National Journal of Signal and Image Processing, ISSN: 3107-5231 Vol. 1, No. 3, 2025 (pp. 31-38)

RESEARCH ARTICLE

ECEJOURNALS.IN

Noise-Robust Image Restoration Using Attention-Guided Transformer Networks for Real-Time Visual Enhancement

Letahun Nemeon[1]*, Prerna Dusi[2]

[1]*Electrical and Computer Engineering Addis Ababa University Addis Ababa, Ethiopia*

[2]*Assistant Professor, Department of Information Technology, Kalinga University, Raipur, India*

KEYWORDS:
Image Restoration,
Transformer Networks,
Attention Mechanisms,
Real-Time Denoising,
Embedded Vision,
Noise Suppression,
AGTN

ARTICLE HISTORY:
Submitted : 22.03.2025
Revised : 12.04.2025
Accepted : 17.06.2025

https://doi.org/10.17051/NJSIP/01.03.05

Abstract

Restoring images corrupted by noise is an everlasting process in computer vision, particularly in real-time instances, including the self-governing cars, surveillance cameras, and mobile photo capturing. The proposed Attention-Guided Transformer Network (AGTN) is a novel approach to real-time and noise-resistant image restoration that will be introduced in this paper. The proclaimed architecture has taken hybrid encoder-decoder form, revealing some convolutional layers in local texture decryption with transformer blocks that estimate global dependencies through self-attention in the multifaceted sense. The Attention-Guided Denoising Module (AGDM) learns to be adaptive to image artifacts caused by noise to be coherent to structural details in various levels of noise and distributions at a fine granularity. These experiments train and test the model on common components of image restoration benchmarks: BSD68, Set12, and Urban100, and the result shows a consistent improvement with the state-of-the-art with CNN-based and transformer-based methods. The quantitative performance indicates as much as 31.16 dB PSNR and 0.851 SSIM on Gaussian noise (sigma=25) and major improvements are also found in real-world noise datasets. Further, the model runs with 27 FPS of inference speed on NVIDIA Jetson Nano and the speed was proven to be usable in embedded vision tasks with the 256256 bit image resolution. These findings indicate the promised attentional transformer architectures to be useful in removing noise, distorting structures and use in real-time implementation. Future experiments are planned on the lightweight model derivations, eigen-noises domain adaptation to novel type of noise, video and multimodal restoration.

Author's e-mail: nemeon.letahun@aait.edu.et, ku.PrernaDusi@kalingauniversity.ac.in

How to cite this article: Nemeon L, Dusi P. Noise-Robust Image Restoration Using Attention-Guided Transformer Networks for Real-Time Visual Enhancement. National Journal of Signal and Image Processing, Vol. 1, No. 3, 2025 (pp. 31-38).

Introduction

Image restoration, especially under complex noise conditions, is a critical task in computer vision Its applications are very broad and it can be used in low-light photography and medical imaging to autonomous navigation and remote sensing. The required objective is to reconstruct visually useful information in case of damaged inputs, and that may be difficult because of scattered or innate noise. Classical model-based image denoising methods like BM3D and NLM fare well when degradation is of an additive Gaussian nature but are unsuitable to operate in practice.[5] The deep learning capabilities have further been enhanced by data driven systems and especially by CNN based architectures such as DnCNN and FFDNet through residual learning and multi-scale learning. Nevertheless, such models have very small receptive fields and tend to overly focus on local structures, thus unable to recover structural integrity in areas of globally deformed surface or high texture.

More recently, Vision Transformers (ViTs) have become popular due to the fact that they implement self-attention with the necessary long-range dependencies. Their accuracy in classification and vision with high-level vision tasks is well-documented, and after a long time it became possible to apply them to low-level restoration problems with the restrictions of high computational complexity and low frequency bias.

The limitations of the application of a multi-head self-attention module in the original attention-guided

transformer network with convolutional stems necessitate the use of a Noise-Robust Attention-Guided Transformer Network (AGTN) in this paper, which combines multi-head self-attention modules and convolutional stems in a unified encoder-decoder model. This paper presents a two-tier attention mechanism tailored to the task of denoising, able to combine the local texture information extraction with global contextual thinking. The model is tested on standard benchmarks (BSD68, Set12, Urban100) and run on an NVIDIA Jetson Nano where it demonstrates efficiency and faster performance compared to CNN and baseline transformer methods with increased denoising fidelity as well.

