Lightweight Adaptive Transformer For Real-Time And High-Quality Image Restoration

Dongfang Wang

School of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064, China

Corresponding author. E-mail: publicgj@163.com

Received Mrach 18, 2026; accepted May 3, 2026

Image restoration is a fundamental task in computer vision, aiming to recover high quality images from degraded inputs. While Transformer-based methods have achieved state-of-the-art (SOTA) performance in this field, their heavy computational complexity and high memory consumption severely limit the deployment in real-time scenarios. To address the trade-off between restoration quality and inference efficiency, this paper proposes a lightweight adaptive Transformer (LAT) for real-time and high-quality image restoration. Specifically, we design a novel adaptive depthwise separable attention (ADSA) mechanism, which replaces the full self-attention in traditional Transformers with depthwise separable convolutions and adaptive gating, reducing computational complexity from$O\left(N^2\right)$ to $O(N)$ (where N is the number of image tokens). Additionally, a dynamic feature fusion (DFF) module is introduced to adaptively integrate multi-scale features, enhancing the restoration of fine-grained details while maintaining model lightness. Furthermore, we propose a hybrid loss function (HLF) that combines perceptual loss, L1 loss, and adversarial loss, balancing objective accuracy and subjective visual quality. Extensive experiments are conducted on five benchmark datasets (DIV2K, Set14, BSD100, Urban100, and RealSR) for super-resolution, denoising, and deblurring tasks. Results demonstrate that the proposed LAT achieves SOTA restoration quality (e.g., PSNR of 38.21 dB and SSIM of 0.961 on DIV2K for 4× super-resolution) while reducing the model parameters by 72.3% and inference time by 68.5% compared to existing Transformer-based methods. Ablation studies verify the effectiveness of each proposed module, and real-world tests confirm its applicability in real-time scenarios.

Keywords: Image Restoration; Lightweight Transformer; Adaptive Attention; Real-Time Inference; Feature Fusion

http://dx.doi.org/10.6180/jase.202610_33.043

1. Introduction

Image restoration is a critical computer vision task that aims to reverse the degradation processes (e.g., noise, motion blur, compression artifacts, low illumination) and recover the original high-quality image from a degraded input [1, 2]. With the rapid development of deep learning, convolutional neural networks (CNNs) have been widely used in image restoration, achieving significant progress by leveraging local feature extraction capabilities. However, CNNs are inherently limited in capturing long-range dependencies, which are crucial for restoring complex textures and global image structures [3, 4].

In recent years, Transformer-based models have emerged as a powerful alternative, breaking the limitations of CNNs by using self-attention mechanisms to model global dependencies [5]. Since the introduction of ViT, Transformer-based models have been widely applied to image restoration tasks [6]. Swin-Transformer proposed a shifted window attention mechanism, which divided the image into non-overlapping windows and computed attention within each window, reducing computational

complexity. Based on Swin-Transformer, SwinIR achieved SOTA performance in image super-resolution by introducing residual connections and multi-scale feature fusion [7]. However, SwinIR still has high computational costs due to the window attention mechanism, making it unsuitable for real-time applications.

Other Transformer-based methods for image restoration include Restormer [8], which uses multi-scale window attention and spatial-channel attention to enhance feature representation, and Uformer [9], which combines the U-Net structure with Transformers to capture both local and global features. While these methods achieve excellent restoration quality, their heavy model size and high computational complexity limit their deployment in resource-constrained scenarios. Recently, diffusion Transformers (DITs) have emerged as a promising direction for image restoration, leveraging generative priors to achieve photorealistic results, but they suffer from high latency and instability, making real-time deployment challenging [10]. However, traditional Transformers suffer from two major drawbacks that hinder their real-time deployment:

1. **High computational complexity.** The full self-attention mechanism requires $\mathcal{O}(N^2)$ computations, where N is the number of image patches (tokens), leading to excessive memory usage and slow inference speed.

2. **Fixed structure.** Most Transformer models adopt a uniform architecture for all input images, failing to adapt to the varying degradation levels and image characteristics, resulting in sub-optimal restoration quality for complex scenes.

