Frequency-Aware Adaptive Fusion Gate for Single Image Super-Resolution

QiXin Liu and Ka-Cheng Choi *

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1446.v1

Keywords: super-resolution; dense-residual-connected transformer; discrete cosine transform; adaptive gating; frequency-aware learning

 doi:10.20944/preprints202604.1446.v1

Article

Frequency-Aware Adaptive Fusion Gate for Single Image Super-Resolution

QiXin Liu and Ka-Cheng Choi *

Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR., China

* Correspondence: rebeccachoi@mpu.edu.mo

Abstract

The Dense-Residual-Connected Transformer (DRCT) has established a new state-of-the-art in single image super-resolution by mitigating the information bottleneck in deep networks. However, its feature aggregation mechanism relies on a suboptimal Static Addition strategy, where residual features are scaled by a fixed, learnable scalar, regardless of the image content. This content-agnostic approach treats high-frequency textures and low-frequency noise indiscriminately, limiting the model's representational capability. To address this, we propose a Frequency-Aware Adaptive Fusion Gate (FAFG) to replace the static scaling. Unlike spatial-only gating mechanisms, FAFG integrates the Discrete Cosine Transform (DCT) to explicitly perceive the frequency distribution of feature maps. By decomposing features into frequency components, our gate acts as an intelligent valve, dynamically amplifying valid structural details while suppressing redundant background noise. Extensive experiments on standard benchmarks demonstrate that our proposed FAFG-integrated model consistently outperforms the static-scaling and other state-of-the-art methods. Specifically, our method achieves a significant PSNR improvement of 0.31dB on the texture-rich Urban100 dataset at $\times 4$ scale. Visual results further confirm that our frequency-aware gating mechanism effectively recovers more sharp edges and fine textures, providing a superior trade-off between reconstruction accuracy and model complexity.

Keywords: super-resolution; dense-residual-connected transformer; discrete cosine transform; adaptive gating; frequency-aware learning

MSC: 68U10; 94A08

1. Introduction

In recent years, deep learning technology has profoundly revolutionized the field of computer vision, enabling machines to perceive, process, and reconstruct complex visual data with unprecedented accuracy [1,2]. A prominent and highly challenging task within this broad domain is Single Image Super-Resolution (SISR), which aims to reconstruct a high-definition, high-resolution image from a single degraded, low-resolution observation [3,4]. Developing effective SISR models directly addresses essential needs across multiple dimensions, as high-quality image reconstruction is indispensable for numerous critical downstream applications. For instance, accurately enhancing the resolution of medical scans can significantly assist doctors in making precise, life-saving clinical diagnoses, while recovering fine details from degraded satellite imagery and surveillance videos is absolutely crucial for robust object identification and public security. Despite its profound practical value, from a strict mathematical and physical perspective, SISR remains an inherently ill-posed inverse problem [5]. This fundamental difficulty arises because a single low-resolution image can theoretically be produced by an infinite number of high-resolution images through various irreversible degradation processes, including downsampling, optical blurring, compression artifacts, and sensor noise corruption [6]. Consequently, forcing a neural network to invert this complex

 doi:10.20944/preprints202604.1446.v1

degradation process and hallucinate the completely missing high-frequency details requires exceptionally strong structural priors and expressive feature representation capabilities.

Beyond the sheer mathematical complexity, human visual perception adds another critical layer of challenge to the SISR task. The human visual system is exquisitely sensitive to high-frequency components, such as sharp geometric edges, intricate textures, and clear structural boundaries. Conversely, human vision is relatively tolerant of minor variations in low-frequency regions, such as smooth skies or flat background walls. Historically, early deep learning architectures for SISR heavily relied on standard pixel-wise objective functions, such as Mean Squared Error (MSE), to minimize the distance between the reconstructed high-resolution image and its ground truth. However, optimizing exclusively for MSE inherently favors the generation of overly smooth and perceptually blurry outputs, significantly failing to satisfy human aesthetic expectations [7]. Therefore, an ideal super-resolution model must not only minimize pixel-wise mathematical errors but also intelligently perceive the underlying content of the image, applying distinct, adaptive restoration strategies to different local regions to achieve optimal perceptual fidelity.

Dong et al. [8] pioneered the field by introducing the widely recognized Super-Resolution Convolutional Neural Network (SRCNN) model, establishing the initial foundation by learning a direct, end-to-end nonlinear mapping from low to high resolution. Subsequent classical architectures further pushed the quantitative performance boundaries. For instance, Lim et al. [9] proposed the Enhanced Deep Super-Resolution network (EDSR), elegantly utilizing deep residual learning by explicitly removing unnecessary batch normalization layers to preserve color distributions. As the field rapidly matured, researchers began exploring diverse generative paradigms to further enhance perceptual realism. Ledig et al. [10] pioneered the Super-Resolution Generative Adversarial Network (SRGAN), introducing adversarial losses to synthesize highly realistic textures that traditional pixel-loss methods completely failed to generate. Building upon this, Wang et al. [11] developed the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), successfully enhancing the adversarial framework with Residual-in-Residual Dense Blocks to achieve visually stunning structural reconstructions. More recently, Saharia et al. [12] introduced Image Super-Resolution via Iterative Refinement (SR3), innovatively adapting the revolutionary Denoising Diffusion Probabilistic Models to the super-resolution task. However, while these diverse generative approaches excel in perceptual realism, they frequently introduce severe hallucinatory artifacts and fundamentally demand exorbitant computational resources and extended inference times, critically limiting their deployment in efficiency-sensitive applications.

Consequently, preserving the high fidelity of structural reconstruction while strictly optimizing computational efficiency remains the most dominant and practical research trajectory. Within this trajectory, dynamically weighting and intelligently fusing extracted features using attention mechanisms has proven to be a highly effective strategy. Zhang et al. [13] designed the Residual Channel Attention Networks (RCAN), which successfully integrated channel-wise attention mechanisms to selectively focus on informative features using spatial average pooling. Recognizing the need for capturing higher-order feature correlations, Dai et al. [14] proposed the Second-order Attention Network (SAN), utilizing adaptive feature pooling to powerfully enhance the non-linear learning capabilities of the model. Furthermore, Niu et al. [15] introduced the Holistic Attention Network (HAN), modeling the comprehensive dependencies across various network depths to optimize global feature aggregation. To overcome the intrinsic local receptive fields of these standard convolutional operations, recent research witnessed a massive paradigm shift toward Vision Transformers (ViTs) [16]. Chen et al. [17] proposed the Image Processing Transformer (IPT), heavily utilizing self-attention to capture long-range global interactions across discrete image patches. Building upon this, Liang et al. [18] developed SwinIR, which elegantly adapted the shifted-window attention mechanism to achieve superior reconstruction quality with a highly manageable computational complexity. Advancing this structural design, Chen et al. [19] introduced the Hybrid Attention Transformer (HAT), explicitly integrating overlapping cross-attention to activate a significantly larger range of pixels, thereby establishing a solid new standard in the image restoration

 doi:10.20944/preprints202604.1446.v1

domain. While these attention-driven and Transformer-based architectures have revolutionized long-range dependency modeling and significantly expanded effective receptive fields, their final residual aggregation strategies remain predominantly static. Specifically, by relying on standard element-wise addition to merge activated features with identity mappings, these advanced models miss the critical opportunity to dynamically modulate the fusion process based on distinct frequency components, leaving room for further textural refinement.

Beyond the evolution of generative models and attention-based transformers, researchers have also extensively explored lightweight architectural designs to explicitly alleviate the computational burden associated with high-resolution image reconstruction. Furthermore, carefully considering the environmental impact, the exponential and unchecked growth in the size of deep learning models has raised valid, urgent concerns regarding their massive computational costs and subsequent carbon footprints [29]. Therefore, there is a strong societal mandate for Green AI solutions that improve model performance without simultaneously introducing heavy parameter overheads. Driven by this efficiency paradigm, Dong et al. [20] introduced the Fast Super-Resolution Convolutional Neural Network (FSRCNN), which dramatically accelerated the reconstruction process by utilizing a compact hourglass structure and performing upsampling at the very end of the network. Advancing this trajectory, Ahn et al. [21] developed the Cascading Residual Network (CARN), elegantly employing a cascading mechanism at both local and global levels to achieve highly efficient feature representations without sacrificing structural fidelity. Alternatively, explicitly incorporating structural priors has garnered significant attention as a mechanism to constrain the ill-posed nature of super-resolution. Ma et al. [22] proposed the Structure-Preserving Super-Resolution (SPSR) framework, actively leveraging gradient maps to guide the neural network in accurately recovering sharp geometric edges and preventing visual distortion. Although these efficiency-driven and structure-aware frameworks successfully mitigate computational overhead and sharpen geometric boundaries, they typically rely on rigid, spatially invariant feature aggregation. Consequently, they lack the adaptive capacity to dynamically differentiate between informative high-frequency textures and redundant low-frequency backgrounds, occasionally leading to over-smoothed details or amplified noise. Ultimately, to deliver an environmentally sustainable and exceptionally effective algorithmic solution, it is imperative to definitively transition from these rigid, content-agnostic paradigms to a highly dynamic, intelligent, frequency-aware adaptive mechanism.

