SYMMETRIC ATTENTION IMPROVES OPTIMIZATION GEOMETRY FOR IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-attention has become the workhorse of Transformer-based image restoration, yet its *optimization behavior* remains largely underexplored. We identify two structural issues arising from the standard asymmetric score matrix $S = QK^\top$: (i) the gradients with respect to W_Q and W_K are governed by *structurally different* operators, allowing the two projections to drift along uncoupled directions and creating flat valleys in the loss landscape; (ii) the asymmetric score matrix may admit *complex eigenvalues*, whose imaginary components induce rotational dynamics during gradient descent. Together, these effects lead to ill-conditioned optimization and unstable training. Our main contribution, the **Balanced Kernel (BK)**, resolves both issues through a lightweight symmetric score construction that couples the gradient dynamics of W_Q and W_K and guarantees a real eigenvalue spectrum. BK introduces **negligible additional FLOPs** while achieving $\sim 3\times$ faster convergence, $> 4\times$ oscillation reduction, and a substantially more compact optimization basin. When integrated into existing architectures such as Restormer and Uformer, BK consistently improves performance, demonstrating largely **architecture-agnostic** optimization benefits. We further introduce **ChannelRouter**, which generates instance-adaptive head weights from degradation-aware channel statistics, and **Gated Feature Modulation (GFM)**, which enables spatially adaptive processing through multiplicative gating. Integrating these components, our **Balformer** achieves state-of-the-art performance on deraining, deblurring, and denoising benchmarks.

1 INTRODUCTION

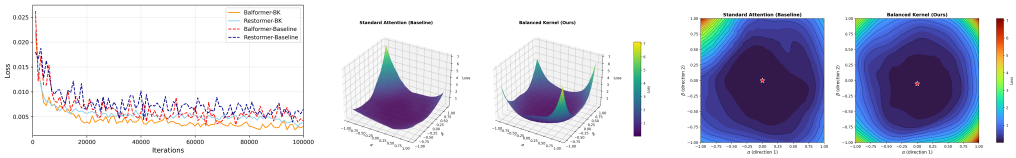
Transformers now dominate image restoration, with SwinIR (Liang et al., 2021), Restormer (Zamir et al., 2022), Uformer (Wang et al., 2022), and DAT (Chen et al., 2023c) achieving strong performance across denoising, deraining, and deblurring benchmarks. Improvements over these models have primarily focused on three directions: enlarging the effective receptive field, refining tokenization, and reducing the quadratic complexity of attention. In this work, we explore a fourth direction that has received comparatively little attention in the image restoration literature: *the optimization geometry of the attention operator itself*. As previewed in Fig. 1, standard attention consistently exhibits optimization issues during training, including slow, oscillatory convergence and a poorly conditioned loss landscape, across multiple restoration architectures.

Structural issues in standard attention. The scaled dot-product score matrix $S = QK^\top$ is asymmetric by construction ($S_{ij} \neq S_{ji}$ in general). While this asymmetry does not hinder forward representation learning, we show that it introduces two coupled issues during optimization. First, differentiating the loss with respect to the projection matrices yields

$$\frac{\partial \mathcal{L}}{\partial W_Q} = \tau X^\top G X W_K, \qquad \frac{\partial \mathcal{L}}{\partial W_K} = \tau X^\top G^\top X W_Q, \tag{1}$$

where $G = \partial \mathcal{L}/\partial S$ is generally asymmetric. The two projections are governed by *structurally different* operators ($X^\top G X$ *vs.* $X^\top G^\top X$), allowing W_Q and W_K to evolve along directions that do not mutually compensate. This leads to extended flat valleys in the loss landscape (Fig. 1(b,c)). Second, because S is asymmetric, it may admit complex eigenvalues, whose imaginary components induce rotational dynamics in the gradient flow and manifest as persistent oscillations in the training loss

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b) (c)