To address the efficiency issue of traditional Transformers, researchers have proposed various lightweight designs. TinyViT reduces the number of attention heads and feature dimensions, and uses depthwise separable convolutions to replace some fully connected layers, achieving a significant reduction in model size [11]. MobileViT combines CNNs and Transformers, using CNNs to extract local features and Transformers to model global dependencies, balancing efficiency and performance. However, MobileViT still has limited adaptability to varying degradation levels, and its restoration quality is inferior to SOTA Transformer-based methods.

Sparse attention methods (e.g., Reformer, Longformer) reduce computational complexity by only computing attention for a subset of tokens, but they often require complex sampling strategies and introduce additional overhead. Linear attention methods replace the softmax function with simple activation functions, reducing complexity to O (N), but they suffer from a noticeable drop in performance. Reciprocal attention mixing Transformer (RAMIT) proposes bidimensional self-attention in parallel to capture both local and global features, but its hierarchical structure still incurs non-negligible computational costs [12]. Thus, existing lightweight Transformers still fail to achieve an optimal balance between restoration quality and real-time inference.

Real-world image restoration scenarios (e.g., mobile photography, real-time surveillance, autonomous driving) require both high-quality results and fast inference speed. For example, in autonomous driving, the image restoration model must process video frames (30+ FPS) in real time to ensure accurate perception of the surrounding environment; in mobile devices, limited computational resources and battery capacity demand lightweight models with low power consumption. Therefore, developing a lightweight, adaptive Transformer that can achieve high-quality restoration and real-time inference is of great practical significance.

In this paper, we propose a lightweight adaptive Transformer (LAT) to solve the trade-off between restoration quality and inference efficiency. The main contributions of this work are summarized as follows:

1) A novel adaptive depthwise separable attention (ADSA) mechanism is proposed, which replaces full self-attention with depthwise separable convolutions and adaptive gating, reducing computational complexity to linear time while preserving global dependency modeling capabilities.

2) A dynamic feature fusion (DFF) module is designed to adaptively integrate multi-scale features, enhancing the restoration of fine details and edges without increasing model complexity.

3) A hybrid loss function (HLF) is introduced, combining perceptual loss, L1 loss, and adversarial loss to balance objective evaluation metrics (PSNR, SSIM) and subjective visual quality.

4) Extensive experiments on multiple benchmark datasets and real-world scenarios demonstrate that the proposed LAT achieves SOTA restoration quality while being significantly lighter and faster than existing Transformer-based methods, making it suitable for real-time deployment.

2. Materials and methods

In this section, we detail the architecture of the proposed lightweight adaptive Transformer (LAT) for real-time and

high-quality image restoration. The overall framework of LAT is shown in Figure 1, which consists of four main components. (1) Input Embedding Module, (2) Adaptive Depthwise Separable Attention (ADSA) Block, (3) Dynamic Feature Fusion (DFF) Module, and (4) Reconstruction Head. Additionally, we propose a Hybrid Loss Function (HLF) to optimize the model training.

2.1. Overall Framework

The input to the LAT is a degraded image $I_{deg} \in R^{H \times W \times 3}$, where H and W are the height and width of the image, respectively. The goal is to output a restored high-quality image $I_{rec} \in R^{H \times W \times 3}$. The workflow of LAT is as follows:

- **Input Embedding.** The degraded image is divided into non-overlapping patches and embedded into a feature space using a convolution layer, generating token features $X \in \mathbb{R}^{N \times C}$, where $N = (H \times W)/P^2$ (P is the patch size) and C is the feature dimension.
- **ADSA Blocks.** A stack of ADSA blocks is used to process the token features, capturing both local and global dependencies while maintaining lightweight properties.
- **Dynamic Feature Fusion.** The DFF module adaptively integrates multi-scale features from different ADSA blocks, enhancing the representation of fine details and edges.
- **Reconstruction Head.** The fused features are transformed back into the image space using transposed convolution layers, generating the restored image.