Furthermore, the effective utilization of multi-scale hierarchical features represents another crucial investigative direction within the super-resolution community. Lai et al. [23] designed the Laplacian Pyramid Super-Resolution Network (LapSRN), which progressively predicts high-frequency residuals across multiple cascading resolution scales, effectively managing the restoration of diverse structural frequencies. Similarly, Li et al. [24] proposed the Multi-scale Residual Network (MSRN), dynamically combining local and global hierarchical features to fully exploit the rich contextual information embedded within the degraded image. While such hierarchical designs successfully capture versatile receptive fields to adapt to objects of varying sizes, most of these architectures treat the extracted representations through uniform concatenation or summation. This fails to account for the varying importance of different frequency components across distinct spatial regions, often leading to a suboptimal allocation of computational resources on redundant background information. However, while these diverse methodologies have undeniably propelled the field forward—whether through generative perceptual realism, advanced long-range attention mechanisms, efficient lightweight architectural designs and hierarchical multi-scale feature fusion—our comprehensive investigation reveals a critical consensus. The fundamental mechanism of aggregating residual features within the deepest functional blocks remains rigidly static or frequency-agnostic across almost all these paradigms.

A prominent example of this neglected feature aggregation bottleneck becomes acutely apparent as modern Transformer-based models scale deeper to pursue even better performance. For instance, Hsu et al. [25] recently proposed the Dense-Residual-Connected Transformer (DRCT) to actively stabilize the hierarchical information flow across deep layers. Given its exceptional efficacy in

 doi:10.20944/preprints202604.1446.v1

modeling complex spatial dependencies and its state-of-the-art reconstruction performance among recent Transformer paradigms, DRCT serves as an ideal and formidable foundation for our research. We chose DRCT as our main model to illustrate that even the most optimized Transformer model can benefit from adopting the dynamic aggregation approach we proposed instead of its traditional static fusion strategy.

To definitively overcome this fundamental limitation, exploring frequency-aware modulation mechanisms has rapidly emerged as a highly promising avenue. By mathematically projecting spatial features into the frequency domain using discrete spectral transformations, such as the Discrete Cosine Transform (DCT), a network can selectively amplify high-frequency structural details while simultaneously suppressing redundant low-frequency noise. Motivated by this, we propose the Frequency-Aware Adaptive Fusion Gate (FAFG), a highly intelligent and computationally lightweight mechanism designed to explicitly replace conventional rigid static scaling. By seamlessly integrating this spectral-based module into the powerful DRCT architecture, we have expanded the boundaries of dynamic feature aggregation in deep super-resolution architectures. Ultimately, our proposed framework powerfully discriminates between vital textures and background noise, achieving superior reconstruction performance without incurring prohibitive parameter overheads.

2. Literature Review

As highlighted in the introduction, the trajectory of SISR has experienced a profound paradigm shift from Convolutional Neural Networks (CNNs) to ViTs. While attention-driven architectures like SwinIR[18] have successfully leveraged shifted-window mechanisms to capture long-range structural dependencies, they frequently encounter severe optimization hurdles when scaled to greater depths. Specifically, Hsu et al. [25] identified a critical architectural flaw inherent in modern deep networks: the severe information bottleneck and progressive feature degradation. As transformer-based models continuously stack self-attention layers to expand their receptive fields and achieve superior restoration performance, the crucial high-frequency details extracted at the early stages are frequently diluted, over-smoothed, or entirely lost during forward propagation. To actively stabilize the hierarchical information flow and explicitly prevent this decay, the DRCT [25] was proposed as a robust topological solution. It maximizes representation capacity and sustains signal variance without demanding the exorbitant computational complexity of heavily modified attention windows.

As explicitly illustrated in Figure 1, the macroscopic architecture of the DRCT [25] framework strictly follows a highly effective three-stage pipeline to ensure structural fidelity. The network commences with a Shallow Extraction module, utilizing a standard 3×3 Convolution to project the Input low-resolution (LR) image into a higher-dimensional latent space. This initial projection is indispensable, as it provides a stable, translation-invariant foundation and preserves essential low-frequency priors. These foundational features are subsequently passed into the core computational engine of the network: the Residual Deep Feature Extraction (RDFE) module. Ultimately, the network concludes with the High-Quality Image Reconstruction stage. This terminal module safely merges the deeply processed representations with the original shallow features—bypassed via a long-range Global Identity connection—and employs sub-pixel convolutions to upsample the aggregated tensors, producing the final Output super-resolved (SR) image.

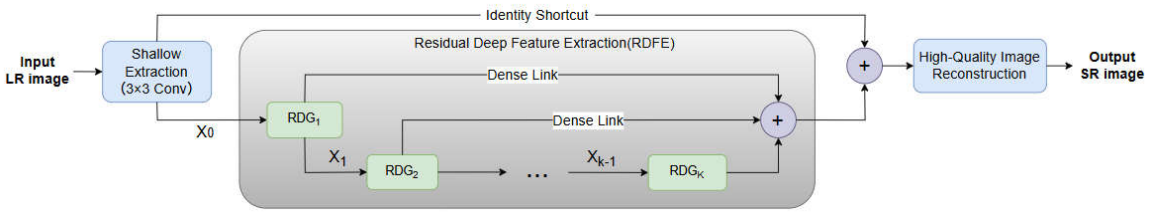

Figure 1. The overall architecture of the DRCT model.

 doi:10.20944/preprints202604.1446.v1

The primary architectural superiority of DRCT [25] over conventional sequential models resides internally within its RDFE module, which is constructed by stacking multiple Residual Dense Groups ($RDG_1, RDG_2, \ldots, RDG_K$). To definitively counteract the catastrophic decay of activation variance across these groups, DRCT [25] abandons simple sequential chaining. Instead, it utilizes continuous Dense Links to forge a multi-lane computational highway. This dense macro-topology creates expansive, uninterrupted pathways that allow high-frequency texture priors from early functional groups to safely bypass intermediate transformations.

At the micro-level, as detailed in the internal schematic of Figure 2, the processing dynamics within each Residual Dense Group (RDG) are primarily driven by cascaded Swin Dense Residual Connected Blocks (SDRCB 1 and SDRCB 2). These blocks are fundamentally constructed utilizing Swin Transformer Layers, which strictly alternate between Window-based Multi-head Self-Attention (W-MSA) and Shifted-Window Multi-head Self-Attention (SW-MSA). This alternating attention mechanism empowers the network to efficiently capture long-range structural dependencies and cross-window connections while maintaining a strictly linear computational complexity. Within the RDG topology, the foundational low-frequency spatial information is perfectly preserved and propagated forward through an uninterrupted Identity Shortcut. Simultaneously, the cascaded SDRCB actively process the incoming features to extract complex high-frequency textural variations, ultimately generating the unmodulated spatial tensor denoted as Residual R.

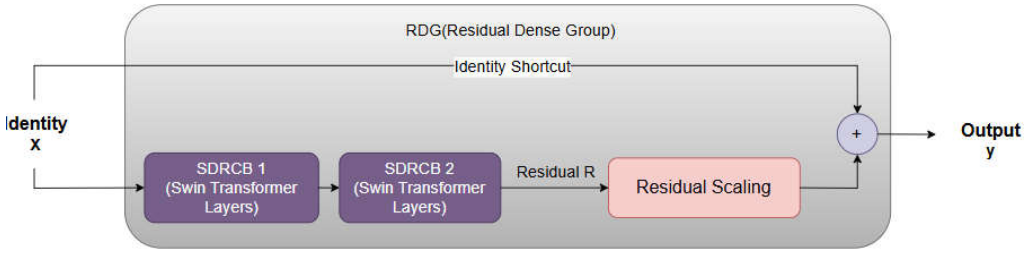

Figure 2. The internal architecture of the Residual Dense Group.

However, despite its exceptional topological advantages and stabilized information flow, the DRCT fundamentally relies on a suboptimal fusion strategy to aggregate these parallel branches. As explicitly highlighted by the warning indicator in Figure 2, the integration of the extracted Residual R with the Identity Shortcut is governed entirely by a rigid Residual Scaling ($\times\alpha$). In this formulation, the residual tensor is multiplied by a globally shared, learnable scalar parameter before being added to the identity branch. While this static scaling operation prevents gradient explosion and physically stabilizes the optimization process of exceptionally deep networks, it acts as a severe Content-Agnostic Bottleneck.

Because this scalar applies an identical, fixed numerical weight across all spatial coordinates and all feature channels indiscriminately, it strictly forces the network to treat completely flat backgrounds and dense, high-frequency textures identically. In texture-rich regions, insufficient residual integration leads to over-smoothed boundaries. in uniform background regions, indiscriminate integration frequently amplifies noise and causes color shifting. To definitively eliminate this specific aggregation flaw while preserving the powerful DRCT backbone, our research proposes to replace this rigid static scalar with a dynamic, frequency-aware adaptive gating mechanism. By doing so, we explicitly empower the network to modulate the residual fusion based strictly on localized spectral energy, effectively unlocking the ultimate reconstructive potential of dense-residual architectures.

2.1. Channel Attention

Recognizing the severe representational limitations imposed by static, content-agnostic aggregation, the logical progression in architectural design is to explore dynamic routing mechanisms. The most intuitive alternative to static scaling is channel-wise feature modulation. In

 doi:10.20944/preprints202604.1446.v1

the broader computer vision domain, architectures like SENet **Error! Reference source not found.** and RCAN **Error! Reference source not found.** have achieved profound success by explicitly modeling interdependencies between convolutional channels, allowing the network to actively recalibrate feature maps based on global context.