Figure 1: **Optimization issues in standard attention and their resolution by the Balanced Kernel.** We compare standard attention (dashed, "Baseline") with the proposed Balanced Kernel (solid, "BK") on two architectures (Restormer and our Balformer) using identical training settings on Rain200L (Yang et al., 2017). **(a) Training loss.** The baselines exhibit slow convergence and persistent high-amplitude oscillations, consistent with the rotational gradient dynamics induced by complex eigenspectra (Sec. 3.2). In contrast, BK converges $\sim 3\times$ faster while reducing oscillation amplitude by more than $4\times$. **(b) 3D loss landscape** (filter-normalized (Li et al., 2018a)). The baseline exhibits an elongated, poorly conditioned valley, reflecting the decoupled optimization of W_Q and W_K. In contrast, BK produces a compact, near-isotropic basin. **(c) Contour projection.** BK yields smooth, nearly concentric iso-loss contours, whereas the baseline exhibits irregular contours indicative of poor optimization conditioning. Together, these visualizations illustrate the two structural issues analyzed in Sec. 3.2—gradient decoupling and complex eigenspectra—and show how the proposed symmetric score construction, $S_{\text{sym}} = (QK^\top + KQ^\top) \cdot \tau$, improves optimization behavior. Similar improvements are observed for both architectures, indicating that the benefits of BK generalize across different Transformer backbones.

(Fig. 1(a), dashed curves). Together, these two properties result in an **ill-conditioned optimization landscape**, slowing convergence and reducing training stability across diverse restoration tasks and architectures.

Balanced Kernel: efficient symmetrization. We address both issues through a lightweight algebraic construction:

$$S_{\text{sym}} \;=\; (QK^\top + KQ^\top) \cdot \tau. \tag{2}$$

This construction (i) causes W_Q and W_K to share the same symmetric gradient operator $\tilde{H} = \tau X^\top (G + G^\top) X$, eliminating gradient decoupling; and (ii) guarantees a real eigenvalue spectrum with an orthonormal eigenbasis, thereby removing the rotational dynamics induced by complex eigenvalues. The implementation requires only one matrix transpose and one addition per attention computation, introducing negligible computational overhead. Empirically, the Balanced Kernel achieves $\sim 3\times$ faster convergence, more than $4\times$ lower oscillation, and a substantially more compact loss basin (Fig. 1). Cross-architecture experiments on Restormer and Uformer further show that these improvements consistently transfer across different Transformer backbones, demonstrating that the optimization benefits of BK are largely architecture-independent (Sec. 4, Sec. 4.3.2).

Complementary architectural enhancements. While the Balanced Kernel improves the optimization behavior of attention, standard Transformer blocks also lack **input-adaptive capacity allocation** along two complementary dimensions. Along the *channel dimension*, the static output projection assigns fixed relative weights to the N_h attention heads regardless of the degradation characteristics of the input. Along the *spatial dimension*, feed-forward branches apply identical transformations across all locations despite spatially varying restoration difficulty. To address these limitations, we introduce two lightweight modules. **ChannelRouter** (Sec. 3.3) generates instance-adaptive head weights from degradation-aware channel statistics (mean, standard deviation, and energy), while **Gated Feature Modulation** (Sec. 3.4) performs content-adaptive multiplicative gating in both the MLP and multi-scale convolution branches. These modules are orthogonal to BK and provide complementary performance gains (Sec. 4.3, Appendix A.7).

Balformer. Integrating BK, ChannelRouter, and Gated Feature Modulation into a U-shaped encoder–decoder yields **Balformer**, which achieves state-of-the-art performance on deraining (Rain200L/H, DID-Data, DDN-Data), single-image defocus deblurring (DPDD), and real-image denoising (SIDD).

Contributions.

- **(Theoretical & Methodological)** We identify two structural optimization issues in standard attention—gradient decoupling between W_Q and W_K and the possibility of complex eigenspectra—and propose the **Balanced Kernel**, a simple symmetric score construction that addresses both through a lightweight algebraic symmetrization with negligible computational overhead. The resulting optimization geometry yields $\sim 3\times$ faster convergence and a substantially better-conditioned loss landscape. Cross-architecture experiments demonstrate that these improvements consistently generalize across different Transformer backbones.
- **(Architectural)** We introduce **ChannelRouter** and **Gated Feature Modulation** as lightweight, complementary enhancements that improve channel- and spatial-level adaptivity in standard Transformer blocks.
- **(Empirical)** Integrating these components into **Balformer**, we achieve state-of-the-art performance across three representative image restoration tasks. A dedicated repeated-structure benchmark (Appendix A.8) further shows that the improved optimization geometry of BK leads to more consistent reconstruction of visually similar patches.