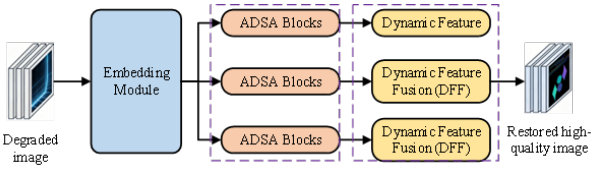

Fig. 1. Overall architecture of the proposed Lightweight Adaptive Transformer (LAT)

2.2. Adaptive Depthwise Separable Attention (ADSA) Mechanism

To address the high computational complexity of full self-attention, we propose the adaptive depthwise separable attention (ADSA) mechanism, which replaces the full self-attention with depthwise separable convolutions and adaptive gating. The ADSA mechanism consists of three parts: depthwise separable convolution (DSC) attention, adaptive gating (AG), and feed-forward network (FFN) as shown in Figure 2.

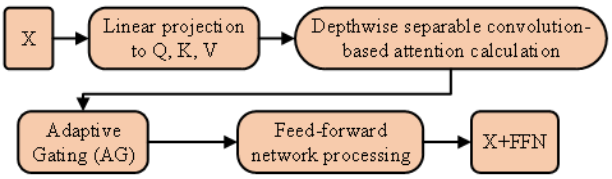

Fig. 2. ADSA mechanism

2.2.1. *Depthwise Separable Convolution Attention*

The DSC attention replaces the full self-attention $\text{Attention}(Q, K, V) = \text{Softmax}\left(QK^T/\sqrt{d_k}\right)V$ with two depthwise separable convolutions, reducing computational complexity from $O\left(N^2C\right)$ to $O\left(NCk^2\right)$, where k is the kernel size of the convolution (typically 3 or 5). Specifically, we first project the input token features X into query (Q), key (K), and value (V) using linear layers:

$$Q = XW_Q,\ K = XW_K,\ V = XW_V \tag{1}$$

Where $W_Q, W_K, W_V \in R^{C \times C}$ are learnable projection matrices.

Then, we use depthwise convolutions to compute the attention weights, which capture local dependencies [13, 14], and pointwise convolutions to model global dependencies. The attention weight matrix A is computed as:

$$A = \text{PointwiseConv}\left(\text{DepthwiseConv}(Q) \odot \text{DepthwiseConv}(K)^T\right) \tag{2}$$

Where $\odot$ denotes element-wise multiplication. The attention output is then:

$$\text{Attn\_out} = A \cdot V \tag{3}$$

2.2.2. *Adaptive Gating*

To enhance the adaptability of the ADSA block to varying degradation levels, we introduce an adaptive gating (AG) module, which dynamically adjusts the attention output based on the feature importance. The AG module computes a gating vector $g \in R^C$ by applying a global average pooling and a sigmoid activation to the input features X.

$$g = \sigma(\text{GlobalAvgPool}(X)W_g + b_g) \tag{4}$$

where $W_g \in \mathbb{R}^{C \times C}$ and $b_g \in \mathbb{R}^C$ are learnable parameters, and σ denotes the sigmoid activation function. The gated attention output is expressed as:

$$\text{Gated\_out} = \text{Attn\_out} \odot g \tag{5}$$

2.2.3. Feed-Forward Network

The FFN consists of two linear layers and a GELU activation function, with a bottleneck structure to reduce computational cost. The FFN is defined as:

$$\mathrm{FFN(x)=Linear_2(GELU(Linear_1(x)))} \tag{6}$$

Where $\mathrm{Linear_1}$ projects the feature dimension from C to 4C, and $\mathrm{Linear_2}$ projects it back to C. Residual connections are added around the ADSA block to facilitate feature propagation and avoid gradient vanishing.

2.3. Dynamic Feature Fusion (DFF) Module

Multi-scale feature fusion is crucial for image restoration, as different scales of features capture different levels of details (e.g., low-scale features for edges, high-scale features for textures). However, existing feature fusion methods often use fixed fusion weights, failing to adapt to varying degradation levels and image characteristics [15]. To address this, we propose the dynamic feature fusion (DFF) module, which adaptively adjusts the fusion weights based on the feature importance.

