

AKDT: Adaptive Kernel Dilation Transformer for Effective Image Denoising

Alexandru Brateanu¹ ^a, Raul Balmez¹ ^b, Adrian Avram² and Ciprian Orhei² ^c

¹University of Manchester, Manchester, U.K.

²Politehnica University of Timișoara, Timișoara, Romania

{alexandru.brateanu, raul.balmez}@student.manchester.ac.uk, {adrian.avram, ciprian.orhei}@upt.ro

Keywords: Image Denoising, Transformer Models, Dilated Convolutions, Deep Learning, Computer Vision.

Abstract: Image denoising is a fundamental yet challenging task, especially when dealing with high-resolution images and complex noise patterns. Most existing methods rely on standard Transformer architectures, which often suffer from high computational complexity and limited adaptability to varying noise levels. In this paper, we introduce the Adaptive Kernel Dilation Transformer (AKDT), a novel Transformer-based model that fully harnesses the power of learnable dilation rates within convolutions. AKDT consists of several layers and custom-designed blocks, including our novel Learnable Dilation Rate (LDR) module, which is utilized to construct a Noise Estimator module (NE). At the core of AKDT, the NE is seamlessly integrated within standard Transformer components to form the Noise-Guided Feed-Forward Network (NG-FFN) and Noise-Guided Multi-Headed Self-Attention (NG-MSA). These noise-modulated Transformer components enable the model to achieve unparalleled denoising performance while significantly reducing computational costs. Extensive experiments across multiple image denoising benchmarks demonstrate that AKDT sets a new state-of-the-art, effectively handling both real and synthetic noise. The source code and pre-trained models are publicly available at <https://github.com/albrateanu/AKDT>.

1 INTRODUCTION

Image enhancement encompasses various techniques aimed at improving the quality, clarity, and perceptibility of images. The main goal is to create visually appealing images, correct an image, or manipulate certain features for specific use cases. Specific categories of enhancements can be highlighted as such: image sharpening (Orhei and VasIU, 2023; Bogdan et al., 2024), low-light enhancement (Wang et al., 2020; Brateanu et al., 2024), de-hazing (Ancuti et al., 2017), denoising (Liang et al., 2021).

Image denoising is a crucial field within image restoration which aims to enhance image quality by eliminating noise introduced by digital and natural factors alike. The noise, manifesting as random variations of brightness and color information, often complicates the task of maintaining image characteristics such as sharpness and texture. As such, the complexity of this problem has led to the adoption of deep

neural networks, which have shown success in various image denoising tasks, as evidenced by recent benchmarks

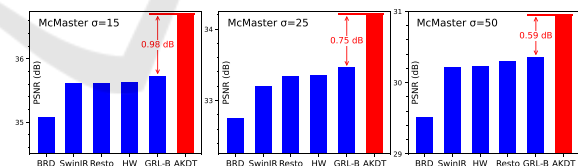





Figure 1: AKDT outperforms SOTA, including BRDNet (Tian et al., 2020), SwinIR (Liang et al., 2021), and Restormer (Zamir et al., 2022), GRL-B (Li et al., 2023), and HWFormer (Tian et al., 2024) on denoising benchmark McMaster at various noise levels.

The field of image denoising has seen a paradigm shift with the introduction of Transformer models (Dosovitskiy et al., 2021), which excel in capturing extensive contextual information, having larger receptive fields when compared to earlier Convolutional Neural Networks (CNN) based methods. However, the high computational cost of Transformer architectures remains a barrier for high-resolution tasks such as image denoising.

^a  <https://orcid.org/0009-0001-2752-2357>

^b  <https://orcid.org/0009-0003-6446-2669>

^c  <https://orcid.org/0000-0002-0071-958X>

Dilated kernels, also known as atrous convolution or convolution with holes, are a type of convolutional kernel used in deep learning architectures or classical processing. The hypothesis behind this operation is that, by dilating, rather than expanding filters, the region covered by the transformation increases in terms of pixel distance and not pixel density. In recent years, dilated filters proved particularly useful in domains ranging from image segmentation (Yu and Koltun, 2016; Yu et al., 2017) to edge detection (Bogdan et al., 2020; Orhei et al., 2021).

In this paper we propose the Adaptive Kernel Dilation Transformer (**AKDT**), a Transformer architecture that incorporates convolutions within the standard Transformer components. **AKDT** innovates through dilated convolutions that employ a mechanism which allows them to learn and adapt the dilation rate of kernels. Through the use of learnable dilation rate kernels, we can harness a weighted expansion of the informational domain without increasing computational cost.