2 Related Work

2.1 Image Restoration

CNN-based methods such as SRCNN (Dong et al., 2016), EDSR (Lim et al., 2017), and RCAN (Zhang et al., 2018) significantly improved image restoration through residual learning and channel attention, but remain limited by the inherently local receptive field of convolution. Transformers (Vaswani et al., 2017) overcome this limitation through self-attention and have become the dominant paradigm for image restoration, with representative models including IPT (Chen et al., 2021), SwinIR (Liang et al., 2021), Restormer (Zamir et al., 2022), Uformer (Wang et al., 2022), and DAT (Chen et al., 2023c).

Alternative architectures also have notable limitations. Mamba (Gu & Dao, 2023) achieves linear complexity, but its causal formulation is not well aligned with the non-causal nature of image restoration. MambaIR (Guo et al., 2024) and MambaIRv2 (Guo et al., 2025) alleviate this issue through improved scanning strategies, yet still underperform well-optimized Transformer models. Diffusion-based methods (Wang et al., 2024; Saharia et al., 2022) deliver impressive perceptual quality but require iterative inference and substantially higher computational cost. Motivated by these trade-offs, we focus on improving Transformer-based image restoration.

2.2 Attention as Kernel Methods

Tsai et al. (2019) showed that scaled dot-product attention can be interpreted as a kernel smoother:

$$\text{Attention}(x_q; \mathcal{S}_{x_k}) = \sum_{x_k \in \mathcal{S}} \frac{k(x_q, x_k)}{\sum_{x_{k'} \in \mathcal{S}} k(x_q, x_{k'})} v(x_k), \tag{3}$$

where $k(\cdot, \cdot)$ is a non-negative kernel measuring pairwise similarity, and $\mathcal{S} \subseteq \mathcal{S}_{x_k}$. This perspective decomposes attention into four design components: the kernel function, positional encoding, value mapping, and set filtering, providing a unified framework for diverse Transformer architectures. Their analysis further shows that infinite-dimensional kernels (e.g., exponential and RBF kernels) consistently outperform finite-dimensional alternatives, while symmetric kernels achieve comparable performance with fewer parameters.

Subsequent work has focused on approximating the softmax kernel to improve attention efficiency. Performer (Choromanski et al., 2020) employs positive random features, Linformer (Wang et al., 2020b) adopts low-rank key-value projections, Reformer (Kitaev et al., 2020) exploits locality-sensitive hashing, and Katharopoulos et al. (2020) reformulate attention as a linear-complexity recurrence using kernel feature maps. These methods generally assume symmetric Mercer kernels and therefore do not consider the intrinsic asymmetry of the attention score matrix ($S_{ij} \neq S_{ji}$). Chen et al. (2023b) revisited this issue through asymmetric Kernel SVD in Reproducing Kernel Banach spaces, showing that standard attention exploits only the right singular projections. Their Primal-Attention recovers both left and right projections through a primal formulation without explicitly constructing the $\mathcal{O}(N^2)$ attention matrix.

Unlike Primal-Attention, which addresses asymmetry through dual-space decomposition, our work directly symmetrizes the attention score matrix using a lightweight algebraic construction. Rather than improving representation capacity, our goal is to improve the optimization geometry induced by attention by coupling the optimization of W_Q and W_K and ensuring a real eigenvalue spectrum (Sec. 3.2, Sec. 4.3.2). Our formulation is also fully compatible with the kernel perspective of Tsai et al. (2019) and can be naturally combined with existing kernel approximation techniques for efficient long-sequence modeling.

