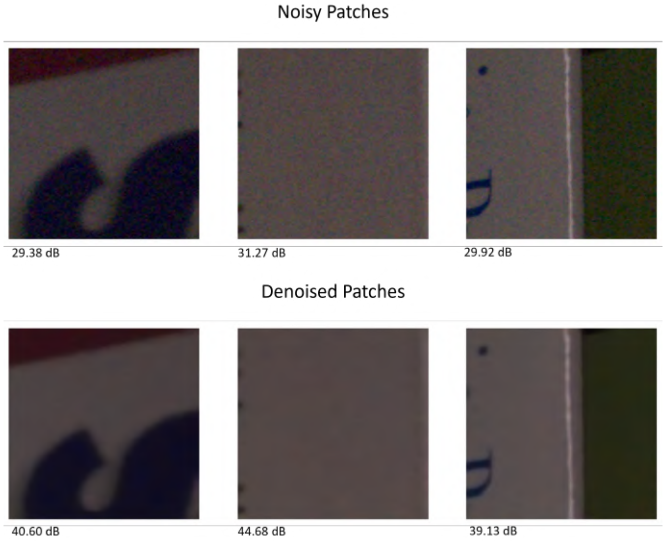

Fig. 6. Denoising results on SIDD dataset images

VI. Conclusion

In conclusion, this research presents SRIDNet, a novel deep learning architecture for image denoising that effectively addresses the limitations of current state-of-the-art models, such as high computational demands, extensive training data requirements, and slow inference times. By designing a lightweight model that integrates super-resolution and denoising tasks, SRIDNet achieves competitive results with significantly reduced resource consumption. Extensive testing on both synthetic (AGWN) and real-world noisy images from the Urban100 and SIDD datasets demonstrates the model's robustness and efficiency. With PSNR scores of 34.23 on

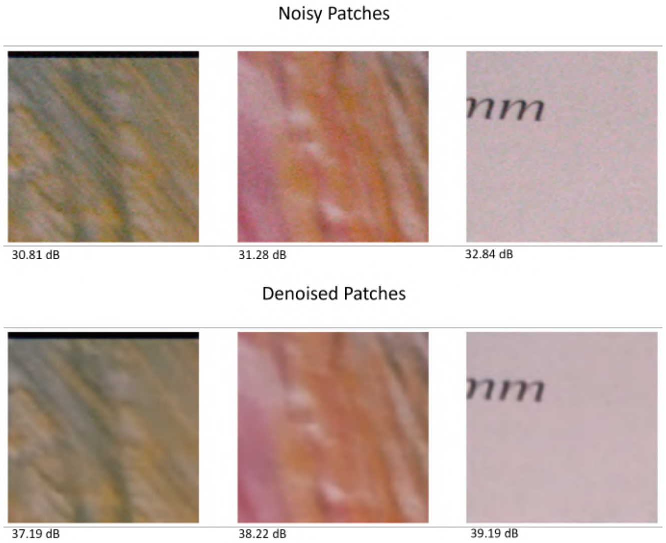

Fig. 7. Denoising results on SIDD dataset images

Model	PSNR (dB)
CBDNet	30.78
RIDNet	38.71
VDN	39.28
DANet	39.47
DnCNN	26.21
FFDNet	29.20
Restomer	40.02
SRIDNet (without Super-Resolution)	26.32
SRIDNet	**39.50**

TABLE II
COMPARISON OF RESULTS WITH STATE OF THE ART MODELS ON SIDD [5] DATASET

Urban100 and 39.50 on SIDD, SRIDNet delivers near state-of-the-art performance while maintaining the smallest model size and fastest inference time. This balance of efficiency and effectiveness positions SRIDNet as a promising solution for practical image denoising applications, particularly in environments with limited computational resources.

VII. ACKNOWLEDGMENT

We would like to extend our heartfelt gratitude to those who made this project possible through their constant guidance and support. Our deepest appreciation goes to our guide, Dr. V. Padmanabha Reddy, Assistant Professor, for his invaluable guidance and continuous cooperation throughout the project. We also want to thank the teaching and non-teaching staff for their support, and express our deepest gratitude to our parents, friends, and well-wishers for their assistance in completing this project report successfully.

Model	Inference Time (s)
DnCNN	0.0314
FFDNet	0.0071
DRUNet	0.0733
CBDNet	0.4
RIDNet	0.2
SRIDNet	**0.0208**

TABLE III
COMPARISON OF INFERENCE TIME WITH DIFFERENT MODELS.

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, 26:3142–3155, 2017.

[2] S. Guo, Z. Yan, K. Zhang, W. Zuo, L. Zhang: Toward Convolutional Blind Denoising of Real Photographs. In: CVPR (April 2019).

[3] S. Anwar, N. Barnes.: Real Image Denoising with Feature Attention. In: ICCV (March 2020).

[4] Stefan Roth and Michael J Black. Fields of experts. IJCV, 2009. 1, 2, 5, 6, 7

[5] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In CVPR, 2018. 5, 8

[6] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. arXiv preprint arXiv:1707.01313, 2017. 5, 7

[7] S. Cheng1, Y. Wang1, H. Huang, D. Liu, H. Fan and S. Liu.: NBNet: Noise Basis Learning for Image Denoising with Subspace Projection. In: CVPR (May 2021).

[8] Sara, U. , Akter, M. and Uddin, M. (2019) Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. Journal of Computer and Communications, 7, 8-18. doi: 10.4236/jcc.2019.73002.

[9] J.-B. Huang, A. Singh, and N. Ahuja, ”Single image super-resolution from transformed self-exemplars,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), June 2015.

[10] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, ”Learning photographic global tonal adjustment with a database of input/output image pairs,” in Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2011.

[11] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, ”Waterloo Exploration Database: New challenges for image quality assessment models,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 1004-1016, Feb. 2017.

[12] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. TIP, 2018. 1, 2, 5, 6, 7, 8

[13] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. 1, 2, 7, 8

[14] 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.

[15] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, ”Variational denoising network: Toward blind noise modeling and removal,” in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 1690-1701. Available: http://papers.nips.cc/paper/8446-variational-denoising-network-toward-blind-noise-modeling-and-removal.pdf.

[16] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, ”Dual adversarial network: Toward real-world noise removal and noise generation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Aug. 2020.

[17] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, ”Restormer: Efficient Transformer for High-Resolution Image Restoration,” Inception Institute of AI, Mohamed bin Zayed University of AI, Monash University, University of California, Merced, Yonsei University, Google Research, March 2022.