Through the use of image denoising benchmarks, with real and synthetic noise, we demonstrate **AKDT** has state-of-the-art (SOTA) performance and highlight its computational efficiency.

The main contributions of our work can be summarized as follows:

- **AKDT**, a Transformer architecture tailored to perform highly-effective image denoising by employing the strengths of dilated convolutions.
- A novel approach of using dilated convolutions in Transformer architectures in order to produce dynamic and adaptable modules that tailor to the specifics of the task.
- Two novel components in the Transformer architecture: Noise-Guided Feed-Forward Network (**NG-FFN**) and Noise-Guided Multi-Headed Self-Attention (**NG-MSA**).
- Through quantitative and qualitative experiments, **AKDT** demonstrates SOTA performance on standardized denoising benchmarks.

2 RELATED WORK

Image denoising is a critical area in computer vision that aims to remove noise from corrupted images to recover the original, uncorrupted content (Fattal, 2007; HeK and SUNJ, 2011; Kopf et al., 2008; Michaeli and Irani, 2013). Traditional approaches, often based on CNNs, have been pivotal in advancing early image restoration techniques. These methods leverage spatial hierarchies of learned filters to

address various levels of degradation in images (Tu et al., 2022; Zamir et al., 2022; Zhang et al., 2020; Zamir et al., 2020b; Chen et al., 2021; Zamir et al., 2020a). CNNs have been effective due to their ability to enforce local connectivity and share weights across spatial domains.

With the advent of Transformer architectures, the computer vision paradigm has drastically shifted. Originally developed for natural language processing tasks (Vaswani et al., 2017), Transformers apply self-attention mechanisms to capture long-range dependencies in data, a significant advantage over CNNs when dealing with complex image structures (Dosovitskiy et al., 2021; Ramachandran et al., 2017). In image denoising, Transformers tokenize images and learn intricate relationships, offering enhanced capabilities for handling detailed textures and patterns in high-resolution images (Touvron et al., 2021; Yuan et al., 2021).

Despite their advantages, the computational demands of Transformers increase quadratically with the input size, presenting challenges for practical applications. Recent research has focused on developing more efficient Transformer models to mitigate these issues. Techniques such as locality-constrained self-attention mechanisms introduced in Swin Transformers (Liu et al., 2021) and innovative attention schemes like those in CAT (Chen et al., 2022), which employ rectangular-window self-attention, and channel-wise attention, proposed in Restormer (Zamir et al., 2022), have shown promising results in enhancing the computational efficiency of Transformers.

3 PROPOSED METHOD

In Figure 2, we detail the architecture of **AKDT**, which features a foundational Transformer structure. However, **AKDT** diverges from conventional models by incorporating two novel components: the Noise-Guided Multi-headed Self-attention (**NG-MSA**) and the Noise-Guided Feed-Forward Network (**NG-FFN**). Both blocks utilize our proposed Noise Estimator Module (**NE**), which consists of two Learnable Dilation Rate (**LDR**) submodules.

The process begins as the input image in RGB format is first subjected to a $conv_{3 \times 3}$ layer, which projects the image into a high-dimensional feature space, setting the stage for more complex transformations. Following this initial projection, the enhanced feature map is introduced into the core of the Transformer architecture, which is organized in a U-shaped configuration utilizing skip connections (Ronneberger et al., 2015). Up-sampling and