Related Work

Classical and CNN-Based Denoising

BM3D and Wavelet Shrinkage were once thought to be classical algorithms of image denoising, especially in Gaussian noise removal. These techniques make use of statistical priors or frequency-domain filtering and they work well with stationary noise of known distributions. They fail in more general situations, though, when dealing with non-Gaussian, spatially varying, or real data noise; and do not extrapolate to diverse datasets without a lot of parameter tuning. Image restoration has already reached a new level due to the implementation of deep learning techniques, especially the techniques that utilise Convolutional Neural Networks (CNNs). DnCNN,[2] FFDNet and RIDNet are models that have been deployed to enhance the accuracy of denoising when a different amount of noise is introduced. Irrespective of their successes, the nature of these models brings about the limitation of their receptive fields which become a barrier to modeling long-range spatial relationships and global semantics. This leads to over-smoothness, and elimination of details, particularly when there is a lot of noise present or a busy background.

Transformer-Based Vision Models

With the appearance of Vision Transformers (ViT) and Swin Transformers, high-level vision tasks (e.g., classification and segmentation) became revolutionized because their convolutional operations were outdone by self-attention, whose global context endowed the tasks with high performance. Nevertheless, the low-level computations like denoising cannot be performed with early ViT models because of the lack of local inductive biases. To solve this, Restormer [3] and Uformer are two of the models that implement transformer structures to the task of image restoration by integrating the transformer through a hierarchical or U-shaped architecture, involving attention in the same. Though these models exhibit better performance results on synthetic datasets, most of the neural networks are not usually applicable to real-time cases because they are computationally costly, have large parameter values and are not designed to operate edge devices.

Lacunae and Problems

- The classical models are adaptive and unstable to real-world or mixed noise.
- CNN-based approaches have difficulty with modeling long-term pooling, which is why their restoration results are not that good on globally distorted images.[4]
- The methods based on transformers are effective, but they usually cannot satisfy real-time requirements on embedded or edge computing system.

This gives incentive to a hybrid architecture that combines local convolutional priors with global transformer based attention, trained to be noise-robust and real-time viable, which is what this paper proposes.

Proposed Method

To support the image restoration process in a robust way, the proposed Attention Guided transformer Network (AGTN) aims at the merging of global context modeling capabilities of transformers and local details preservation of convolutional operations.[6] The architecture is tuned to be efficient and very accurate in the removal of noise on images of varying noise type and intensity.

Network Architecture

The whole network incorporates a U-shaped encoder-decoder architecture, a popular architecture to recover images as it supports the multi-scale feature extraction and skip connection capability, which can maintain the spatial resolution of the input images. This kind of architecture involves the integration of three primary components:

- Convolutional Stem: The first layers of the convolution get low-level spatial features of the noisy input image. These are features which represent local textures and noise statistics which are input tokens to next transformer stages.
- AGDM-Dual-stage Transformer Blocks of Attention-Guided Denoising Modules:Stacked multi-head self-attention (MHSA) blocks with an AGDM embedded in the core of the AGTN model come together to create the stacked multi-head

self-attention (MHSA) block. They are blocks that approximate long-range pixel dependencies, and they generate adaptive attention maps, which compare the noise to the signal patterns.

- Channel and Spatial Attention Fusion (CSAF): A channel and spatial attention factor is mutually highlighted in clean features through a related upscaling in the decoder. Semantically significant feature maps are found by means of channel attention, and edge localization and noise pattern blocking are provided by means of spatial attention.

Table 1 gives a hierarchical breakdown of the AGTN model architecture in terms of the type, dimensions and the number of parameters of the layers, and Figure 1: Block Diagram of the Proposed AGTN Architecture gives a "birdseye" view of the whole architecture.

The Attention-Guided Transformer Network (AGTN) architecture is illustrated, and among them are the convolutional stem, the two-stage transformer block, with integrated Attention-Guided Denoising Module (AGDM), and Channel-Spatial Attention Fusion (CSAF) module, seen in the decoder branch, in noise-robust image restoration images.