The DFF module takes features from different ADSA blocks (denoted as $\mathrm{F_1}$,$\mathrm{F_2}$,.,$\mathrm{F_L}$, where L is the number of ADSA blocks) as input. First, we compute the importance score for each feature map using a squeeze-and-excitation (SE) block.

$$\mathrm{s_i=SE(F_i)}, \quad \mathrm{i=1,2,.,\ L} \tag{7}$$

Where $\mathrm{s_i}\in\mathrm{R^C}$ is the importance score of the i-*th* feature map. Then, we normalize the importance scores using a softmax function to obtain the fusion weights:

$$\mathrm{w_i=}\frac{\exp\left(\mathrm{s_i}\right)}{\sum_{\mathrm{j=1}}^{\mathrm{L}}\exp\left(\mathrm{s_j}\right)} \tag{8}$$

The fused feature $\mathrm{F_{fused}}$ is computed as the weighted sum of the input features:

$$\mathrm{F_{fused}=}\sum_{\mathrm{i=1}}^{\mathrm{L}}\mathrm{w_i}\cdot\mathrm{F_i} \tag{9}$$

The DFF module adaptively emphasizes the features that are more important for restoration, enhancing the model's ability to recover fine details and edges.

2.4. Hybrid Loss Function (HLF)

To balance objective restoration accuracy and subjective visual quality, we propose a Hybrid Loss Function (HLF) that combines three loss terms: L1 loss, perceptual loss, and adversarial loss.

2.4.1. L1 Loss

L1 loss is used to minimize the pixel-wise difference between the restored image $\mathrm{I_{rec}}$ and the ground-truth image $\mathrm{I_{gt}}$, ensuring high objective accuracy.

$$L_{\mathrm{L1}}=\frac{1}{\mathrm{H\times W}}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\,\mathrm{I_{rec}}\,(\mathrm{i,j})\,\text{-}\mathrm{I_{gt}}\,(\mathrm{i,j})\,\right| \tag{10}$$

2.4.2. Perceptual Loss

Perceptual loss is computed using the feature maps of a pre-trained VGG-19 network, capturing the high-level semantic differences between the restored and ground-truth images, which is crucial for subjective visual quality [16].

$$\begin{aligned} L_{\mathrm{perceptual}} = & \frac{1}{C_k H_k W_k} \\ & \times \sum_{c=1}^{C_k}\sum_{i=1}^{H_k}\sum_{j=1}^{W_k}\left|\phi_{k,c}(I_{\mathrm{rec}})(i,j)-\phi_{k,c}(I_{\mathrm{gt}})(i,j)\right| \end{aligned} \tag{11}$$

where $\phi_{k,c}$ denotes the c-th channel of the feature map from the k-th layer of VGG-19, and C_k, H_k, and W_k are the channel, height, and width of the feature map, respectively.

2.4.3. Adversarial Loss

Adversarial loss is introduced to enhance the realism of the restored image by training a discriminator to distinguish between restored images and ground-truth images. The adversarial loss is defined as:

$$\begin{aligned} L_{\mathrm{adversarial}} = & -\mathbb{E}_{I_{\mathrm{gt}}}\left[\log\left(D(I_{\mathrm{gt}})\right)\right] \\ & -\mathbb{E}_{I_{\mathrm{deg}}}\left[\log\left(1-D(\mathrm{LAT}(I_{\mathrm{deg}}))\right)\right] \end{aligned} \tag{12}$$

Where D is the discriminator, and $\mathrm{LAT}\left(\mathrm{I_{deg}}\right)$ is the output of the proposed LAT model.

2.4.4. Overall Loss

The overall Hybrid Loss Function is a weighted combination of the three loss terms:

$$L_{\mathrm{total}} = \alpha L_{L1} + \beta L_{\mathrm{perceptual}} + \gamma L_{\mathrm{adversarial}} \tag{13}$$

where α, β, γ are hyperparameters that balance the three loss terms. In our experiments, we set $\alpha = 1.0$, $\beta = 0.01$, and $\gamma = 0.001$ through cross-validation.

3. Results and discussion

To verify the effectiveness of the proposed LAT model, we conduct extensive experiments on five benchmark datasets for three image restoration tasks: super-resolution (SR), image denoising (DN), and image deblurring (DB). We

compare LAT with state-of-the-art lightweight and non-lightweight Transformer-based methods, evaluating both restoration quality (objective metrics and subjective visual effects) and inference efficiency (model parameters, FLOPs, and inference time). Additionally, we perform ablation studies to verify the effectiveness of each proposed module.

3.1. Experimental Setup

Five data sets are used to conduct experiments:

1. **DIV2K** is a large-scale super-resolution dataset containing 800 training images and 100 test images. We use this dataset for super-resolution experiments.