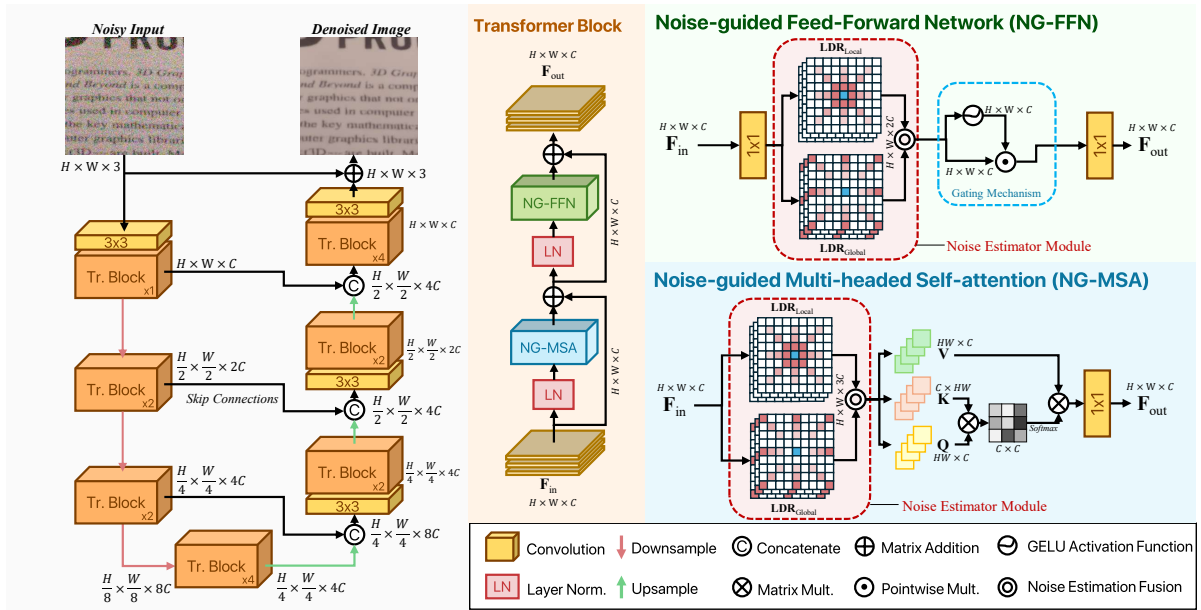


Figure 2: Framework of AKDT. Our Transformer architecture utilizes the proposed Noise Estimator module (NE) within the Noise-Guided Feed-Forward Network (NG-FFN) and the Noise-Guided Multi-headed Self-attention (NG-MSA).

down-sampling are achieved through pixel-shuffle and pixel-unshuffle operations (Shi et al., 2016).

After the feature map traverses through the U-shaped Transformer sequence, it enters an advanced refinement stage. This stage employs additional Transformer blocks that further refine the features, ensuring that even subtle nuances are captured and enhanced. This refinement is critical in restoration, particularly in the case of heavily degraded images.

The refined feature map is then processed through another $conv3 \times 3$ layer, which compresses the high-dimensional features back into the standard RGB color space. Additionally, a residual connection from the original input is incorporated at this final stage. This connection aids in preserving essential image details by allowing the original data to flow directly into the output, thereby enhancing the fidelity of the restored image and ensuring that important textural and color details are maintained.

3.1 Learnable Dilation Rate Module

The **LDR** Module, is defined as the weighted concatenation (\odot) of N convolutions with N different dilation rates. The input feature map is first projected through a $conv1 \times 1$. The projected input $\mathbf{F}_{in} \in \mathbb{R}^{H \times W \times C}$ is then passed through multiple parallel dilated convolutions. Each of the dilated convolution operations has a learnable weight that is adjusted during training. As such, the output feature map \mathbf{F}_{LDR} can be expressed as:

$$\mathbf{F}_{LDR} = conv1 \times 1 \left(\odot_{i=1}^N \alpha_i \times conv3 \times 3_i(\mathbf{F}_{in}) \right) \quad (1)$$

The choice of dilation rates in the **LDR** is important, as they directly influence how the module captures and integrates information from different spatial scales of the input data. This structured approach to combining multiple dilated convolutions, each adjusted for a specific scale of feature extraction, enhances the ability of the model to discern finer details and contributes to more robust and scale-invariant feature representations.

This implementation ensures that the **LDR** Module is easily adjustable for different use-cases. In consequence, our Global and Local **LDR** Modules have the same implementation but employ different-scale dilation rates.

3.2 Noise Estimator Module

The **NE** in the proposed architecture serves a critical role by integrating both global and local context understanding through its unique structure. This module consists of two distinct parallel components: the Global and Local **LDR** modules.

The \mathbf{LDR}_{Global} module is designed to employ higher-scale dilation rates, which enables it to grasp the broader context and underlying patterns across the entire image. We define $\mathbf{F}_{LDR}^{Global}$ as follows:

$$\mathbf{F}_{LDR}^{Global} = \mathbf{LDR}_{Global}(\mathbf{F}_{in}), \quad \mathbf{F}_{LDR}^{Global} \in \mathbb{R}^{H \times W \times C} \quad (2)$$

Conversely, the \mathbf{LDR}_{Local} module utilizes lower-scale dilation rates, focusing on capturing finer, more detailed aspects of the image. This attention to detail