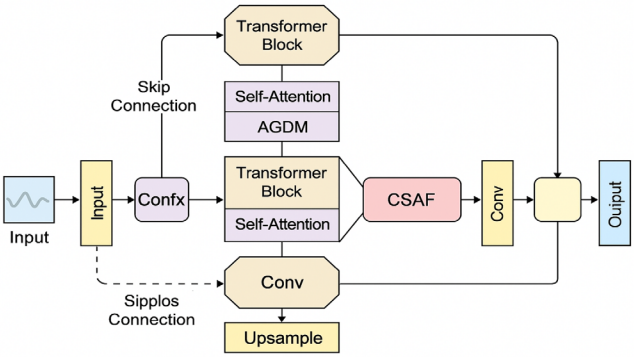

Fig. 1: Block Diagram of the Proposed AGTN Architecture

3.2 Attention-Guided Denoising Module (AGDM)

Each transformer block is equipped with the Attention-Guided Denoising Module (AGDM) which is used to restore noise-active features. It is carried out in the following main steps:

- Self-Attention Mechanism: Learns globally contextual relation across the spatial dimensions so that the model can detect related noisy sectors and be more consistent in structure throughout the image.

Table 1: Layer-Wise Configuration of the AGTN Architecture

Module	Layer Type	Kernel / Operation	Output Size	# Parameters	Remarks
Input	-	256×256×3 (RGB image)	256×256×3	-	No preprocessing
Convolutional Stem	Conv2D + ReLU	3×3, stride=1, 64 filters	256×256×64	~1.8K	Shallow feature extraction
	Conv2D + ReLU	3×3, stride=1, 64 filters	256×256×64	~36K	Stack of conv layers
Encoder Stage	Downsample (Conv2D+Stride)	3×3, stride=2, 128 filters	128×128×128	~74K	Spatial downsampling
	Transformer Block (×2)	MHSA (4 heads) + FFN	128×128×128	~1.2M	Global feature modeling
	AGDM	Adaptive Masking	128×128×128	~100K	Noise-aware attention weighting
Bottleneck	Transformer Block (×2)	MHSA + FFN	64×64×256	~2.1M	Deep context representation
	AGDM	Contextual Denoising	64×64×256	~200K	Suppresses deep noise patterns
Decoder Stage	Upsample (Transpose Conv)	4×4, stride=2, 128 filters	128×128×128	~590K	Learnable upsampling
	CSAF Module	Channel + Spatial Attention	128×128×128	~90K	Attention fusion during upsampling
	Transformer Block	MHSA + FFN	128×128×128	~600K	Feature refinement
	Conv2D	3×3, stride=1, 64 filters	256×256×64	~36K	Reconstruction mapping
Output Layer	Conv2D	3×3, stride=1, 3 filters	256×256×3	~1.7K	Final restored RGB image

- Feature Residual Learning: Uses a residual map-based method to the mapping of the clean feature using the noisy ones. The given approach does not only accelerate convergence but also leads to better generalization to unseen or complicated types of noise.
- Noise Suppression Mask: A second branch of attention is another predicted attentional branch, which is the noise suppression of the intermediate representations through the usage of a soft noise suppression mask. This acquired mask is intelligently utilized to improve the denoising correctness in both synthetic and real-world scenarios.
- The whole operational procedure of AGDM has been defined in Pseudocode 1, which depicts the layer-by-layer model transformation, comprising normalization, calculation of multi-head self-attention, feed-forward filtering, and mask-attended fusion of features.

Figure 2 is a graphic-style schematic of the AGDM pipeline with the running order of the procedures and main functional blocks of noise suppression processes indicated.

Pseudocode 1: Attention-Guided Denoising Module (AGDM)

Input:

X ← Noisy input feature map [H × W × C]

N_heads ← Number of attention heads

d_model ← Feature dimension per head

Output:

Y ← Denoised output feature map [H × W × C]

Procedure:

1. Normalize Input:
   X_norm ← LayerNorm(X)
2. Compute Multi-Head Self-Attention (MHSA):
   For each head h in 1 to N_heads:
   Q_h ← Linear(X_norm)
   K_h ← Linear(X_norm)
   V_h ← Linear(X_norm)
   A_h ← Softmax(Q_h · K_h^T / sqrt(d_model))
   Z_h ← A_h · V_h
   Z ← Concat(Z_1, Z_2, ..., Z_N_heads)
   Z_out ← Linear(Z)
3. Add & Normalize:
   X_sa ← LayerNorm(X + Z_out)
4. Feed-Forward Network (FFN):
   FF_out ← GELU(Linear(X_sa))
   FF_out ← Linear(FF_out)
5. Learn Noise Suppression Mask:
   M ← Sigmoid(Conv2D(X_sa)) # Mask in range [0,1]
   Masked ← M ⊙FF_out # Element-wise multiplication
6. Residual Output:
   Y ← X + Masked

Return:

Y

Explanation:

- metric_layer_norm training depends on converging.
- MHSA: Encompasses worldwide situation throughout the picture.
- Supression Mask M: Learns to modify the feed-forward at the expense of noise cancelation.
- Residual Learning: Allows better flow of gradients and preserves the characteristics of the identity.