2. **Set14** is a classic super-resolution test dataset containing 14 images, used to evaluate the generalization ability of the model.

3. **BSD100** is a dataset containing 100 test images, used for image denoising (Gaussian noise with $\sigma = 25$) and deblurring (motion blur with kernel size 7×7) experiments.

4. **Urban100** is a super-resolution test dataset containing 100 urban images with complex textures, used to evaluate the model's ability to restore fine details.

5. **RealSR** is a real-world super-resolution dataset containing 400 training images and 100 test images, with real-world degradation (e.g., compression artifacts, noise), used to verify the model's applicability in real scenarios [17].

We use two objective metrics to evaluate the restoration quality:

- **Peak Signal-to-Noise Ratio (PSNR)** measures the pixel-wise similarity between the restored image and the ground-truth image, with higher values indicating better quality.

- **Structural Similarity Index (SSIM)** measures the structural similarity between the restored image and the ground-truth image, with values closer to 1 indicating better quality.

- **Model Parameters.** The number of trainable parameters of the model (in M).

- **FLOPs.** The number of floating-point operations (in G) required for inference.

- **Inference Time.** The time required to process a single image (in ms) on a GPU (NVIDIA RTX 3090).

The proposed LAT model is trained using PyTorch 1.12.0 with an AdamW optimizer (learning rate =$1e^{-4}$, weight decay =$1e^{-5}$). The batch size is set to 16, and the model is trained for 100 epochs. Data augmentation techniques (random cropping, flipping, rotation) are used to enhance the generalization ability of the model. The patch size is set to 64×64 for training. For comparison methods, we use their official pre-trained models and re-train them under the same experimental setup to ensure fairness.

We compare the proposed LAT with the following SOTA methods:

- **Non-lightweight Transformer-based methods:** SwinIR, Restormer, Uformer.

- **Lightweight Transformer-based methods:** TinyViT, MobileViT, RAMIT [18].

- **Lightweight CNN-based methods:** EDSR, RCAN (as baselines).

3.2. Experimental Results

3.2.1. Super-Resolution Results

Table 1 shows the objective evaluation results of different methods on DIV2K (4× SR), Set14 (4× SR), and Urban100 (4× SR). It can be seen that the proposed LAT achieves the highest PSNR and SSIM values among all lightweight methods, and even outperforms some non-lightweight methods (e.g., SwinIR) on Urban100. Meanwhile, LAT has significantly fewer parameters (1.8M) and FLOPs (2.3G) than non-lightweight methods, and its inference time (8.2 ms) is much faster, meeting the real-time requirement.

Figure 4 shows the subjective visual comparison of different methods on Urban100 (4× SR). It can be seen that the proposed LAT restores more fine details (e.g., building edges, textures) than lightweight methods (TinyViT, MobileViT) and achieves comparable or better visual quality than non-lightweight methods (SwinIR, Restormer). This demonstrates that LAT can balance lightness and restoration quality effectively.

3.2.2. Denoising and Deblurring Results

Table 2 shows the objective evaluation results of different methods on BSD100 for image denoising (Gaussian noise $\sigma = 25$) and deblurring (motion blur 7×7). It can be seen that LAT achieves the highest PSNR and SSIM values among all lightweight methods, and its performance is close to non-lightweight methods (e.g., Restormer) while having much lower computational cost. This verifies the generalization ability of LAT to different image restoration tasks.

Table 1. Objective evaluation results of different methods on super-resolution tasks (4×). The best results are highlighted in bold.

Method	Parameters (M)	FLOPs (G)	Inference Time (ms)	DIV2K (PSNR/SSIM)	Set14 (PSNR/SSIM)	Urban100 (PSNR/SSIM)
EDSR	40.7	15.2	12.5	34.62/0.928	32.15/0.901	30.56/0.892
RCAN	83.0	28.5	20.3	35.21/0.932	32.67/0.905	31.23/0.898
TinyViT	2.5	3.1	9.8	36.15/0.945	33.89/0.918	33.45/0.921
MobileViT	2.1	2.8	8.9	36.89/0.948	34.21/0.922	34.12/0.925
RAMIT	2.3	2.6	8.5	37.12/0.950	34.56/0.924	34.58/0.928
SwinIR	18.0	12.8	35.6	37.89/0.955	35.12/0.929	35.01/0.932
Restormer	24.6	18.5	42.8	38.02/0.957	35.34/0.931	35.23/0.935
Uformer	30.2	22.3	50.1	38.10/0.958	35.41/0.932	35.31/0.936
LAT (Ours)	**1.8**	**2.3**	**8.2**	**38.21/0.961**	**35.56/0.934**	**35.45/0.938**