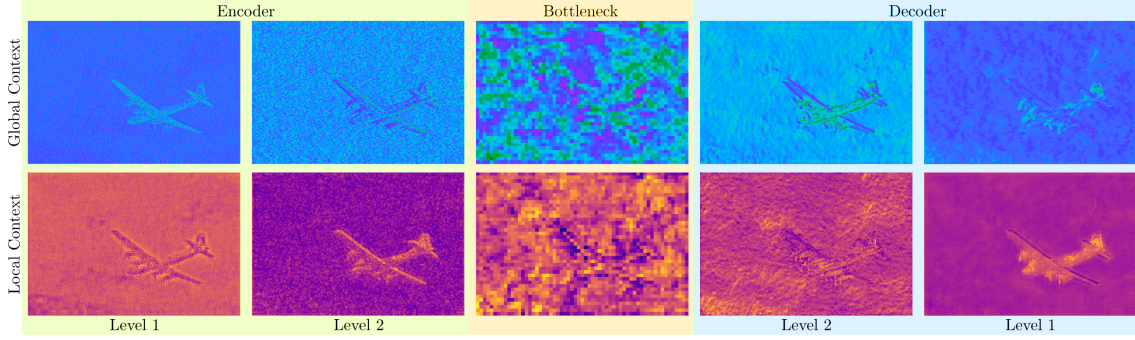


Figure 3: Context captured by the NE at various levels.

is crucial for restoring specific features and textures. $\mathbf{F}_{\text{LDR}}^{\text{Local}}$ is defined as:

$$\mathbf{F}_{\text{LDR}}^{\text{Local}} = \text{LDR}_{\text{Local}}(\mathbf{F}_{\text{in}}), \quad \mathbf{F}_{\text{LDR}}^{\text{Local}} \in \mathbb{R}^{H \times W \times C} \quad (3)$$

Where $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H \times W \times C}$ is the input feature map in both Equations 2 and 3.

Both of these modules operate in parallel, allowing the NE to efficiently combine insights from both the global and local perspectives. The resulting feature map \mathbf{F}_{NE} is represented in equation 4:

$$\mathbf{F}_{\text{NE}} = \mathbf{F}_{\text{LDR}}^{\text{Global}} \odot \mathbf{F}_{\text{LDR}}^{\text{Local}}, \quad \mathbf{F}_{\text{NE}} \in \mathbb{R}^{H \times W \times C} \quad (4)$$

Where \odot is the Noise Estimation Fusion operation, consisting of a convolutional block that merges both local and global noise-modulated contexts.

In Figure 3, we illustrate the context captured by the NE at various stages in the network. AKDT leverages these NE across all stages of processing, ensuring that both local nuances and global patterns are consistently considered, preventing the overspecialization of the model.

3.3 Noise-Guided Attention Block

Given an input image $H \times W \times C$, traditional attention mechanisms reach quadratic complexity with respect to the spatial resolution (i.e., $O(H^2W^2)$). In this work, we utilize a channel-wise attention mechanism (Zamir et al., 2022) to reduce the complexity to $O(C^2)$, allowing our model to work seamlessly with high-resolution images.

In our novel transformer architecture, we prioritize the integration of the NE to refine the attention mechanism. Starting with a layer-normalized input feature $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H \times W \times C}$, we apply the NE to produce an enhanced feature representation \mathbf{F}_{NE} . Utilizing \mathbf{F}_{NE} , we then compute the query (\mathbf{Q}) key (\mathbf{K}), and value (\mathbf{V}) components essential for the attention mechanism as:

$$\mathbf{Q} = W^Q \mathbf{F}_{\text{NE}}, \quad \mathbf{K} = W^K \mathbf{F}_{\text{NE}}, \quad \mathbf{V} = W^V \mathbf{F}_{\text{NE}}, \quad (5)$$

where W^Q , W^K , and W^V denote the learnable parameters for projecting \mathbf{F}_{NE} into the respective components.

This method underscores that our architecture departs from conventional convolution-based approaches, spotlighting the NE role in facilitating a nuanced, context-aware generation of attention components. Following the computation of \mathbf{Q} , \mathbf{K} , and \mathbf{V} , the attention operation is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \odot \text{Softmax}\left(\frac{\mathbf{K}^T \mathbf{Q}}{\alpha}\right), \quad (6)$$

where α is a learnable scaling parameter that fine-tunes the influence of the dot product between \mathbf{K} and \mathbf{Q} prior to the softmax application. Channels are divided into multiple 'heads', resulting in multiple feature maps, as seen in the traditional multi-headed self-attention (Dosovitskiy et al., 2021). This approach ensures that our attention mechanism is both computationally efficient and capable of capturing extensive contextual interactions within the input features.