3 METHODOLOGY

3.1 OVERALL ARCHITECTURE

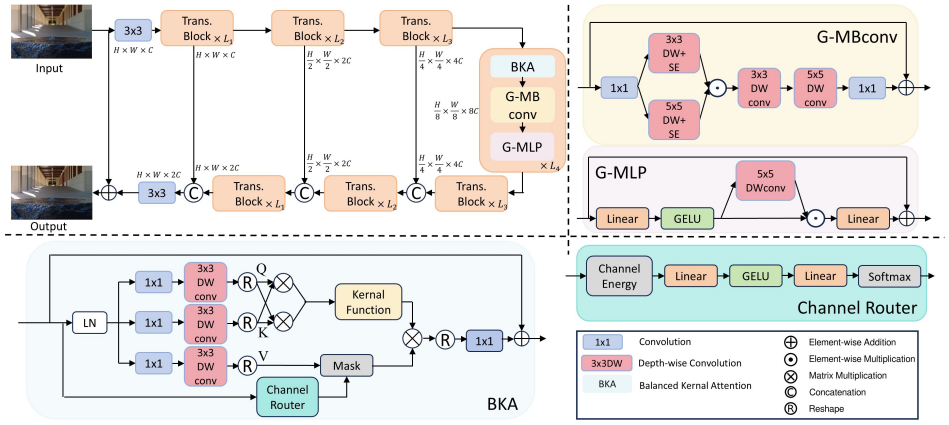


Figure 2: **Overview of Balformer.** Top left: the U-shaped encoder–decoder architecture with skip connections. Each stage consists of stacked Transformer Blocks, each comprising Balanced Kernel Attention (BKA), Gated MBConv (G-MBConv), and Gated MLP (G-MLP). The remaining panels illustrate the detailed architectures of these modules.

Balformer adopts a U-shaped encoder–decoder architecture (Fig. 2). A 3×3 convolution first projects the input to feature maps of size $H \times W \times C$. The encoder progressively downsamples the features through three stages to a bottleneck of size $\frac{H}{8} \times \frac{W}{8} \times 8C$, while the decoder mirrors this process using skip connections. Each stage stacks $[4, 6, 6, 8]$ Transformer Blocks (from shallow to deep), each comprising three residual blocks.

BKA. After Layer Normalization, the input is projected into Q, K, and V through three parallel 1×1 convolutions, each followed by a 3×3 depth-wise convolution. Q and K are fed into the Balanced Kernel (Sec. 3.2) to compute symmetric attention scores. The aggregated value features are then modulated by instance-adaptive head routing weights (Sec. 3.3) before a final 1×1 projection.

G-MBConv. This block extends MBConv (Tan & Le, 2019) with Squeeze-and-Excitation (Hu et al., 2018) and a learned spatial gate that adaptively fuses dual-scale (3×3 and 5×5) depth-wise convolutions according to local content.

G-MLP. Features are split into two streams, one of which is modulated by a spatially adaptive gate generated from the other through a depth-wise convolution, enabling location-dependent feature selection. Full architectural details are provided in Appendix A.3.

3.2 BALANCED KERNEL

Motivation: Structural Asymmetry in Gradient Flow. As discussed in Sec. 1, the asymmetric score matrix $S = QK^\top$ introduces two coupled optimization issues: structurally different gradient operators for W_Q and W_K (Eq. 1), and a generally complex eigenspectrum whose imaginary com-

ponents induce rotational dynamics during optimization. We show that both issues can be resolved through a simple algebraic symmetrization.

Method: Symmetric Score Construction. We construct the symmetric score matrix as

$$S_{\text{sym}} = \tau(QK^\top + KQ^\top) = \tau X M X^\top, \quad M = W_Q W_K^\top + W_K W_Q^\top. \tag{4}$$

The matrix M is symmetric by construction, yielding the corresponding kernel

$$K_{\text{sym}}(x_i, x_j) = \exp(\tau x_i^\top M x_j). \tag{5}$$

This implementation requires only one additional matrix transpose and one addition per attention computation, introducing negligible computational overhead.

Gradient Coupling. Differentiating Eq. 4 gives

$$\frac{\partial \mathcal{L}}{\partial W_Q} = \tilde{H} W_K, \qquad \frac{\partial \mathcal{L}}{\partial W_K} = \tilde{H} W_Q, \quad \tilde{H} = \tau X^\top (G + G^\top) X, \quad G = \partial \mathcal{L}/\partial S. \tag{6}$$

Both projection matrices are therefore updated through the same symmetric operator $\tilde{H}$, coupling their optimization and preventing the independent drift that leads to poorly conditioned optimization landscapes.

Real Eigenvalue Spectrum. Since S_{sym} is symmetric, it admits a real eigendecomposition with an orthonormal eigenbasis. This removes the rotational dynamics induced by complex eigenvalues in the asymmetric formulation and leads to more stable optimization.

Practical Simplicity. The Balanced Kernel requires only a one-line modification to existing attention implementations, replacing $QK^\top$ with $(QK^\top + KQ^\top)$. It introduces no additional learnable parameters, requires no hyperparameter tuning, and incurs negligible computational overhead. Unlike Primal-Attention (Chen et al., 2023b), which addresses asymmetry through a primal–dual decomposition to improve representation completeness, BK directly symmetrizes the score matrix to improve optimization geometry. As a drop-in replacement, it can be readily incorporated into existing Transformer backbones, as demonstrated on Restormer and Uformer (Sec. 4).