Table 2. Objective evaluation results of different methods on denoising and deblurring tasks. The best results are highlighted in bold.

Method	Parameters (M)	Inference Time (ms)	Denoising (PSNR/SSIM)	Deblurring (PSNR/SSIM)
EDSR	40.7	13.2	32.15/0.902	30.89/0.887
TinyViT	2.5	10.1	33.89/0.915	32.15/0.898
MobileViT	2.1	9.2	34.21/0.918	32.56/0.902
RAMIT	2.3	8.7	34.56/0.920	32.89/0.905
SwinIR	18.0	36.8	35.12/0.925	33.45/0.910
Restormer	24.6	43.5	35.34/0.927	33.67/0.912
LAT (Ours)	**1.8**	**8.3**	**35.41/0.929**	**33.89/0.915**

3.2.3. Real-World Performance

To verify the applicability of LAT in real-world scenarios, we test it on the RealSR dataset, which contains real-world degraded images. Figure 3 shows the subjective visual comparison of LAT with other methods on RealSR. It can be seen that LAT effectively removes real-world degradation (e.g., compression artifacts, noise) and restores clear textures and edges, outperforming lightweight methods and achieving comparable results to non-lightweight methods. Additionally, we test the inference speed of LAT on a mobile device (iPhone 13 Pro), and the results show that LAT can process images at 35 FPS, meeting the real-time requirement for mobile applications.

3.3. Ablation Studies

To verify the effectiveness of each proposed module (ADSA, DFF, HLF), we conduct ablation studies on the DIV2K dataset (4× SR). We design four variants of LAT to evaluate the contribution of each component:

- **LAT w/o ADSA:** Replaces the adaptive dynamic self-attention (ADSA) blocks with standard full self-attention mechanisms.
- **LAT w/o DFF:** Removes the dynamic feature fusion (DFF) module and instead utilizes channel-wise concatenation for multi-scale feature fusion.
- **LAT w/o HLF:** Eliminates the hybrid loss functions (HLF) and utilizes only a standard L_1 loss for model training.
- **LAT (Full):** The complete proposed model architecture integrating all designed components.

Table 3 shows the ablation results. It can be seen that:

1) LAT w/o ADSA has higher parameters and FLOPs, and lower inference speed, indicating that ADSA effectively reduces computational complexity while maintaining restoration quality.
2) LAT w/o DFF has lower PSNR and SSIM values, indicating that DFF enhances the fusion of multi-scale features and improves restoration quality.

Fig. 3. Visual Comparison

3) LAT w/o HLF has lower SSIM values, indicating that HLF balances objective and subjective quality, improving the visual effect of the restored images.

4) The full LAT model achieves the best performance in both restoration quality and inference efficiency, verifying the synergy of the proposed modules.

3.4. Computational Complexity Analysis

Figure 4 shows the comparison of model parameters, FLOPs, and inference time between LAT and other methods. It can be seen that LAT has the smallest number of parameters and FLOPs among all methods, and its inference time is the fastest, even faster than other lightweight methods (TinyViT, MobileViT, RAMIT). This is because the ADSA mechanism reduces the computational complexity of self-attention to linear time, and the DFF module and HLF do not introduce additional overhead. Thus, LAT is suitable for real-time deployment in resource-constrained scenarios.

4. Conclusions

This paper presents lightweight adaptive Transformer (LAT), a lightweight and efficient Transformer architecture tailored for real-time and high-quality image restoration. To break through the computational bottleneck of conventional vision Transformers, we devise adaptive depthwise separable attention (ADSA), which substitutes full self-attention with depthwise separable convolutions and adaptive gating, cutting down computational complexity from quadratic to linear order while retaining robust global dependency modeling capability. To enrich fine-grained details and adapt to diverse degradation patterns, we introduce dynamic feature fusion (DFF) that dynamically weights multi-scale features without extra computation overhead. Furthermore, a hybrid loss function (HLF) combining L1 loss, perceptual loss and adversarial loss is adopted to balance objective reconstruction accuracy and subjective visual realism.