These conventional channel attention mechanisms universally rely on Global Average Pooling (GAP) to compress spatial statistics before generating modulation weights. While GAP provides a degree of dynamic adaptability, its direct mathematical implementation in high-fidelity super-resolution introduces a critical analytical flaw. From the rigorous perspective of digital signal processing, computing the arithmetic mean of a two-dimensional signal via GAP is formally equivalent to applying an extreme low-pass filter—specifically, extracting the zero-frequency, or Direct Current, component. Consequently, the resulting pooled descriptor merely reflects the overall brightness or the average color intensity of that particular channel, retaining absolutely zero deterministic information regarding the spatial distribution or textural complexity of the original region. Faced with this Spatial Ambiguity, the network fundamentally fails to distinguish between a perfectly flat background and a highly dense geometric pattern if their average intensities are identical.

To mitigate the excessive information loss caused by GAP, advanced channel attention variants frequently employ Dual-Pooling strategies. By concatenating GAP with Global Max Pooling (GMP), the network theoretically captures the most extreme activation values, which often correspond to sharp edges and salient boundaries. However, as empirically demonstrated in our ablation studies, this heuristic statistical combination still fundamentally remains trapped within the spatial domain. While GMP detects the presence of strong gradients, it completely fails to encode the precise frequency orientations, geometric periodicities, and directional boundaries essential for reconstructing sharp structural details.

2.2. Spatial Attention

To counteract the spatial ambiguity inherent in channel-wise pooling, another extensively explored trajectory is Spatial Attention. Inspired by the Convolutional Block Attention Module (CBAM) [31] and recent advanced gating mechanisms—such as the Nonlinear Activation Free Network (NAFNet) [35] and the Gated Feed-Forward Network in Restormer [36]—spatial gating utilizes localized convolutional filters to explicitly perceive spatial context. Instead of compressing the spatial dimensions into a scalar, spatial attention generates a dense, pixel-wise attention map, theoretically instructing the network on exactly where to emphasize high-frequency residual signals.

We initially hypothesized that incorporating localized spatial convolution modules could explicitly guide the residual fusion process to differentiate between complex textures and smooth backgrounds. However, applying spatial attention within a dense Transformer-based pipeline presents severe structural redundancies. The preceding Swin Transformer Layers (STLs) within the network have already executed highly sophisticated, dense spatial token mixing via window-based self-attention mechanisms. Appending additional spatial convolutions for the sole purpose of gating introduces bloated learnable parameters and significant computational latency.

More detrimentally, forcing the network to learn these additional, redundant spatial convolutions frequently disrupts the carefully calibrated feature representations produced by the self-attention heads. This structural redundancy leads to severe optimization conflicts during the early stages of training. The network struggles to balance the spatial routing computed by the STLs and the localized filtering computed by the spatial attention gate, ultimately resulting in sub-optimal super-resolution performance compared to purely spectral approaches.

The comprehensive analysis of both channel-wise and spatial-wise methodologies highlights a critical academic consensus: while dynamic modulation is absolutely essential, both heuristic channel pooling, which suffers from spatial ambiguity, and redundant spatial convolutions, which cause optimization conflicts, are fundamentally misaligned with the specific demands of residual fusion in deep SISR networks. It becomes evident that any dynamic gating mechanism strictly confined to the

 doi:10.20944/preprints202604.1446.v1

spatial domain will inevitably encounter a representational bottleneck. To decisively resolve these limitations, extracting a complete spectrum of orthogonal frequency components emerges as the sole rigorous mathematical solution. This profound theoretical realization directly motivates our architectural shift towards a DCT based frequency-aware gating mechanism.

3. Materials and Methods

3.1. FAFG Overall Architecture

Hsu et al. [25] designed the DRCT framework with a highly optimized macroscopic information flow. To preserve this powerful global topology—including the shallow feature extraction, the macroscopic dense connections between consecutive groups, and the final image reconstruction modules—our architectural modifications are strictly confined to the internal feature aggregation phase. For a comprehensive view of this macroscopic routing and the unchanged external topological structure, please refer to Figure **1** in Chapter 2. By explicitly locking the outer architecture, we ensure that our proposed FAFG acts as an internal, plug-and-play enhancement designed specifically to replace the suboptimal static scaling operation residing deep within each RDG.

The precise internal mechanics of this upgraded module are explicitly illustrated in Figure 3. Within the bounding box of a single RDG, the data flow initiates from a singular input tensor denoted as Identity X. From this origin, the architecture bifurcates into two distinct pathways. The upper pathway forms the Identity Shortcut, a continuous, unhindered gradient highway. The primary purpose of this shortcut is to safely propagate the foundational, low-frequency representations directly to the end of the block, thereby preserving the basic structural identity of the image and actively preventing the vanishing gradient problem during backpropagation.

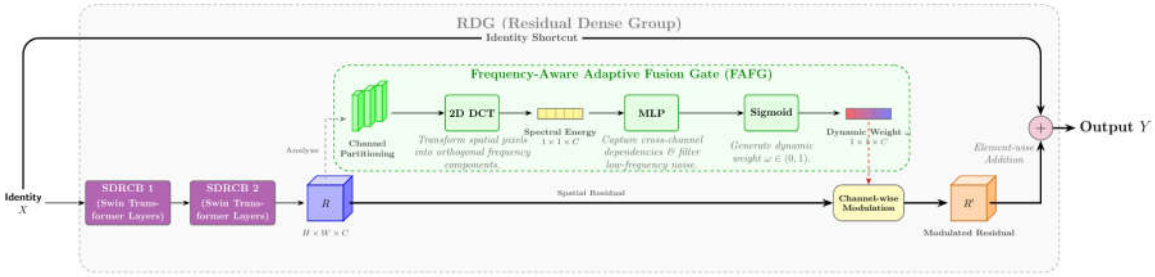

Figure 3. The detailed architectural pipeline of the proposed FAFG module.

Simultaneously, the primary structural information flows into the lower residual extraction branch. The Identity X tensor is sequentially processed by two cascaded functional blocks: SDRCB 1 and SDRCB 2. These highly expressive transformer layers execute dense spatial token mixing to capture long-range visual dependencies and extract complex high-frequency textural variations, ultimately generating an unmodulated 3D spatial tensor identified as Residual R, which possesses the dimensions of $H \times W \times C$.

In conventional architectures, this Residual R would be indiscriminately multiplied by a fixed static scalar. However, in our proposed design, the residual tensor undergoes a sophisticated multi-spectral assessment before any fusion occurs. As depicted by the dashed arrow labeled Analyze in Figure **3**, a copy of Residual R is intercepted and routed directly into the green bounding box of our core innovation: the FAFG.

The internal pipeline of the FAFG operates through a continuous sequence of meticulously designed transformations. Initially, the incoming 3D spatial tensor undergoes Channel Partitioning, where it is sliced into multiple discrete sub-groups. This physical division is crucial, as it forces the network to concurrently analyze diverse frequency bands across different channel dimensions, preventing high-frequency edge signals from being overwhelmed by dominant low-frequency background signals. Following this, the partitioned features are passed through a 2D DCT. As

 doi:10.20944/preprints202604.1446.v1

denoted by the architectural caption, this operation transforms spatial pixels into orthogonal frequency components, successfully compressing the 3D spatial dimensions into a 1D vector representing the absolute Spectral Energy ($1 \times 1 \times C$). Unlike Global Average Pooling (GAP), which suffers from spatial ambiguity, the 2D DCT precisely quantifies the magnitude of energy at different structural scales, providing the network with a deterministic metric to differentiate between sharp geometric edges and flat backgrounds. This extracted spectral energy vector is subsequently fed into a lightweight multi-layer perceptron (MLP) bottleneck. By utilizing an intentional dimension reduction and expansion strategy, this module is engineered to capture non-linear cross-channel dependencies, effectively filtering out uncorrelated spectral noise while retaining vital geometric cues. Finally, the refined signal passes through a Sigmoid activation function. The explicit role of the Sigmoid is to mathematically bound the output modulation vector strictly between 0 and 1, ensuring numerical stability. This process transforms raw spectral features into the ultimate control signal of the module: the Dynamic Weight ω, a 1D gradient-colored vector of dimensions $1 \times 1 \times C$ that represents the precise structural demands of each channel.

Following the generation of this intelligent spectral routing signal, the network proceeds to the terminal fusion stage. The original 3D Spatial Residual R and the newly computed 1D Dynamic Weight ω converge at the yellow block labeled Channel-wise Modulation. Here, an explicit multiplication is executed, utilizing the 1D frequency weights to dynamically scale the 3D spatial residual along the channel dimension. This modulation acts as an intelligent valve, dynamically amplifying channels that contain valid structural details while severely suppressing redundant artifacts, which physically produces the orange 3D tensor, the Modulated Residual R'. In the final step of the RDG, this refined Modulated Residual R' merges with the pristine features traveling along the upper Identity Shortcut. This integration occurs at the purple Element-wise Addition (+) node, resulting in the final Output Y tensor. This terminal addition safely integrates the highly refined textural details back into the main architectural flow, ensuring absolute consistency with the macroscopic topology and guaranteeing stable convergence. By executing this exact architectural sequence, the FAFG seamlessly upgrades the RDG to a dynamic, frequency-aware state.