Downstream Impact on Restoration. The coupled gradient flow and real eigenspectrum lead to substantially faster convergence ($\sim 3\times$, Sec. 4.3.2) and better-conditioned optimization landscapes (Fig. 1(b,c)). A Lipschitz analysis (Appendix A.2) further relates these optimization improvements to inference-time robustness: the sensitivity of the attention output is bounded by $\|M\|_{\text{op}}\|W_V\|_{\text{op}}$. By coupling the optimization of W_Q and W_K, BK prevents unnecessary growth of $\|M\|_{\text{op}}$, resulting in tighter empirical Lipschitz constants. On a repeated-structure benchmark (Appendix A.8), this translates into an 8–9% reduction in within-group PSNR variance across four deraining datasets.

3.3 Architectural Enhancement I: Channel-Statistical Routing

Motivation. In channel-domain multi-head attention, the channel space $\mathbb{R}^C$ is partitioned into N_h subspaces processed independently by different heads. However, the static output projection assigns fixed relative weights to all heads regardless of the input content or degradation level. Moreover, each head only observes its own (C/N_h)-dimensional feature subspace, limiting cross-head interaction. Consequently, standard multi-head attention lacks an input-adaptive mechanism for reallocating capacity across heads.

Statistics-Driven Routing. We generate instance-specific head weights from three channel-wise statistics computed from the input features: the mean μ_c (low-frequency response), the standard deviation σ_c (texture complexity), and the energy e_c (signal magnitude). These descriptors are concatenated and mapped by a lightweight two-layer MLP followed by a softmax function to obtain routing weights w (see Appendix A.3 for details). Although $e_c = \mu_c^2 + \sigma_c^2$, explicitly providing the quadratic term relieves the MLP from learning this interaction. The energy statistic is particularly

informative: for additive noise $n \sim \mathcal{N}(0, \sigma_n^2)$, $\mathbb{E}[e_c]$ increases uniformly by σ_n^2 across channels, providing a direct estimate of degradation severity.

Discussion. As detailed in Appendix A.3, the value modulation separates two complementary roles. The static mask $\mathcal{M}$ captures channels that are generally important for restoration, while the dynamic routing weights w adapt the contribution of each attention head to the current input. Applying the modulation before attention aggregation preserves the learned Q–K similarity structure while restoring cross-head coordination.

3.4 Architectural Enhancement II: Gated Feature Modulation

Motivation. Image degradation is inherently spatially non-uniform: the severity of noise, blur, and compression artifacts varies across locations, and so does the desired restoration strength. While ChannelRouter (Sec. 3.3) adapts capacity allocation at the instance level, it cannot adjust processing strength spatially. Standard MLP and convolution branches likewise apply identical transformations across all locations.

Multiplicative Gating. We introduce multiplicative gates into both the MLP and multi-scale convolution branches (Sec. 3.1). The gate is generated by a depth-wise convolution over the input features, allowing the modulation strength to vary across spatial locations. In the multi-scale branch, the gate further determines how the 3×3 and 5×5 pathways are fused: finer textures favor the smaller kernel, whereas larger structures rely more on the larger receptive field.

Discussion. ChannelRouter and Gated Feature Modulation operate at complementary levels. The former reallocates capacity across attention heads according to the input, while the latter adapts processing strength across spatial locations and receptive-field scales. Together, they replace the fixed head weighting and spatially uniform processing of standard Transformer blocks with input- and location-adaptive alternatives. These modules are orthogonal to the Balanced Kernel: BK improves the optimization geometry of attention, whereas ChannelRouter and GFM enhance its input adaptivity. Ablation studies (Sec. 4.3 and Appendix A.7) show that their effects are complementary.

4 Experiments

We evaluate Balformer on three representative image restoration tasks: image deraining, single-image motion deblurring, and image denoising. Training configurations and hyperparameters are provided in Appendix A.4. In all tables, the best and second-best results are highlighted in bold and underlined, respectively. Our evaluation has two objectives: (i) to verify that the proposed Balanced Kernel consistently improves existing Transformer backbones when used as a drop-in replacement, and (ii) to demonstrate that the full Balformer achieves state-of-the-art performance across diverse image restoration tasks.