3.4 Noise-Guided Feed-Forward Network

Leveraging the NE to enrich the input features, our NG-FFN significantly refines the feature transformation process. After applying the NE to the input feature map $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H \times W \times C}$ to obtain $\mathbf{F}_{\text{NE}} = \text{NE}(\mathbf{F}_{\text{in}})$, the NG-FFN employs a gating mechanism (Zamir et al., 2022) which enhances control over the feature transformation. The gating mechanism is articulated through the element-wise multiplication of two parallel transformation paths, where one of the paths incorporates the GELU (Hendrycks and Gimpel, 2016) non-linearity.

The process is mathematically represented as:

$$\mathbf{F}_{\text{NG-FFN}} = \phi(W_1 \mathbf{F}_{\text{NE}}) \odot W_2 \mathbf{F}_{\text{NE}} + \mathbf{F}_{\text{NE}}, \quad (7)$$

where ϕ denotes the GELU activation function, \odot represents element-wise multiplication, and W_1 ,

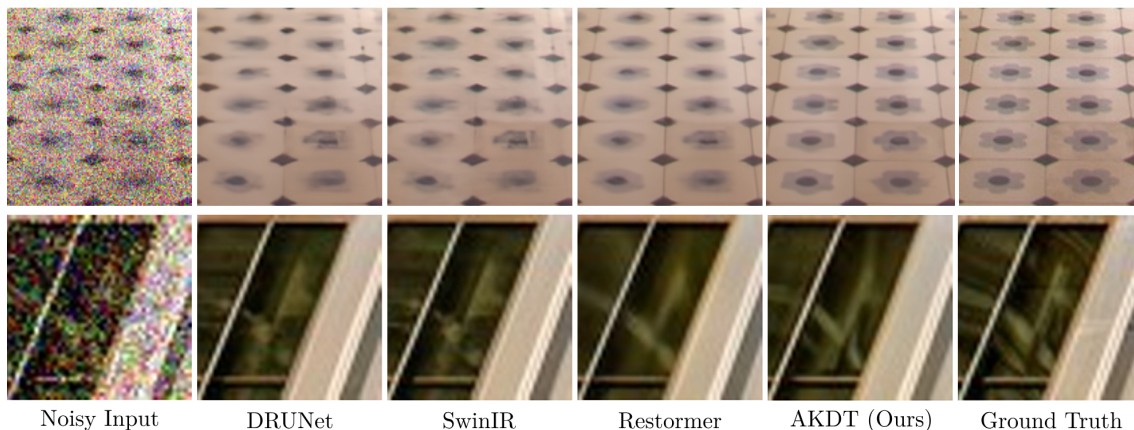


Figure 4: Qualitative results on Color Gaussian Denoising - Urban100 dataset.

W_2 are the learnable parameters of the parallel paths. Here, $F_{\text{NG-FFN}}$ symbolizes the output feature map of the feed-forward network, showcasing an enhanced representation for subsequent processing stages.

This approach distinctly bypasses the conventional reliance on convolutions or fully-connected layers, focusing instead on the synergy between the NE-enhanced features and the gating mechanism.

4 EXPERIMENTS AND RESULTS

4.1 Implementation Details

In our evaluation, we select image denoising benchmarks like CBSD68 (Martin et al., 2001), Urban100 (Huang et al., 2015), and McMaster (Zhang et al., 2011) for synthetic cases, alongside SIDD (Abdelhamed et al., 2018) for real noise environments.

Our transformer-based model is architecturally devised with a quad-level framework, hosting $[1, 2, 2, 4]$ transformer blocks, complemented by $[1, 2, 4, 8]$ attention heads at respective levels. The attention dimensions are assigned as $[34, 68, 134, 268]$ across the levels. A set of 3 refinement blocks are employed at the end of the encoder-decoder setup.

A distinctive feature of our methodology is the adoption of progressive training, starting with image patches of 128×128 pixels and gradually advancing through sizes of 194×194 , 256×256 , 320×320 , and ultimately 384×384 pixels. This strategy accommodates a comprehensive learning spectrum from local to global image characteristics and enhances the model’s adaptability to diverse noise patterns.

To improve the training dynamics, random rotations and flipping are employed as data augmentation techniques, strengthening the robustness of the

model. We utilize AdamW as the optimizer, parameterized by $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $1e-4$, across 300K iterations. The learning rate begins at $3e-4$ and is methodically tapered to $1e-6$ via cosine annealing (Loshchilov and Hutter, 2017), ensuring a smooth and effective model refinement.