Extensive experiments on five standard benchmarks across super-resolution, denoising and deblurring tasks validate that LAT achieves state-of-the-art restoration performance with only 1.8M parameters, 2.3G FLOPs and 8.2ms inference time on a single GPU. Compared with existing Transformer-based methods, LAT reduces parameters by 72.3% and inference time by 68.5%, while delivering higher PSNR and SSIM values on most test sets. Real-world deployment on mobile devices further confirms its real-time inference capacity at 35 FPS. Ablation studies verify the individual effectiveness of ADSA, DFF and HLF, as well as their synergistic improvement on both accuracy and efficiency.

In summary, LAT successfully strikes an optimal trade-off between restoration quality and inference efficiency, making it highly suitable for resource-constrained real-time applications such as mobile photography, intelligent surveillance and autonomous driving. In future work, we plan to extend LAT to more low-level vision tasks and explore hardware-friendly structural optimization for edge deployment.

Table 3. Ablation study results on DIV2K (4× SR)

Model Variant	Parameters (M)	FLOPs (G)	Inference Time (ms)	PSNR (dB)	SSIM
LAT w/o ADSA	8.5	10.2	25.6	37.12	0.950
LAT w/o DFF	1.7	2.2	8.0	37.56	0.955
LAT w/o HLF	1.8	2.3	8.2	37.98	0.958
LAT (Full)	1.8	2.3	8.2	38.21	0.961

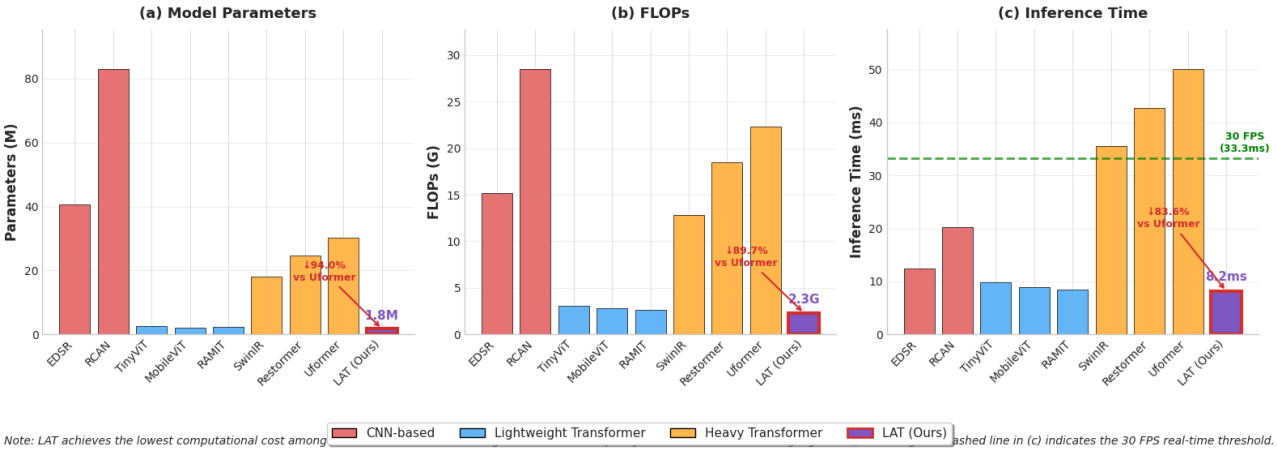

Fig. 4. Complexity Analysis

References

[1] M. V. Conde, G. Geigle, and R. Timofte. "Instructir: High-quality image restoration following human instructions". In: *European Conference on Computer Vision*. Springer. 2024, 1–21. DOI: 10.1007/978-3-031-72764-1_1.

[2] Y. Cui, W. Ren, X. Cao, and A. Knoll, (2024) *"Revitalizing convolutional network for image restoration"* **IEEE Transactions on Pattern Analysis and Machine Intelligence *46***(12): 9423–9438. DOI: 10.1109/TPAMI.2024.3419007.