4.1 Single Image Deraining Results

As shown in Table 1, Balformer achieves the best PSNR and SSIM on all four deraining benchmarks. The largest improvement is observed on Rain200H, where dense rain streaks heavily obscure image structures. Fig. 3 presents a challenging example containing repeated staircase patterns. Existing methods either leave visible rain residuals or over-smooth structural details, whereas Balformer restores the repeated structures more faithfully while preserving their regularity.

Effect of the Balanced Kernel. An important objective of this work is to evaluate whether the Balanced Kernel consistently benefits different Transformer backbones. Table 1 provides clear evidence: replacing only the attention score computation in Uformer (Wang et al., 2022) with $S_{\text{sym}} = (QK^\top + KQ^\top) \cdot \tau$ (Uformer-BK) improves PSNR by 0.34, 0.44, 0.21, and 0.17 dB on the four datasets, respectively. Applying the same modification to Restormer (Restormer-BK) yields additional gains of 0.22, 0.24, 0.03, and 0.09 dB. No other component, including the output projection, feed-forward network, or normalization layers, is changed. The consistent improvements across two representative Transformer architectures suggest that the performance gain primarily arises from the proposed symmetric score construction rather than architecture-specific design choices, consistent with the analysis in Sec. 3.2 and Sec. 4.3.2.

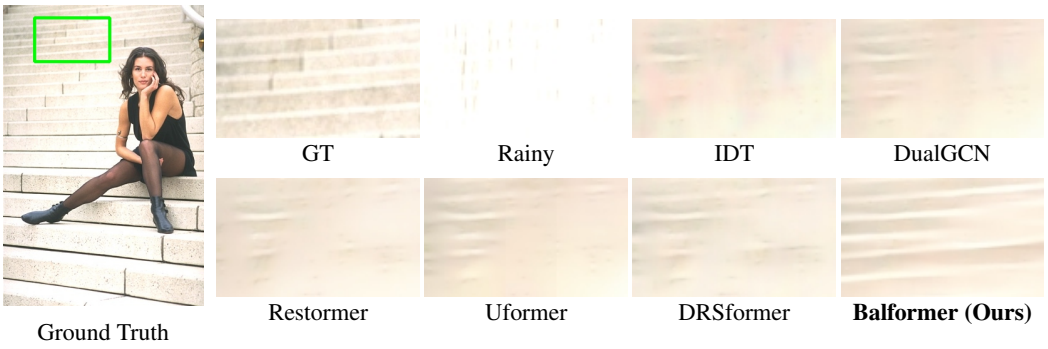


Figure 3: **Qualitative comparison on image deraining.** An example containing repeated staircase structures. Competing methods either leave residual rain streaks or restore visually similar structures with inconsistent fidelity, whereas Balformer produces more consistent reconstructions while better preserving the regularity of the repeated pattern. Additional qualitative comparisons are provided in Appendix A.5.

Table 1: **Quantitative comparison on multiple deraining datasets** (Rain200L/H (Yang et al., 2017), DID-Data (Zhang & Patel, 2018), and DDN-Data (Fu et al., 2017)). **Bold** and underlined values indicate the best and second-best results, respectively. † Results obtained by retraining the original architectures under identical settings, with only the attention score computation replaced by $S_{\text{sym}} = (QK^\top + KQ^\top) \cdot \tau$.

	Datasets	Rain200L		Rain200H		DID-Data		DDN-Data	
	Metrics	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Prior-based methods	DSC (Luo et al., 2015)	27.16	0.8663	14.73	0.3815	24.24	0.8279	27.31	0.8373
	GMM (Li et al., 2016)	28.66	0.8652	14.50	0.4164	25.81	0.8344	27.55	0.8479
CNN-based methods	DDN (Fu et al., 2017)	34.68	0.9671	26.05	0.8056	30.97	0.9116	30.00	0.9041
	RESCAN (Li et al., 2018b)	36.09	0.9697	26.75	0.8353	33.38	0.9417	31.94	0.9345
	PReNet (Ren et al., 2019)	37.80	0.9814	29.04	0.8991	33.17	0.9481	32.60	0.9459
	MSPFN (Jiang et al., 2020)	38.58	0.9827	29.36	0.9034	33.72	0.9550	32.99	0.9333
	RCDNet (Wang et al., 2020a)	39.17	0.9885	30.24	0.9048	34.08	0.9532	33.04	0.9472
	MPRNet (Zamir et al., 2021)	39.47	0.9825	30.67	0.9110	33.99	0.9590	33.10	0.9347
	DualGCN (Fu et al., 2021)	40.73	0.9886	31.15	0.9125	34.37	0.9620	33.01	0.9489
	SPDNet (Yi et al., 2021)	40.50	0.9875	31.28	0.9207	34.57	0.9560	33.15	0.9457
Transformer-based methods	Uformer (Wang et al., 2022)	40.20	0.9860	30.80	0.9105	35.02	0.9621	33.95	0.9545
	Uformer-BK†	40.54	0.9871	31.24	0.9186	35.23	0.9634	34.12	0.9561
	Restormer (Zamir et al., 2022)	40.99	0.9890	32.00	0.9329	35.29	0.9641	34.20	0.9571
	Restormer-BK†	41.21	0.9893	32.24	0.9341	35.32	0.9645	34.29	0.9584
	IDT (Xiao et al., 2022)	40.74	0.9884	32.10	0.9344	34.89	0.9623	33.84	0.9549
	DRSformer (Chen et al., 2023a)	41.23	0.9894	32.18	0.9330	35.38	0.9647	34.36	0.9590
	Balformer	**41.50**	**0.9901**	**32.40**	**0.9384**	**35.48**	**0.9657**	**34.39**	**0.9606**