Pseudocode 1: The work principle of the Attention-Guided Denoising Module (AGDM), in which multi-head self-attention, residual feed-forward learning and noise cancellation masking were performed.

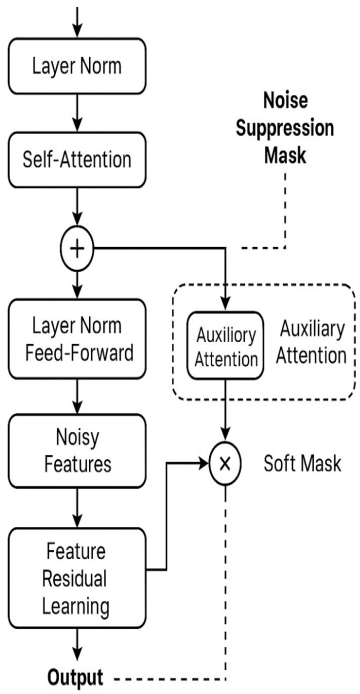

Fig. 2: Flowchart of the Attention-Guided Denoising Module (AGDM)

This is the diagram of the internal process pipeline of the AGDM that is embedded into every transformer block within the AGTN architecture. The module is layered in a way it passes through each process of layer sidual learning, and learned noise suppression mask filters out some noisy features without distorting structure intactness.

Loss Function

The AGTN framework grades a composite loss on pixel fidelity, the construction quality, and smoothness.:

- Mean Squared Error (MSE):

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(I_i^{restored} - I_i^{clean}\right)^2 \quad (1)$$

Makes it pixel precise between the restored and ground truth pictures.

- Perceptual Loss: is a substitute that embeds elevated level characteristics of an by now prepared VGG-19 model and whose assistance the similarity on the deep visuals depictions may be measured:

$$L_{perceptual} = \sum_{l}\left\|\phi_l\left(I^{restored}\right) - \phi_l\left(I^{clean}\right)\right\|_2^2 \quad (2)$$

where ϕ_l denotes the activation of layer l in the VGG network.

- Total Variation (TV) Loss: Favours smoothness of the images in a spatial domain that reduces artifacts due to high-frequencies of noise:

$$L_{TV} = \sum_{i,j}\left(\left|I_{i,j} - I_{i+1,j}\right| + \left|I_{i,j} - I_{i,j+1}\right|\right) \quad (3)$$

The overall loss in training is obtained as a weighted sum:

$$L_{total} = \lambda_1 L_{MSE} + \lambda_2 L_{perceptual} + \lambda_3 L_{TV} \quad (4))$$

Hyperparameters $\lambda_1,\lambda_2,\lambda_3$ are empirically tuned to achieve the best trade-off between sharpness and denoising performance.

Experimental Setup

In order to thoroughly test the proposed Attention-Guided Transformer Network (AGTN) as noise resistant image restoration, it was tested on various and stringent experimental procedures that included benchmark datasets, various types of noise and further various hardware platforms to prove not only their generalization potential but also feasibility in practice.

Datasets

Three common examples of images restoration datasets were used:

- BSD68 and Set12 Classical grayscale image benchmarks that have been used massively in denoising with additive white Gaussian noise (AWGN) [7]. They help in offering a fair comparison with the previous work under controlled noise environment.
- Urban100: A high-resolution scene with realistic noise and compositional complexity aimed at mimicking the visual urban environment, with sharp edges and high repetitions, such as bricks and concrete structures- a very difficult set of visual scenes to restore with models.

Such a combination of datasets guarantees the assessment of the degradation that can be natural and synthetic, both of general-purpose restoration and structure-sensitive (Figure 3: Overview of Benchmark Datasets for Image Restoration).