Our evaluation framework employ the classical PSNR (peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measure) metrics to quantitatively measure the denoising capability of the model, ensuring a thorough appraisal of its capability to reconstruct clean images from noisy inputs. We also utilize GMACs (Giga Multiply-Accumulate Operations) to highlight the computational complexity of different methods.

4.2 Results

Real Image Denoising. Table 2 shows the performance of multiple models on SIDD dataset. Restormer achieves the highest PSNR of 40.02, making it potentially the best choice. On the other hand, our **AKDT** model showcases outstanding efficiency with the lowest computational demand, measured at 56.15 GMACs, and the highest SSIM of 0.961. The results highlight our method and showcase its SOTA efficiency. Qualitative results in Fig. 4 further illustrate the performance of our method despite its reduced computational load.

Gaussian Denoising. Table 1 shows PSNR scores of different models on various benchmark datasets. We test color gaussian denoising capabilities, by including well determined synthetic noise levels (σ) 15, 25 and 50. Qualitative results are presented in Fig. 5.

From the data presented we can conclude that **AKDT** emerges as a standout, particularly in the CBSD68, McMaster, and Urban100 datasets, where it consistently outperforms other models across all

Table 1: Gaussian Color Denoising benchmarks results (PSNR). “-”: not reported. **Red** and **blue** represent best and second-best values, respectively.

Model	CBSD68			McMaster			Urban100			MACs
	$\sigma=15$	$\sigma=25$	$\sigma=50$	$\sigma=15$	$\sigma=25$	$\sigma=50$	$\sigma=15$	$\sigma=25$	$\sigma=50$	
DnCNN(Zhang et al., 2017)	33.90	31.24	27.95	33.45	31.52	28.62	32.98	30.81	27.59	37 G
FFDNet(Zhang et al., 2018)	33.87	31.21	27.96	34.66	32.35	29.18	33.83	31.40	28.05	-
DSNet(Peng et al., 2019)	33.91	31.28	28.05	34.67	32.40	29.28	-	-	-	-
BRDNet(Tian et al., 2020)	34.10	31.43	28.16	35.08	32.75	29.52	34.42	31.99	28.56	-
DRUNet(Zhang et al., 2021)	34.30	31.69	28.51	35.40	33.14	30.08	34.81	32.60	29.61	144 G
SwinIR(Liang et al., 2021)	34.42	31.78	28.56	35.61	33.20	30.22	35.13	32.90	29.82	759 G
Restormer(Zamir et al., 2022)	34.40	31.79	28.60	35.61	33.34	30.30	35.13	32.96	30.02	141 G
NAFNet-RCD(Zhang et al., 2023)	34.14	31.49	28.26	35.11	32.84	29.81	34.45	32.12	29.09	-
GRL-B(Li et al., 2023)	34.45	31.82	28.62	35.73	33.46	30.36	35.54	33.35	30.46	-
DCANet(Wu et al., 2024)	34.05	31.45	28.28	34.84	32.62	29.59	-	-	-	75 G
HWFormer(Tian et al., 2024)	-	-	-	35.64	33.36	30.24	35.26	33.10	30.14	303 G
AKDT (Ours)	34.64	31.94	28.68	36.71	34.21	30.95	35.63	33.14	29.82	56 G

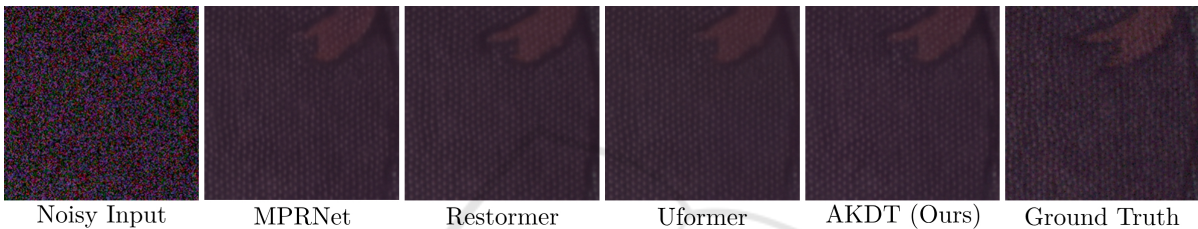


Figure 5: Qualitative results on Real Image Denoising - SIDD dataset.

Table 2: Performance of different models on the SIDD dataset. **Red** and **blue** text indicate best and second-best values, respectively.