3.2. Static Scaling and Its Limitation

However, scaling up these architectures introduces profound optimization challenges. The unregulated sequential addition of residual branches inevitably leads to an exponential accumulation of activation variances across the network depth. This unchecked variance growth frequently triggers numerical instability and gradient explosion during the continuous backpropagation phase. To mitigate these optimization hurdles and ensure a stable training trajectory for exceedingly deep transformer networks, the DTCT model incorporates a static residual scaling strategy. Following the architectural design pioneered by Liang et al. **Error! Reference source not found.**, who utilized the Swin Transformer Layer (STL) [38] for super-resolution, let the input to a given STL be denoted by an identity tensor X. The aforementioned layer processes this input to produce a transformed residual tensor R. The static scaling operation dictates that this residual tensor is multiplied by a single, globally shared scalar parameter before being aggregated with the identity branch. For mathematical rigor, all interacting tensors X, R and the final output Y share identical spatial dimensions $H \times W$ and channel configuration C. We establish consistent subscript definitions for the spatial coordinates i and j, alongside the channel index c, which are strictly retained in all subsequent fusion formulations. The output tensor Y resulting from this residual block is formally defined by the following equation:

$$Y_{i,j,c} = X_{i,j,c} + \alpha \cdot R_{i,j,c} \tag{1}$$

In this formulation, the subscripts i and j represent the two-dimensional spatial coordinates corresponding to the vertical height and the horizontal width of the feature map, respectively. The subscript c indicates a particular channel index within the multi-dimensional tensor volume. The parameter α represents the globally shared, learnable scalar weight. Notably, this scalar α applies

 doi:10.20944/preprints202604.1446.v1

an identical scaling factor to all spatial locations and all feature channels of the residual tensor R, making the fusion strategy completely content-agnostic. The fixed modulation of R is the only variable in this residual fusion architecture, which is the core bottleneck we aim to address with our adaptive gating mechanism. While this elegant simplicity guarantees computational efficiency and successfully prevents gradient anomalies, it simultaneously introduces a severe representational bottleneck. The core structural limitation stems directly from the fact that the scalar multiplier remains entirely agnostic to the underlying visual content. As unequivocally evidenced by the mathematical formulation provided above, the identical numerical weight is uniformly enforced across all spatial coordinates and every distinct feature channel. This rigid mechanism dictates that the network must apply an indiscriminate degree of feature fusion regardless of whether the target pixel belongs to a completely flat background area or an intricately detailed textural region. In domains demanding extreme precision such as high-resolution image reconstruction, this inflexible paradigm inevitably leads to severely compromised visual fidelity. Homogenous regions often suffer from amplified noise because the unvarying scalar indiscriminately incorporates redundant residual perturbations. Conversely, complex geometries and sharp edges exhibit noticeable over-smoothing artifacts due to the insufficient integration of high-frequency restorative signals.

Recognizing the severe limitations imposed by the content-agnostic static multiplier, the immediate logical progression is to engineer a dynamic gating mechanism. The primary objective of such a module is to compute a customized weighting factor for each individual channel, thereby allowing the network to selectively emphasize informative features while robustly suppressing redundant noise. Hu et al. [30] and Zhang et al. [13] pioneered channel attention mechanisms in contemporary deep learning paradigms, where spatial statistics are predominantly compressed using GAP, a fundamental operation introduced by Lin et al. [39]. The application of this pooling strategy involves condensing the entire spatial plane of a given feature map into a solitary representative scalar per channel. Assuming the spatial height and the spatial width of the incoming residual tensor R are denoted by H and W respectively, the pooled spatial descriptor z for a target channel c is computed by calculating the arithmetic mean of all pixel intensities distributed across that specific dimension. The representation of this compression step is provided by the following expression:

$$z_c = \frac{1}{H \cdot W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} R_{i,j,c} \tag{2}$$

Upon deriving this compact descriptor, conventional gating mechanisms subsequently employ a lightweight perceptron to map the pooled statistic into a gating weight constrained between zero and one. However, its deployment in low-level image restoration tasks introduces a fatal analytical flaw widely recognized as spatial ambiguity. From the rigorous perspective of digital signal processing, computing the arithmetic mean of a two-dimensional signal is not merely a heuristic aggregation. It is formally equivalent to applying an extreme low-pass filter. A fundamental analysis reveals that this spatial averaging operation is strictly proportional to computing the zero-frequency component of a spectral decomposition. Consequently, the resulting descriptor z_c merely reflects the overall brightness or the average color intensity of that particular channel. It retains absolutely zero deterministic information regarding the spatial distribution, the structural density, or the textural complexity of the original region. Consider two distinctly different regions within an image: one containing a perfectly uniform background and another featuring a highly dense geometric pattern. If the average pixel intensity of the flat background happens to be identical to the average intensity of the complex pattern, the pooling operation will yield the exact same numerical descriptor for both regions. Faced with identical input statistics, the perceptron is forcibly constrained to generate an identical gating weight. Therefore, relying exclusively on spatial averaging completely incapacitates the network from distinguishing between low-frequency noise and high-frequency details.

To further elucidate this catastrophic loss of structural information, one must consider the fundamental statistical moments of a spatial distribution. While the arithmetic mean successfully captures the primary central tendency, it fundamentally fails to encode the dispersion or the spatial

 doi:10.20944/preprints202604.1446.v1

variance of the pixel intensities. Let the spatial variance of the given channel c be denoted by the mathematical symbol v_c. This second-order statistic is formally defined by the double summation of the squared deviations from the previously computed mean z_c, mathematically expressed as follows:

$$v_c = \frac{1}{H \cdot W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left(R_{i,j,c} - z_c\right)^2 \tag{3}$$

In highly textured regions encompassing sharp edges and complex geometries, this variance metric exhibits a substantially large numerical magnitude. Conversely, in uniform and flat background areas, this calculated variance approaches zero. While appending this continuous variance descriptor to the existing pooling mechanism could theoretically alleviate the spatial ambiguity to a certain degree, it merely provides a rudimentary estimation of global spatial contrast. It still comprehensively fails to capture the precise frequency orientations, the geometric periodicities, and the directional boundaries that are essential for high-fidelity image restoration. Extracting a complete spectrum of orthogonal frequency components emerges as the sole rigorous solution to achieve genuine content-aware modulation.

To decisively resolve the aforementioned spatial ambiguity without incurring the prohibitively high computational burdens typically associated with complex frequency domain transformations, our FAFG methodology completely discards the simplistic spatial averaging approach in favor of a robust spectral transformation. Qin et al. [26] profoundly explored the foundational principles of spectral analysis in channel representations. Inspired by their mathematical insights, the core of our innovation lies in the integration of the two-dimensional DCT, fundamentally formulated by Ahmed et al. [40]. This continuous transformation enables the network to explicitly perceive the intricate frequency distribution of the incoming residual features, acting as a highly discerning spectral analyzer. The operation commences immediately after the shifted-window attention mechanism [38] concludes its token mixing process. At this juncture, instead of indiscriminately passing the tensor forward, the module intercepts the residual map to commence the multi-spectral decomposition.

3.3. FAFG Implementation

A critical architectural design within our proposed module is the strategic implementation of a multi-spectral routing mechanism. Instead of applying a single mathematical transformation to the entire tensor volume indiscriminately, the architecture intelligently divides the channels into multiple contiguous groups. This group-wise processing strategy is essential because it allows the network to concurrently analyze multiple distinct frequency bands, capturing a holistic spectrum of structural information ranging from low-frequency structural contours to extremely high-frequency sharp edges. Assume the total number of channels within the residual tensor is denoted by the integer C. We systematically partition these continuous channels into a specific number of independent groups, denoted by the integer N. Consequently, the channel indices belonging to a specific target group g are rigorously confined within a mathematically defined interval. The precise formulation for this channel partitioning strategy is expressed as follows:

$$c_g \in \left[\frac{g \cdot C}{N}, \frac{(g+1) \cdot C}{N} - 1\right] \tag{4}$$

In this interval definition, the variable g serves as the group index iterating from zero up to one less than the total number of groups N. This equitable division ensures that each independent group contains exactly an identical number of channels. By isolating the channels into these distinct groups, we can assign an entirely different spectral analysis task to each specific partition. The discrete cosine transform operates entirely within the real number domain. This characteristic ensures seamless and highly efficient integration with standard neural network feature mapping processes. To implement this spectral transformation within our grouped channel partitions, we define a comprehensive set of fundamental two-dimensional basis functions. The formulation for the specific basis function

 doi:10.20944/preprints202604.1446.v1

corresponding to the vertical frequency index u and the horizontal frequency index v evaluated at the spatial coordinates i and j is expressed as follows:

$$B_{u,v,i,j} = \cos\left(\frac{\pi \cdot (2i+1) \cdot u}{2H}\right) \cdot \cos\left(\frac{\pi \cdot (2j+1) \cdot v}{2W}\right) \tag{5}$$

In this rigorous mathematical definition, the variables H and W consistently represent the absolute spatial height and width of the localized feature window. The variables $i \in [0, H-1]$and$j \in [0, W-1]$ iterate through the spatial height and width coordinates, respectively. The integer parameters u and v fundamentally determine the respective frequencies of the cosine waves along the vertical and horizontal axes. A higher value of the vertical index u signifies rapid vertical oscillations along the i-axis, which effectively captures horizontal edges within the image geometry. Conversely, a higher value of the horizontal index v signifies rapid horizontal oscillations along the j-axis, which is instrumental in detecting vertical structural boundaries.