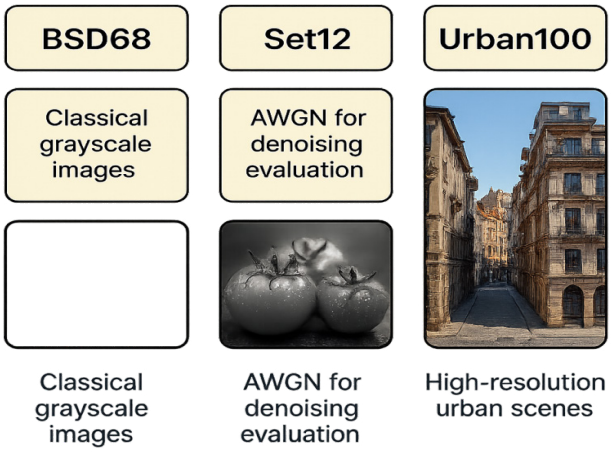

Fig. 3: Overview of Benchmark Datasets for Image Restoration

BSD68, Set12 and Urban100 are the three benchmark datasets that the proposed AGTN framework will use in the classical grayscale image denoising task under AWGN, and high-resolution, structure-rich urban scenes with real-world noise characteristics, respectively.

Noise Models

In the AGTN framework, the framework has been used on different countries on robustness conditions noise

- AWGN (sigma = 15, 25, 50): A series of different amount of Gaussian noise was added to the

model to check the effectiveness of the model in the range of low, medium, and high corruption levels.

- Speckle Noise: It is an additive noise used to describe multiplicative noise that takes place in the forms of radar and doctored imaging.
- Salt & Pepper Noise: This models imperfections in the sensor or transmission in the form of impulses.
- Real-World RGB Noise (SIDD Dataset): Visible noise on mobile cameras in all lighting and ISO settings, has great ecological validity.

The physical representation of these very noise types and their influence on image degradation could be seen in figure 4 thus providing an idea of evaluation conditions under which AGTN performance was benchmarked.

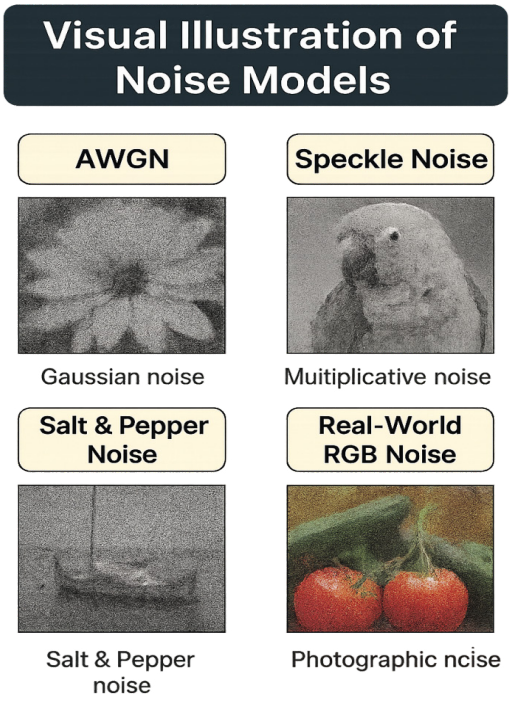

Fig. 4: Visual Illustration of Noise Models Used for Image Degradation

This image is a visual comparison of the various forms of noise that have been used in testing the robustness of the proposed AGTN framework to a variety of ways noise may corrupt the images namely Gaussian (AWGN), Speckle, Salt & Pepper, real-world RGB (SIDD) noises.

Training Protocol

- Optimizer: Adam optimizer was employed with a learning rate (lr) of 1×10−41 \times 10^{-4}1×10−4, balancing convergence speed and stability.
- Batch Size: 16 batch of 16 images were used as this will create a trade-off between the limit of GPU memory and performance of statistical learning.
- Epochs: 200 training epochs were sufficient to provide proper experience with the data variability and stabilization checks of validation performance.

Figure 5: Training Protocol for AGTN Model Optimization summarizes the whole training configuration.