Model	PSNR \uparrow	SSIM \uparrow	GMACs \downarrow
MPRNet (Zamir et al., 2021)	39.17	0.958	588
CycleISP (Zamir et al., 2020a)	39.52	0.957	189.5
HINet (Chen et al., 2021)	39.99	0.23	170.7
MAXIM (Tu et al., 2022)	39.96	0.960	169.5
Restormer (Zamir et al., 2022)	40.02	0.960	141
MIRNet (Zamir et al., 2020b)	39.72	0.959	786
Uformer (Wang et al., 2021)	39.89	0.960	88.8
PCFormer (Wan et al., 2023)	39.80	0.959	152
DCANet (Wu et al., 2024)	39.27	0.956	75.30
AKDT (Ours)	39.70	0.961	56.15

noise levels. This underscores the effectiveness of **AKDT** in handling color denoising tasks, even in challenging noisy environments.

The comparative analysis also brings to light the trade-offs between performance and computational efficiency. For instance, SwinIR, while achieving near top marks, demands a substantial computational cost with 759 GMACs, whereas **AKDT** maintains a competitive edge with significantly lower computational needs (56 GMACs). This aspect is crucial for practical applications where computational resources are limited or cost-effectiveness is a priority. The data suggests that **AKDT** not only provides denoising capabilities but does so with remarkable efficiency.

From a qualitative perspective, **AKDT** demon-

strates its superior performance by producing sharper images while not introducing artifacts, as seen in the outputs of DRUNet and SwinIR, while also not over-smoothing details, as in the case of Restormer.

By leveraging our proposed **NE**, we construct a robust and efficient Transformer architecture (**AKDT**) that sets SOTA performance on various denoising benchmarks.

5 ABLATION STUDY

As ablation studies, we provide details into our architectural choices, accompanied by performance metrics on the SIDD dataset.

In Table 3 we present the impact of various **LDR** configurations within the **NE**, employing specialized paths for local and/or global context (i.e. Dual-Path). The Dual-Path implementation enforces the importance of specialization upon the model, by having dedicated dilation rates for both types of contexts, preventing excessive focus on one of the types of information.

Table 4 shows experiments with different compression/expansion rates within the **NE**. C represents the input dimensions of the **NE**. $C \times K$ means that the **NE** has inner dimensions equal to those of the input multiplied by a factor of K . As the results suggest, performing feature compression within the **NE**

Table 3: Impact of LDR.

Local	Global	PSNR
✓		39.47
	✓	39.32
✓	✓	39.70

Table 4: Impact of feature expansion.

Change	PSNR
$C \rightarrow C \times 0.125$	39.28
$C \rightarrow C \times 0.25$	39.70
$C \rightarrow C \times 0.5$	39.67
$C \rightarrow C \times 2$	38.63

Table 5: Transformer Block variations impact.

MSA	FFN	PSNR	GMACs
V	V	35.29	51.13
NG	V	37.48	53.27
V	NG	36.46	65.35
NG	NG	39.59	67.49
NG	NG+G	39.70	56.16

performs better. This study suggests that the **NE** is able to perform best by extracting only the most relevant features from both the Global and Local contexts, through the $\mathbf{LDR}_{\text{Global}}$ and $\mathbf{LDR}_{\text{Local}}$ respectively.

Table 5 presents a more extensive study on potential Transformer Block implementations. *V* represents vanilla MSA/FFN implementations. **NG** is the noise-guidance integration. **+G** represents the gating mechanism. As illustrated, the addition of the **NE** used for the **NG-MSA** improves performance over the vanilla (*V*) MSA by 2.19 dB. Similarly, the **NG-FFN** outperforms vanilla FFN by 1.17 dB. Furthermore, the gating mechanism (**+G**) improves performance over the non-gated **NG-FFN** by 0.11 dB while also reducing GMACs by 17%.

The ablation studies underscore the critical impact of our architectural choices on the performance of the Adaptive Kernel Dilation Transformer (**AKDT**).

6 CONCLUSION & FUTURE WORKS

We introduced a novel approach of utilising learnable dilation convolutions, the **LDR** block, and used it to develop a mechanism that is able to effectively distinguish and learn between local and global noise patterns, the **NE**. We proposed the integration of the **NE** into the attention and feed-forward mechanisms prevalent in Transformer architectures, producing an efficient SOTA model.

Our comprehensive evaluation demonstrates that our proposed method has superior performance and computational efficiency in comparison to existing denoising methods. Furthermore, our experiments serve as theoretical backing to the proposed design, thereby proving the positive impact of learnable dilation rate convolutions in Transformer architectures.