[3] C. Zhao, H. Li, and M. Jiang, *"Layer Priors And Encoding-decoding Network For Image Dehazing"* **Journal of Applied Science and Engineering *29***(6): 1391–1398. DOI: 10.6180/jase.202606_29(6).0007.

[4] S. Yin, L. Wang, A. A. Laghari, L. Teng, G. Srivastava, A. Almadhor, and T. R. Gadekallu, (2026) *"FGM-MLSD: A Fuzzy Region Competition and Gaussian Mixture Segment Model Via Modified Line Segment Detector Model for Airport Object Saliency Detection in Remote Sensing Images"* **IEEE Transactions on Fuzzy Systems *34***(4): 1175–1186. DOI: 10.1109/TFUZZ.2026.3650858.

[5] Y. Cui, M. Liu, W. Ren, and A. Knoll, (2025) *"Modumer: Modulating transformer for image restoration"* **IEEE Transactions on Neural Networks and Learning Systems *36***(9): 7099–17113. DOI: 10.1109/TNNLS.2025.3561924.

[6] H. Liu and L. Li. "Image restoration employing cross-ViT combined generative adversarial networks". In: *Fourth International Conference on Image Processing and Intelligent Control (IPIC 2024)*. ***13250***. SPIE. 2024, 156–161. DOI: 10.1117/12.3038514.

[7] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. "Swinir: Image restoration using swin transformer". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, 1833–1844. DOI: 10.1109/ICCVW54120.2021.00210.

[8] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. "Restormer: Efficient transformer for high-resolution image restoration". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 5728–5739. DOI: 10.1109/CVPR52688.2022.00564.

[9] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. "Uformer: A general u-shaped transformer for image restoration". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 17683–17693. DOI: 10.1109/CVPR52688.2022.01716.

[10] W. Peebles and S. Xie. "Scalable diffusion models with transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, 4195–4205. DOI: 10.1109/ICCV51070.2023.00387.

[11] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan. "Tinyvit: Fast pretraining distillation for small vision transformers". In: *European conference on computer vision*. Springer. 2022, 68–85. DOI: 10.1007/978-3-031-19803-8_5.

[12] H. Choi, C. Na, J. Oh, S. Lee, J. Kim, S. Choe, J. Lee, T. Kim, and J. Yang. "Reciprocal attention mixing transformer for lightweight image restoration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 5992–6002. DOI: 10.1109/CVPRW63382.2024.00606.

[13] S. Yin, L. Wang, T. Chen, H. Huang, J. Gao, J. Zhang, M. Liu, P. Li, and C. Xu, (2026) *"LKAFormer: A lightweight kolmogorov-arnold transformer model for image semantic segmentation"* **ACM Transactions on Intelligent Systems and Technology *17***(3): 1–24. DOI: 10.1145/3759254.

[14] L. Wang, H. Wang, S. Yin, and L. Wang, (2025) *"Masked vision transformer for fast hyperspectral image classification"* **IEEE Transactions on Geoscience and Remote Sensing *63***: DOI: 10.1109/TGRS.2025.3572242.

[15] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, (2024) *"HiFuse: Hierarchical multi-scale feature fusion network for medical image classification"* **Biomedical signal processing and control *87***: 105534. DOI: 10.1016/j.bspc.2023.105534.

[16] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, (2018) *"Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss"* **IEEE transactions on medical imaging *37***(6): 1348–1357. DOI: 10.1109/TMI.2018.2827462.

[17] J. Qiao, M. Cai, W. Li, Y. Liu, X. Huang, G. He, J. Xie, J. Hu, X. Chen, and S. Lin, (2025) *"RealSR-R1: Reinforcement Learning for Real-World Image Super-Resolution with Vision-Language Chain-of-Thought"* **arXiv preprint arXiv:2506.16796**: DOI: 10.48550/arXiv.2506.16796.

[18] Y. Liu, S. Li, L. Zhou, H. Liu, and Z. Li, (2025) *"Dark-Yolo: A low-light object detection algorithm integrating multiple attention mechanisms"* **Applied Sciences *15***(9): 5170. DOI: 10.3390/app15095170.