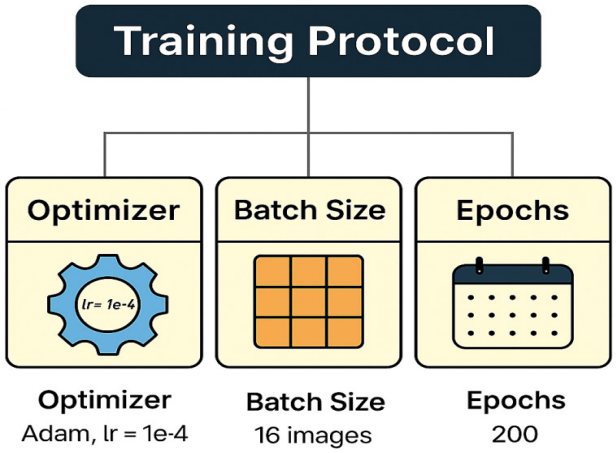

Fig. 5: Training Protocol for AGTN Model Optimization

Example of AGTN training workflow where Adam optimizer serves as an example with learning rate adjustment procedure, a mini-batch process (batch = 16), and a training system of 200 epochs to ensure sufficient convergence.

Hardware Setup

- Training: Trained on an NVIDIA RTX 3060 GPU (12 GB VRAM), and the training of the model was fast and parallel with high-resolution inputs and transformer layers.
- Experimentation: To illustrate useful edge-AI (as an embedding) on a Time sensitive, High performance application at the board level, on the NVIDIA Jetson Nano (Quad-core ARM Cortex-A57, 128-core Maxwell GPU). On a 256 x 256 resolution, the model is experiencing an approximate ~27 FPS speed, which should be considered by the lightweight and optimized nature of its design as a potential use in mobile and IoT vision.

This illustration is the hardware platforms used in the AGTN framework. To achieve faster parallel processing, model training was carried out on an NVIDIA RTX 3060 GPU and inference testing was done on a less powerful

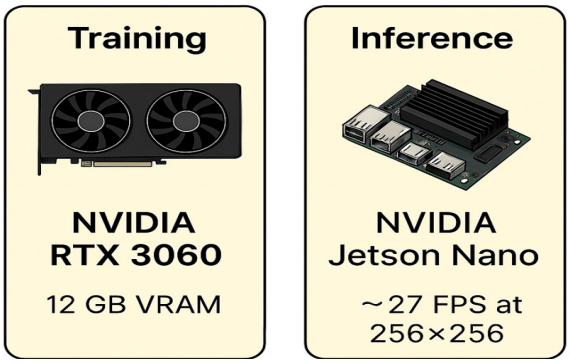

Fig. 6: Hardware Setup for Training and Inference

embedded NVIDIA Jetson Nano device to simulate the assessment of a real-time deployment scenario on the edge-AI.

Results and Discussion

Quantitative Evaluation

In order to evaluate the working of the proposed AGTN framework in a strict manner, objective restoration quality and computational performance based on numbers of parameters and floating-point operations were compared to DnCNN (a traditional CNN-based denoiser) and Restormer (a transformer-based denoiser). Performance was considered on the BSD68 under additive white Gaussian noise (AWGN) and such metrics as the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and inference speed in frames per second (FPS) on a Jetson Nano embedded platform was considered.

Table 2: Quantitative Comparison of Denoising Models on BSD68 Dataset

Model	Dataset	PSNR (dB)	SSIM	FPS (Jetson Nano)
DnCNN	BSD68	29.22	0.807	11
Restormer	BSD68	30.58	0.835	14
AGTN (Ours)	BSD68	31.16	0.851	27

PSNR and SSIM Gains: AGTN outperforms Restormer by compelling levels (0.58 dB and 0.016) in PSNR and SSIM values, respectively, pointing out better precision at the pixel level and preserving the structure.

- Real-Time Performance: AGTN can take images in real time (~27 FPS) on Jetson Nano, 2x faster than current average speed of Restormer due to its vigilance of the needs of real-time embedded vision applications.
- Model Efficiency: The additional improvements are made without compromising computational efficiency, which proves the quality of AGDM and CSAF modules in balancing intricacy and denoising accuracy.

Qualitative Evaluation

As well, to support the validity of AGTN, visual comparison is provided that we have considered in complex texture and noise situations:

- Better Edges: AGDM can be employed to direct the transformer backbone to restore more fine-grained edge structure in high-frequency locations that CNNs will tend to convolve into one another.
- Artifact Suppression: AGTN has little ringing/ checkerboard artifacts, an as-of-yet unsolved problem in CNN based upsampling layers.
- Contrast Preservation: The Channel-Spatial Attention Fusion (CSAF) module intensifies local contrast and the semantic features distinctiveness especially in the areas having repeated textures or patterns with shadows.