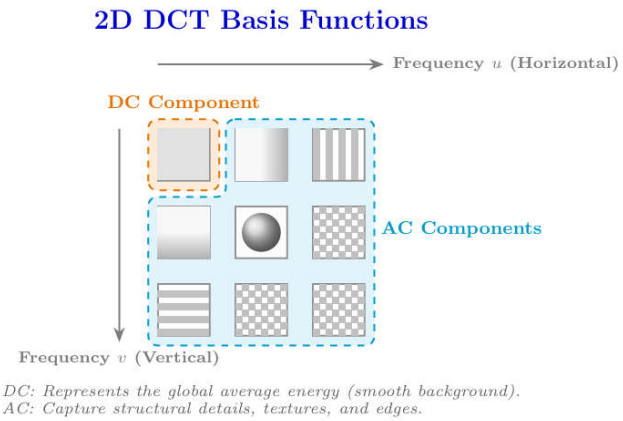

Figure 4. 2D DCT basis functions.

To explicitly visualize this spectral decomposition, Figure 4 illustrates the theoretical basis functions of the 2D DCT utilized within our module. As depicted in the basis function grid, the 2D DCT fundamentally decomposes signals using a dictionary of orthogonal cosine patterns. As explicitly highlighted by the annotated regions, these patterns are mathematically categorized into two distinct groups: the DC Component (orange region), which captures uniform background illumination and global average energy; and the AC Components (blue region), which encode sharp geometric grids, textures, and directional edges. Following the channel partitioning established earlier, we assign the lowest possible frequency components to the first channel group, empowering it to extract the fundamental background illumination and smooth contextual shapes. The subsequent channel groups are respectively assigned distinct high-frequency coordinate pairs. This multi-spectral assignment establishes a robust and comprehensive dictionary of structural patterns.

Building upon these established multi-frequency basis functions, the core aggregation operation involves projecting the incoming spatial tensors onto the specified frequency domain. Instead of calculating a blind spatial average, our module computes the precise mathematical projection of the residual feature map against the selectively assigned distinct spectral patterns. This continuous inner-product operation yields a highly informative descriptor that explicitly quantifies the magnitude of energy present at different structural scales. The scalar spectral descriptor f_c for a given channel c, which is assigned specific basis function indices u_c and v_c, is obtained by computing the spatial inner product over the entire feature window:

$$f_c = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} R_{i,j,c} \cdot B_{u_c,v_c,i,j} \tag{6}$$

 doi:10.20944/preprints202604.1446.v1

In this specific projection equation, the residual pixel values R located at coordinates i and j for a particular channel c are directly multiplied by their corresponding mathematical values derived from the predefined basis function B. This double summation operation cleanly compresses the 2D spatial features into a 1D Spectral Energy vector. Once this robust descriptor is successfully extracted across all channels according to their assigned groups, the resulting values inherently form a highly distinctive one-dimensional column vector f representing the overall multi-spectral composition of the entire image patch.

To accurately capture the non-linear cross-channel interactions and map these complex spectral signatures into decisive attention weights, the vector f is subsequently processed by a specialized non-linear excitation network. This bottleneck sub-network acts as the intelligent central controller, actively interpreting the spectral composition. The architecture specifically comprises two sequential linear transformations separated by an activation layer, strategically designed to capture intricate dependencies while maintaining an exceptionally low computational footprint. The design inherently relies on a strict channel condensation mechanism to force the network to distill the most critical representational features. Assume the integer r denotes the predefined reduction ratio. The entire intelligent weight generation sequence is defined in a vectorized format as follows:

$$\omega = \sigma\left(W_2 \cdot \delta(W_1 \cdot f)\right) \tag{7}$$

In this rigorous formulation, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ represents the learnable weight matrix of the initial down-projection stage. This specific mapping intentionally compresses the incoming one-dimensional spectral vector f from its original capacity C down to a significantly smaller scale. This forced compression creates an artificial information bottleneck, compelling the transformation to actively learn the global correlations and cross-channel dependencies among the varied frequency bands, effectively filtering out any uncorrelated spectral noise. Following this crucial purification step, represents the weight matrix of the subsequent up-projection stage, restoring the refined tensor back to the original full channel dimension C. The intermediate activation is governed by a Rectified Linear Unit (ReLU), denoted by the symbol δ, which introduces essential non-linearity into the feature mapping process. The final dynamic gating weight vector $\omega \in \mathbb{R}^C$ is generated through a terminal Sigmoid function, represented by the mathematical symbol σ. The deployment of the Sigmoid activation is a necessary architectural choice [41] because its smooth non-linear saturation characteristic mathematically bounds the output modulation vector strictly between zero and one.

This meticulously computed weight vector ω continuously encapsulates the precise structural demands of the feature map. Let ω_c denote the specific scalar weight extracted from the vector ω corresponding to the c-th channel. The final dynamic aggregation process utilizes this intelligently generated scalar to selectively amplify or suppress the corresponding residual features before they are permanently fused with the continuous identity tensor. We retain the fundamental residual fusion architecture of the DRCT (Equation (1)) in our formulation, with the only modification being the replacement of the globally shared static scalar α with a channel-wise adaptive weight ω_c generated by the FAFG module. This gradual modification ensures consistency with the architecture, while also enabling frequency-aware dynamic modulation of the remaining feature R. The final dynamically modulated output tensor Y is robustly formulated through an element-wise Hadamard product, followed immediately by the requisite identity addition operation:

$$Y_{i,j,c} = X_{i,j,c} + \omega_c \cdot R_{i,j,c} \tag{8}$$

In this advanced formulation, unlike the globally fixed scalar α utilized in Equation (1), represents a channel-specific adaptive weight derived directly from the spectral energy distribution of the residual tensor. This mechanism ensures that each distinct channel c is intelligently assigned an independent, content-aware modulation factor. The multiplication operation $\omega_c \cdot R_{i,j,c}$ applies precise differential scaling across the various channels of the residual feature, dynamically amplifying valid structural details while suppressing redundant noise. Concurrently, the core

 doi:10.20944/preprints202604.1446.v1

identity mapping, which executes the addition with $X_{i,j,c}$, remains completely consistent with the DRCT topology, thereby safely preserving the macroscopic gradient flow of the network.

3.4. Efficiency Superiority over Conventional Gating

To comprehensively appreciate the elegance of this spectral projection design, a rigorous comparative analysis against conventional spatial attention mechanisms is explicitly required. Traditional dynamic gating modules frequently employ localized spatial filtering operations to perceive spatial context before generating the attention weights. Assuming the square dimension of such a localized kernel footprint is denoted by the integer k, applying this standard transformation across the entire tensor volume demands a computational complexity that scales quadratically with the kernel dimension. The computational cost for a standard feature mapping process is proportional to the product of k squared, the spatial height H, the spatial width W, and the square of the total channels C. Furthermore, this conventional approach forces the neural network to actively learn a massive weight matrix containing an immense volume of individual floating-point parameters, severely exacerbating the risk of overfitting and elevating the memory consumption during both the forward and backward propagation phases. In stark contrast, our methodology completely bypasses these severe computational bottlenecks. By projecting the features directly onto the frequency domain, the spatial aggregation is executed through a strictly localized mathematical inner product. This operation mathematically reduces the computational complexity to an absolute linear scale, demanding a minimal number of multiplications proportional to the feature resolution without introducing any spatial kernel overhead.

The profound architectural advantage of this proposed methodology lies explicitly in its absolute zero-parameter nature regarding the highly complex frequency analysis stage. Unlike conventional dynamic kernel-based approaches or sophisticated spatial attention modules that necessitate the widespread deployment of numerous learnable kernels to perceive structural variations, our methodology is fundamentally distinct. The spectral basis functions utilized in our formulation are strictly predetermined mathematical constants derived directly from the inherent analytical properties of the discrete cosine transform. They require absolutely no internal learnable parameters and fundamentally consume exactly zero memory footprint during the continuous backpropagation optimization process. Because these transform basis functions can be effectively pre-computed and statically cached directly in the graphical device memory prior to the commencement of the global training phase, the forward propagation only entails a highly optimized and extremely rapid tensor multiplication. This elegant analytical property unconditionally guarantees that our dynamic gating mechanism operates as a seamless plug-and-play architectural module. It achieves highly complex, content-based, comprehensive frequency-selective feature aggregation, completely overcoming the limitations of static methods. Consequently, it rigorously maintains the strict computational efficiency and fundamental lightweight characteristics intrinsically mandated by real-world, high-definition super-resolution applications.