REFERENCES

Abdelhamed, A., Lin, S., and Brown, M. S. (2018). A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700.

Ancuti, C. O., Ancuti, C., De Vleeschouwer, C., and Bekaert, P. (2017). Color balance and fusion for underwater image enhancement. *IEEE Transactions on image processing*, 27(1):379–393.

Bogdan, V., Bonchis, C., and Orhei, C. (2024). An image sharpening technique based on dilated filters and 2d-dwt image fusion. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*, pages 591–598. INSTICC, SciTePress.

Bogdan, V., Bonchiş, C., and Orhei, C. (2020). Custom dilated edge detection filters. *Journal of WSCG*, 28:161–168.

Brateanu, A., Balmez, R., Avram, A., and Orhei, C. (2024). Lyt-net: Lightweight yuv transformer-based network for low-light image enhancement. *arXiv preprint arXiv:2401.15204*.

Chen, L., Lu, X., Zhang, J., Chu, X., and Chen, C. (2021). Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 182–192.

Chen, Z., Zhang, Y., Gu, J., Kong, L., Yuan, X., et al. (2022). Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Fattal, R. (2007). Image upsampling via imposed edge statistics. In *ACM SIGGRAPH 2007 papers*, pages 95–es.

HeK, M. and SUNJ, T. X. O. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341.

Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv:1606.08415*.

Huang, J.-B., Singh, A., and Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *CVPR*.

Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., and Lischinski, D. (2008). Deep photo: Model-based photograph enhancement and viewing. *ACM transactions on graphics (TOG)*, 27(5):1–10.

Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., and Van Gool, L. (2023). Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Con-*

- ference on Computer Vision and Pattern Recognition, pages 18278–18289.
- Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). SwinIR: Image restoration using swin transformer. In *ICCV Workshops*.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030*.
- Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *ICLR*.
- Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*.
- Michaeli, T. and Irani, M. (2013). Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952.
- Orhei, C., Bogdan, V., Bonchis, C., and VasIU, R. (2021). Dilated filters for edge-detection algorithms. *Applied Sciences*, 11(22):10716.
- Orhei, C. and VasIU, R. (2023). An analysis of extended and dilated filters in sharpening algorithms. *IEEE Access*, 11:81449–81465.
- Peng, Y., Zhang, L., Liu, S., Wu, X., Zhang, Y., and Wang, X. (2019). Dilated residual networks with symmetric skip connection for image denoising. *Neurocomputing*.
- Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In *MICCAI*.
- Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*.
- Tian, C., Xu, Y., and Zuo, W. (2020). Image denoising using deep cnn with batch renormalization. *Neural Networks*.
- Tian, C., Zheng, M., Lin, C.-W., Li, Z., and Zhang, D. (2024). Heterogeneous window transformer for image denoising. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(11):6621–6632.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *ICML*.
- Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. (2022). Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.
- Wan, Y., Shao, M., Cheng, Y., Meng, D., and Zuo, W. (2023). Progressive convolutional transformer for image restoration. *Engineering Applications of Artificial Intelligence*, 125:106755.
- Wang, W., Wu, X., Yuan, X., and Gao, Z. (2020). An experiment-based review of low-light image enhancement methods. *Ieee Access*, 8:87884–87917.
- Wang, Z., Cun, X., Bao, J., and Liu, J. (2021). Uformer: A general u-shaped transformer for image restoration. *arXiv:2106.03106*.
- Wu, W., Lv, G., Duan, Y., Liang, P., Zhang, Y., and Xia, Y. (2024). Dual convolutional neural network with attention for image blind denoising.
- Yu, F. and Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*.
- Yu, F., Koltun, V., and Funkhouser, T. (2017). Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480.
- Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E., Feng, J., and Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv:2101.11986*.
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*.
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. (2020a). Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2696–2705.
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. (2020b). Learning enriched features for real image restoration and enhancement. In *ECCV*.
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. (2021). Multi-stage progressive image restoration. In *CVPR*.
- Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., and Timofte, R. (2021). Plug-and-play image restoration with deep denoiser prior. *TPAMI*.
- Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., and Li, H. (2020). Deblurring by realistic blurring. In *CVPR*.
- Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*.
- Zhang, K., Zuo, W., and Zhang, L. (2018). FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *TIP*.
- Zhang, L., Wu, X., Buades, A., and Li, X. (2011). Color demosaicking by local directional interpolation and non-local adaptive thresholding. *JEI*.
- Zhang, Z., Jiang, Y., Shao, W., Wang, X., Luo, P., Lin, K., and Gu, J. (2023). Real-time controllable denoising for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14028–14038.