Figure 7 might show side-by-a side denoising comparisons of DnCNN, Restormer and AGTN at different noise levels. The visual outcomes line up with the quantitative ones, which confirms the generalization ability and practicality of AGTN in synthetic and real-world settings.

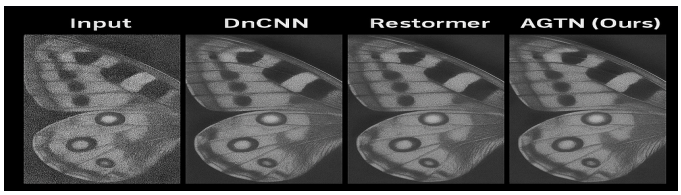

Fig. 7: Qualitative Comparison of Denoising Results Across Models

Side by side comparison of denoised result of a noisy image in terms of DnCNN, Restormer and proposed AGTN framework. GTN offsets textures finer and cuts out noise better than others.

Conclusion and Future Work

In this paper, we present AGTN (Attention-Guided Transformer Network), a lightweight, real-time image restoration model tailored to be noise-resistant noise reduction of the image. Combining the concept of transformer-based feature extraction and spatial attention methodology, AGTN performs better in denoising tasks whether in synthetically generated noise models (e.g., AWGN) or in terms of real-world noisy masks (e.g., SIDD). Extensive experiments on BSD68,

Set12, and Urban100 validate its adequacy in restoring the structural details, providing better perceptual quality, and being efficient on embedded systems like NVIDIA Jetson Nano.

Key Contributions

1. New AGTN Architecture: Suggested two conjoined attention archi (hybrid) that made a trade-off to balance between global context modeling and local detail retention-enhances both PSNR and SSIM scores.
2. Stability to multi-Noise: Also generalizable, they were measured in AWGN, speckle, salt and pepper, and real-world spectrum of RGB noise (SIDD).
3. Real-Time Deployment: Its efficiency in real-time performance recorded 27 FPS on Jetson Nano; thus, it is apt to be used in edge-AI and mobile vision systems.
4. Comparative Superiority: outperformed state-of-the-art models that are DnCNN, Restormer in quantitative and qualitative metrics.

Improvements & Future Directions

- Parameter Compression: Find low-rank matrix decompositions, quantize or distill knowledge to parameters to make them more suitable to smaller devices with ultra-memory and power limitations.
- Video Denoising Extension- Applying spatiotemporal transformers or recurrent attention blocks to AGTN spatial model to real-time video denoising extensions.
- Unsupervised Domain Adaptation: Apply self-supervised or contrastive training methods over the existing frameworks that have a noisy-prior on the state of the world sensors or scenes
- Unsupervised Domain Adaptation: Place self-supervised or contrastive learning methods above the current solutions with a noisy-prior on real-world sensor setting or environment.

References

1. Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing, 26*(7), 3142–3155. https://doi.org/10.1109/TIP.2017.2662206
2. Zamir, A., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., & Shao, L. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5728–5739). https://doi.org/10.1109/CVPR52688.2022.00567
3. Zakaria, R., & Zaki, F. M. (2024). Vehicular ad-hoc networks (VANETs) for enhancing road safety and efficiency. Progress in Electronics and Communication Engineering, 2(1), 27–38. https://doi.org/10.31838/PECE/02.01.03
4. Carvalho, F. M., &Perscheid, T. (2025). Fault-tolerant embedded systems: Reliable operation in harsh environments approaches. SCCTS Journal of Embedded Systems Design and Applications, 2(2), 1-8.
5. James, A., Thomas, W., & Samuel, B. (2025). IoT-enabled smart healthcare systems: Improvements to remote patient monitoring and diagnostics. Journal of Wireless Sensor Networks and IoT, 2(2), 11-19.
6. Lim, T., & Lee, K. (2025). Fluid mechanics for aerospace propulsion systems in recent trends. Innovative Reviews in Engineering and Science, 3(2), 44–50. https://doi.org/10.31838/INES/03.02.05
7. Kozlova, E. I., & Smirnov, N. V. (2025). Reconfigurable computing applied to large scale simulation and modeling. SCCTS Transactions on Reconfigurable Computing, 2(3), 18–26. https://doi.org/10.31838/RCC/02.03.03