3.5. Datasets, Metrics and Implementation

Following the rigorous and standard evaluation protocols universally established in state-of-the-art image restoration literature, such as SwinIR [18] and HAT [19], we conduct all our neural network training on the widely adopted DIV2K dataset. The DIV2K dataset comprises 800 highly diverse, 2K-resolution high-quality training images, covering a comprehensive spectrum of real-world scenarios including natural landscapes, complex architectural facades, flora, and intricate human-made objects. The exceptional resolution and rich textural variety of this dataset make it an ideal foundation for training deep frequency-aware models. For the rigorous evaluation of our proposed methodology, we employ a diverse suite of standard benchmark testing datasets: Set5, Set14, and Urban100, which consist of 5, 14, and 100 images respectively. Each dataset serves a distinct analytical purpose: Set5 and Set14 evaluate basic geometric structures and contours. Urban100 contains immensely complex man-made architectural structures with dense, repetitive

 doi:10.20944/preprints202604.1446.v1

parallel patterns. To synthetically generate the corresponding Low-Resolution input images for training and testing, we apply the standard bicubic downsampling operation, utilizing the MATLAB imresize function, to the High-Resolution (HR) ground truth images. This mathematically defined degradation process guarantees a fair and consistent comparison with all existing baseline methodologies.

To precisely quantify the fidelity of the SR reconstructions, we employ two universally recognized full-reference metrics: Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). PSNR calculates the absolute pixel-level mean squared error between the reconstructed image and the ground truth, providing a strict mathematical evaluation of signal fidelity. Conversely, SSIM evaluates the perceptual quality by explicitly measuring the degradation of structural information, luminance, and contrast, which highly correlates with the human visual system's perception. In strict accordance with standard super-resolution conventions, both PSNR and SSIM are calculated exclusively on the luminance Y channel of the transformed YCbCr color space. The human eye is biologically far more sensitive to structural variations in luminance than to high-frequency shifts in chrominance. Furthermore, to prevent evaluation inaccuracies caused by boundary padding artifacts during convolution, we crop a border of s pixels, where s corresponds to the upscaling factor, from all evaluated images prior to metric computation.

Our proposed network, DRCT-FAFG, is meticulously implemented utilizing the PyTorch deep learning framework and the highly optimized BasicSR open-source toolbox. The foundational architecture of our model strictly aligns with the topological configuration of the original DRCT [25]. Specifically, the deep feature extraction module consists of 6 RDG, maintaining a latent embedding dimension of 180, a localized window size of 16×16, and an attention head configuration of [6,6,6,6,6,6]. The proposed FAFG is seamlessly integrated immediately following the token mixing stage within every STL across all RDG. During the training phase, the LR input patches are randomly cropped to a spatial dimension of 64×64, corresponding to an HR ground-truth patch size of 256×256 for the $\times 4$ super-resolution task. To prevent structural overfitting and heavily enhance the model's spatial robustness, we apply extensive data augmentation techniques, including random horizontal and vertical flips, as well as orthogonal rotations of 90°, 180° and 270°. The network is optimized using the L_1 loss function, which minimizes the absolute pixel-wise differences and has been proven to yield sharper edge reconstructions compared to the L_2 loss. We utilize the Adam optimizer with momentum parameters set to $\beta_1 = 0.9 \text{and} \beta_2 = 0.999$, alongside an epsilon value of $\epsilon = 10^{-8}$. The global batch size is configured to 2.

4. Results

4.1. Ablation Study

To systematically deconstruct our architectural design and rigorously investigate the isolated effectiveness of the proposed FAFG, we conduct comprehensive ablation studies. All ablation variants are evaluated on the Urban100 dataset at a scale factor of 4. Urban100 is intentionally selected for these critical experiments because it is notably rich in high-frequency textures, dense architectural grids, and repetitive geometric shapes. Consequently, it serves as an extremely highly sensitive touchstone for evaluating a network's true capability in feature aggregation and high-frequency restoration.

To validate our design choices, we meticulously construct and evaluate several alternative gating and attention strategies against our proposed DCT-based module. The quantitative results of these extensive experiments are systematically summarized in Table 1. To establish a rigid lower bound, we first evaluate the DRCT model [25], which explicitly relies on a static residual scaling strategy. This method multiplies the entire residual feature map by a singular, constant scalar. As mathematically anticipated, this content-agnostic approach yields the lowest overall performance. The severe underperformance confirms our primary hypothesis: enforcing a fixed scaling parameter across highly heterogeneous spatial regions indiscriminately suppresses critical high-frequency

 doi:10.20944/preprints202604.1446.v1

textures alongside low-frequency background noise, fundamentally bottlenecking the representation capacity.

Table 1. Quantitative ablation results on the DIV2K dataset.

Method	Set5		Set14		Urban100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DRCT**Error! Reference source not found.**	30. 31	0. 8605	27. 36	0. 7512	24. 22	0. 7139
GAP + MLP + Sigmoid	30. 46	0. 8638	27. 45	0. 7533	24. 34	0. 7178
Dual-Pooling (GAP+GMP)+ MLP + Sigmoid	30. 45	0. 8641	27. 47	0. 7535	24. 35	0. 7193
GAP + ECA + Sigmoid	30. 56	0. 8669	27. 53	0. 7557	24. 42	0. 7230
GAP + MLP + Sigmoid + Spatial Attention	30. 57	0. 8664	27. 52	0. 7552	24. 42	0. 7223
Dual-Pooling + MLP + Sigmoid + Spatial Attention	30. 57	0. 8665	27. 53	0. 7553	24. 42	0. 7224
DCT + MLP + Sigmoid + Spatial Attention	30. 60	0. 8674	27. 54	0. 7557	24. 45	0. 7237
DCT + MLP + Sigmoid (Ours)	**30. 62**	**0. 8679**	**27. 58**	**0. 7564**	**24. 46**	**0. 7243**

Recognizing the severe limitations imposed by the static multiplier, we introduce a rudimentary form of dynamic modulation by experimenting with a gating mechanism driven by GAP, which is structurally identical to the standard channel attention mechanisms employed in legacy models like RCAN [13]. While this dynamic variant achieves a marginal improvement over the static method, the gain is severely limited because GAP is mathematically equivalent to an extreme low-pass filter. it merely extracts the mean color intensity, effectively the zero-frequency or DC component, without perceiving intricate structural densities. To validate our design choices, we meticulously construct and evaluate several alternative gating and attention strategies against our proposed DCT-based module. The quantitative results are systematically summarized in Table **1**. To establish a rigid lower bound, we first evaluate the DRCT [25] model, which explicitly relies on a static residual scaling strategy. This method multiplies the entire residual feature map by a singular, constant scalar. As mathematically anticipated, this content-agnostic approach yields the lowest overall performance. The severe underperformance confirms our primary hypothesis: enforcing a fixed scaling parameter across highly heterogeneous spatial regions indiscriminately suppresses critical high-frequency textures alongside low-frequency background noise, fundamentally bottlenecking the representation capacity.

Recognizing the limitations in DRCT [25], we introduce dynamic modulation by experimenting with a gating mechanism driven by GAP (GAP-MLP), which is structurally identical to the standard channel attention. While this variant improves over the static method, the gain is limited because GAP acts as an extreme low-pass filter, merely extracting mean color intensity without perceiving intricate structural densities. To address this, we explore a Dual-Pooling strategy (Dual-MLP) concatenating GAP and GMP. Although GMP captures salient boundaries, it remains trapped in the spatial domain. Furthermore, to test if GAP's bottleneck was due to MLP dimensionality reduction, we replaced the MLP with an Efficient Channel Attention mechanism (GAP-ECA). As shown in Table 1, although GAP-ECA preserves channel correspondences better, its performance remains significantly lower than our proposed method, confirming that the root cause is GAP's spatial ambiguity, not MLP compression.

In a further attempt to counter this ambiguity, we investigated hybrid architectures by appending Spatial Attention to various gating mechanisms. Adding it to spatial-pooling variants (GAP-Spatial and Dual-Spatial) yields only marginal improvements, proving that spatial convolutions cannot fully recover high-frequency information destroyed by pooling. Counterintuitively, when appended to our DCT gate (DCT-Spatial), it yields slightly lower performance than the pure DCT gate alone. This suggests our DCT-based projections already capture sufficient cues. adding redundant convolutions disrupts the calibrated frequency weights, resulting in optimization conflicts. Finally, our proposed pure DCT-based FAFG (Ours) achieves the highest performance, outperforming DRCT [25] by a significant margin. This definitively demonstrates that modeling continuous frequency domain information is essential for accurate super-resolution.

 doi:10.20944/preprints202604.1446.v1

4.2. Efficiency Analysis

A paramount advantage of our proposed FAFG methodology is its strict mathematical elegance and extreme computational efficiency, particularly when compared to other contemporary frequency-aware or heavy self-attention approaches. We specifically analyze the parameter overhead and FLOPs associated with our gating strategy.

As rigorously formulated in Section 3, our integration of the Discrete Cosine Transform relies exclusively on predefined, mathematically fixed basis functions denoted as $B_{u,v,i,j}$. These continuous cosine waves are statically cached in the computational graph and inherently require absolutely zero learnable parameters and no gradient updates during the backpropagation phase. Consequently, the incredibly complex spectral decomposition step imposes zero parameter overhead on the network. The only minimal parameters introduced by our module reside within the lightweight MLP responsible for generating the final Sigmoid gating weights. Because this MLP utilizes a massive channel reduction ratio r, the resulting parameter addition is astronomically small, typically accounting for less than 0. 1% of the total parameters in a standard Swin Transformer block.

Quantitative complexity comparisons, evaluated on the Urban100 dataset for x4 upscaling, are comprehensively presented in Table 2. In stark contrast to networks like FcaNet [26] or HAT [19], which deploy massive overlapping convolutional kernels or complex cross-window attention mechanisms that heavily inflate both memory footprint and latency, our methodology achieves a remarkable surge in performance while maintaining a parameter count and FLOPs measurement nearly identical to the static method. This flawless equilibrium unequivocally validates that FAFG achieves a vastly superior trade-off between reconstructive performance and architectural complexity, rendering it highly viable for real-world deployment on resource-constrained devices.

Table 2. Model complexity and performance comparison on Urban100 and Set14.

Method	Params (M)	FLOPs (G)
DRCT [25]	14. 1396	59. 6449
Ours (DRCT-FAFG)	14. 1883	59. 6512
Difference	**+0. 0487**	**+0. 0063**

4.3. Comparison with State-of-the-Art Methods

To fully contextualize the capabilities of our optimized architecture under a strictly controlled training protocol, we comprehensively compare the proposed DRCT-FAFG with the DRCT [25] alongside powerful recent Transformer-based approaches, specifically SwinIR [18] and HAT [19]. The quantitative benchmark results for super-resolution at a massive upscaling factor of $\times 4$ are systematically summarized in Table 3.

As explicitly detailed in the table, our proposed method consistently and robustly outperforms the static DRCT configuration and competing state-of-the-art models across all evaluated datasets. The performance margin is particularly striking on the challenging Urban100 dataset. Urban100 is widely recognized by the global computer vision community as the ultimate, most rigorous touchstone for Single Image Super-Resolution algorithms due to its unforgiving density of man-made geometries, heavily aliased parallel lines, and complex windows. The substantial performance gain achieved by FAFG on this specific dataset powerfully corroborates our core theoretical hypothesis: dynamic, frequency-aware feature modulation is disproportionately effective and essential for recovering high-frequency textures that are inevitably over-smoothed, blurred, or completely annihilated by static, content-agnostic scaling strategies. Furthermore, the stable performance improvements consistently recorded on standard validation sets like Set5 and Set14 demonstrate the incredible robustness, content-adaptability, and broad generalization capabilities of our frequency-gated network when compared directly to recent highly-optimized Transformer architectures **Error! Reference source not found.Error! Reference source not found.**.

 doi:10.20944/preprints202604.1446.v1

Table 3. Comparison with state-of-the-art methods.

Method	Set5		Set14		Urban100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DRCT**Error! Reference source not found.**	30. 31	0. 8605	27. 36	0. 7512	24. 22	0. 7139
SwinIR **Error! Reference source not found.**	30. 08	0. 8540	27. 15	0. 7432	24. 08	0. 7046
HAT **Error! Reference source not found.**	30. 11	0. 8553	27. 18	0. 7441	24. 09	0. 7055
Ours (DRCT-FAFG)	**30. 62**	**0. 8679**	**27. 58**	**0. 7564**	**24. 46**	**0. 7243**

To definitively illustrate the perceptual and qualitative superiority of our proposed methodology beyond mere numerical metrics, we present detailed visual comparisons derived from the notoriously difficult Urban100 dataset, as well as the standard Set14 benchmark. Figure 5 visualizes the highly magnified $\times 4$ super-resolved outputs of our FAFG network directly against the DRCT [25], alongside recent state-of-the-art Transformer architectures SwinIR [18] and HAT [19].

As clearly observed in the localized image patches, the competing models frequently suffer from catastrophic spectral confusion and severe aliasing artifacts when tasked with reconstructing dense, repetitive high-frequency patterns. Highly illustrative examples of this failure are found in Img_004 and Img_092 (Urban100), which depict complex, repeating grid structures on distant building facades. In these scenarios, SwinIR, HAT, and the static DRCT model entirely fail to distinguish the precise mathematical orientation of the orthogonal stripes. Because their feature aggregation relies on content-agnostic spatial routing or static scaling, it effectively suffocates the high-frequency restorative signals. This results in highly blurred, structurally distorted reconstructions heavily plagued by the infamous Moiré effect—a false pattern generated by improper signal sampling. Similar severe blurring and structural decay are evident in the dense, high-frequency striped fabric of Img_002 (Set14) and the sharp contours of Img_012 (Set14), where legacy attention mechanisms fail to recover crisp, continuous boundaries.

In absolute contrast, our proposed DRCT-FAFG effortlessly neutralizes these severe visual artifacts. Empowered by the explicit multi-spectral DCT projections within the fusion gate, our network continuously perceives the precise dominant directionality and localized spectral energy of the texture components before making any fusion decisions. Consequently, whether reconstructing the precise architectural grids in Img_092 and Img_004, restoring the dense parallel lines of the fabric in Img_002, or preserving the sharp geometric boundaries in Img_012, our model accurately suppresses out-of-band noise while heavily amplifying the valid directional gradients. It successfully reconstructs perfectly straight, pristine lines and highly defined contours, producing a final output that is structurally faithful and phenomenally closer to the High-Resolution Ground Truth. This overwhelming visual superiority serves as the ultimate confirmation that integrating explicit, parameter-free frequency priors enables the deep network to intelligently separate structural textures from degradation noise, ultimately delivering exceptionally accurate and visually pleasing high-fidelity image reconstructions.

 doi:10.20944/preprints202604.1446.v1

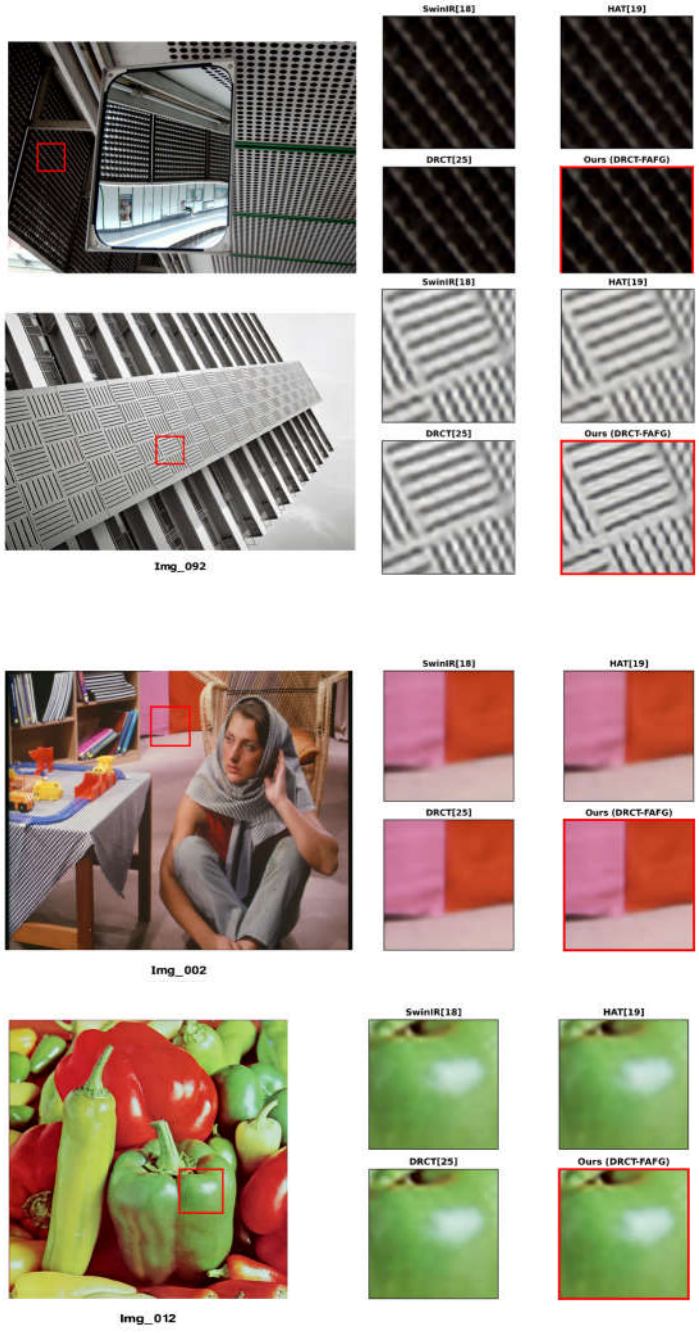

Figure 5. Visual comparison on 4 times Urban100 and Set14.

5. Discussion

5.1. Interpretation of Frequency-Aware Gating

The most compelling evidence for our architectural design stems from the ablation studies detailed in Table **1**. While channel-wise pooling methods (GAP and Dual-Pooling) offered slight improvements over the static DRCT baseline, they remained fundamentally constrained by spatial ambiguity. Because spatial averaging acts as an extreme low-pass filter, these networks struggle to differentiate between dense textures and smooth backgrounds if their mean intensities are similar.

By contrast, the integration of the 2D DCT within the FAFG explicitly decomposes spatial features into orthogonal frequency components. This allows the network to precisely quantify spectral energy. The superior performance of our purely DCT-based module (30.62 dB on Set5 and 24.46 dB on Urban100) confirms that extracting a complete spectrum of geometric periodicities is the most rigorous solution for content-aware modulation. Furthermore, our investigation revealed a

 doi:10.20944/preprints202604.1446.v1

fascinating phenomenon: appending a computationally heavy Spatial Attention module immediately following the DCT gate actually resulted in a slight performance drop (from 24.46 dB to 24.45 dB on Urban100). This counterintuitive result strongly suggests that our DCT-based projections already capture sufficient structural and spatial cues. Forcing the network to learn additional, redundant spatial convolutions disrupts the carefully calibrated frequency weights, introducing optimization conflicts and mild overfitting.

5.2. Robustness and Generalization across Diverse Scenes

The quantitative comparison (Table **3**) and visual evidence (Figure **5**) demonstrate the remarkable generalization capabilities of the DRCT-FAFG. The performance margin is particularly striking on the challenging Urban100 dataset, where our method achieved a substantial 0.31 dB gain over the static DRCT. Urban100 is characterized by an unforgiving density of man-made geometries and heavily aliased parallel lines. The fact that FAFG excels on this specific dataset powerfully corroborates that dynamic, frequency-aware feature modulation is disproportionately effective for recovering high-frequency textures that are inevitably over-smoothed by static scaling strategies. Simultaneously, the stable improvements recorded on Set5 and Set14 indicate that the FAFG does not overfit to complex textures; it intelligently adapts its gating weights to preserve smooth, low-frequency natural contours without amplifying background noise.

5.3. Societal and Environmental Implications

Beyond its technical achievements in structural reconstruction, the proposed DRCT-FAFG architecture demonstrates profound societal and environmental implications across its entire deployment life-cycle. From a societal perspective, by significantly enhancing the visual fidelity of degraded images, this technology directly addresses critical needs in high-stakes professional domains, such as medical diagnostics and public security. In clinical settings, the ability to reconstruct high-frequency details from low-resolution scans can substantially accelerate accurate diagnoses and potentially reduce the need for prolonged patient exposure to harmful imaging radiation. Furthermore, in public safety applications, recovering sharp geometric features from corrupted surveillance footage provides crucial evidentiary support. However, developers must remain ethically vigilant, as highly realistic image reconstruction technologies could potentially be misused to generate deceptive media. Ensuring the responsible deployment of such models requires strict adherence to ethical guidelines and the potential integration of digital watermarking techniques.

From an environmental standpoint, the proposed methodology explicitly addresses the urgent ecological concerns associated with modern deep learning. The exponential growth in parameter counts for contemporary Vision Transformers inherently leads to massive electricity consumption and substantial carbon emissions during both the prolonged training phase and the continuous inference life-cycle. By utilizing predefined, mathematically constant basis functions for spectral decomposition, the FAFG module operates with exactly zero learnable parameters during its complex frequency analysis stage. As demonstrated in the preceding efficiency analysis, this intelligent lightweight design allows the network to achieve state-of-the-art reconstructive performance without inflating the computational footprint. Consequently, deploying this frequency-aware architecture on resource-constrained edge devices drastically reduces real-world energy expenditures, strongly aligning with the global mandate for Green AI and promoting a highly sustainable trajectory for future computer vision research.

6. Conclusions

In this paper, we have identified and addressed a critical limitation in the feature aggregation mechanism of the state-of-the-art Dense-Residual-Connected Transformer or DRCT. We argued that the conventional static residual scaling strategy, which relies on a fixed and learnable scalar, restricts the representational capability of the network by treating high-frequency textures and low-frequency

 doi:10.20944/preprints202604.1446.v1

backgrounds indiscriminately. This content-agnostic approach often leads to the over-smoothing of fine details in complex image regions.

To overcome this bottleneck, we proposed the Frequency-Aware Adaptive Fusion Gate or FAFG. By integrating the Discrete Cosine Transform or DCT into the residual path, our method dynamically modulates feature integration based on spectral characteristics. A distinguishing feature of our design is the use of pre-calculated and fixed DCT basis functions, which ensures that the frequency analysis step remains zero-parameter and computationally efficient. This allows the network to selectively emphasize high-frequency structural components without incurring the heavy parameter costs associated with learnable frequency layers or complex self-attention mechanisms.

Extensive experiments on standard benchmark datasets have demonstrated the superiority of our approach. Specifically, compared to the static method, our approach achieved a significant performance improvement of 0. 31 decibels on the textured Urban100 dataset. Visual comparisons further confirm that FAFG effectively recovers complex structural details and sharpens edges that are otherwise blurred by static scaling. While our experiments were conducted under a lightweight reproduction setting due to computational constraints, the consistent improvements validate the effectiveness of our core hypothesis.

In future work, we intend to expand upon these foundational achievements through several promising avenues. First, we plan to validate our proposed FAFG methodology using a full-scale, unconstrained training schedule to observe its peak performance capabilities against fully saturated generative models. Second, we aim to extend the application of this zero-parameter spectral routing mechanism to other highly challenging low-level vision tasks, such as image denoising and image deblurring, where dynamic frequency-aware modulation could provide similar structural benefits. Ultimately, continuing to explore parameter-free frequency priors presents a highly sustainable and effective trajectory for the future of deep image restoration.

Author Contributions: Conceptualization, Q.L. and R.C.; methodology, Q.L.; software, Q.L.; validation, Q.L.; formal analysis, Q.L.; writing—original draft preparation, Q.L.; writing—review and editing, Q.L. and R.C.; supervision, R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The DIV2K, Set5, Set14, and Urban100 datasets analyzed in this study are publicly available standard benchmark datasets.

Acknowledgments: During the preparation of this manuscript, the author(s) used Gemini for the purposes of language polishing and formatting. The authors have reviewed and edited the output and take full responsibility for the content of this publication. We also acknowledge the computational resources provided by the Faculty of Applied Sciences at Macao Polytechnic University.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SISR	Single Image Super-Resolution
DRCT	Dense-Residual-Connected Transformer
FAFG	Frequency-Aware Adaptive Fusion Gate
DCT	Discrete Cosine Transform
RDG	Residual Dense Group
CNN	Convolutional Neural Network
GAP	Global Average Pooling
GMP	Global Max Pooling
MLP	Multi-Layer Perceptron
SDRCB	Swin Dense Residual Connected Block

 doi:10.20944/preprints202604.1446.v1

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444.
2. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. Comput. Intell. Neurosci. 2018, 2018.
3. Wang, Z.; Chen, J.; Hoi, S.C. Deep learning for image super-resolution: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 43, 3365–3387.
4. Anwar, S.; Khan, S.; Barnes, N. A deep journey into super-resolution: A survey. ACM Comput. Surv. 2020, 53, 1–34.
5. Baker, S.; Kanade, T. Limits on super-resolution and how to break them. IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 1167–1183.
6. Yang, C.Y.; Ma, C.; Yang, M.H. Single-image super-resolution: A benchmark. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014; pp. 372–386.
7. Blau, Y.; Michaeli, T. The perception-distortion tradeoff. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018; pp. 6228–6237.
8. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 38, 295–307.
9. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017; pp. 136–144.
10. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017; pp. 4681–4690.
11. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; et al. ESRGAN: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 2018.
12. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 45, 4713–4726.
13. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018; pp. 286–301.
14. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019; pp. 11065–11074.
15. Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; et al. Single image super-resolution via a holistic attention network. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 2020; pp. 191–207.
16. Dosovitskiy, A.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 2021.
17. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; et al. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021; pp. 12299–12310.
18. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021; pp. 1833–1844.
19. Chen, X.; Wang, X.; Zhou, J.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023; pp. 22367–22377.
20. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016; pp. 391–407.

 doi:10.20944/preprints202604.1446.v1

21. Ahn, N.; Kang, B.; Sohn, K.A. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018; pp. 252–268.
22. Ma, C.; Rao, Y.; Cheng, Y.; Chen, C.; Lu, J.; Zhou, J. Structure-preserving super resolution with gradient guidance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020; pp. 2776–2784.
23. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017; pp. 624–632.
24. Li, J.; Fang, F.; Mei, K.; Zhang, G. Multi-scale residual network for image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018; pp. 517–532.
25. Hsu, C.C.; Lee, C.M.; Chou, Y.S. DRCT: Saving image super-resolution away from information bottleneck. arXiv 2024, arXiv:2404.00722.
26. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021; pp. 783–792.
27. Isaac, J.S.; Kulkarni, R. Super resolution techniques for medical image processing. In Proceedings of the International Conference on Technologies for Sustainable Development (ICTSD), Mumbai, India, 2015; pp. 1–6.
28. Zou, W.W.; Yuen, P.C. Very low resolution face recognition problem. IEEE Trans. Image Process. 2011, 21, 327–340.
29. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. Commun. ACM 2020, 63, 54–63.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018; pp. 7132–7141.
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018; pp. 3–19.
32. Liu, D.; Wen, B.; Fan, Y.; Loy, C.C.; Huang, T.S. Non-local recurrent network for image restoration. Adv. Neural Inf. Process. Syst. 2018, 31.
33. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019; pp. 510–519.
34. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021; pp. 3560–3569.
35. Chen, L.; Chu, X.; Zhang, X.; Sun, J. Simple baselines for image restoration. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 2022; pp. 17–33.
36. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022; pp. 5728–5739.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016; pp. 770–778.
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021; pp. 10012–10022.
39. Lin, M.; Chen, Q.; Yan, S. Network in network. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 2014.
40. Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. IEEE Trans. Comput. 1974, 100, 90–93.
41. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Highway networks. arXiv 2015, arXiv:1505.00387.