Qualitative observations on Balformer. Building upon the improvements brought by BK, integrating ChannelRouter and Gated Feature Modulation further improves restoration quality, enabling Balformer to achieve the best overall performance on all four benchmarks. As illustrated in Fig. 3, Balformer produces visually consistent reconstructions across repeated staircase structures, whereas competing methods often restore similar structural elements with noticeably different fidelity.

4.2 Additional Image Restoration Results

Single-Image Defocus Deblurring. Table 2 reports results on the DPDD dataset (Abuolaim & Brown, 2020). Balformer achieves the highest PSNR on indoor, outdoor, and combined scenes, while maintaining competitive MAE performance. Fig. 4 presents a representative example of an iron fence. Existing methods either retain residual blur or produce over-smoothed tones, whereas Balformer recovers sharper textures and more faithful metallic details.

Image Denoising. Table 3 summarizes real-world denoising results on the SIDD benchmark (Abdelhamed et al., 2018). Balformer achieves 40.22 dB PSNR and 0.961 SSIM, obtaining the best performance among all compared methods.

Table 2: **Quantitative comparison on single-image defocus deblurring.** Results are reported on the DPDD dataset (Abuolaim & Brown, 2020), which contains 37 indoor and 39 outdoor scenes.

Method	Indoor		Outdoor		Combined	
	PSNR ↑	MAE ↓	PSNR ↑	MAE ↓	PSNR ↑	MAE ↓
EBDB (Karaali & Jung, 2017)	25.77	0.040	21.25	0.058	23.45	0.049
DMENet (Lee et al., 2019)	25.50	0.038	21.43	0.063	23.41	0.051
JNB (Shi et al., 2015)	26.73	0.031	21.10	0.064	23.84	0.048
DPDP (Abuolaim & Brown, 2020)	26.54	0.031	22.25	0.056	24.34	0.044
KPAC (Son et al., 2021)	27.97	0.026	22.62	0.053	25.22	0.040
GKMNet (Quan et al., 2021)	28.08	0.027	22.87	0.052	25.41	0.040
Restormer (Zamir et al., 2022)	28.87	0.025	23.24	0.050	25.98	0.038
LaKDNet (Ruan et al., 2023)	<u>29.01</u>	**0.024**	<u>23.62</u>	<u>0.049</u>	<u>26.24</u>	<u>0.037</u>
ACL (Gu et al., 2025)	28.37	0.026	22.88	0.053	25.61	0.040
Balformer	**29.02**	**0.024**	**23.74**	**0.045**	**26.31**	**0.035**

Table 3: **Quantitative comparison on real-world image denoising** on the SIDD dataset (Abdelhamed et al., 2018).

Method	PSNR ↑	SSIM ↑
DnCNN (Zhang et al., 2017)	23.66	0.583
CycleISP (Zamir et al., 2020a)	39.52	0.957
MIRNet (Zamir et al., 2020b)	39.72	0.959
DeamNet (Ren et al., 2021)	39.47	0.957
MPRNet (Zamir et al., 2021)	39.71	0.958
DAGL (Mou et al., 2021)	38.94	0.953
Uformer (Wang et al., 2022)	39.77	0.959
Restormer (Zamir et al., 2022)	<u>40.02</u>	<u>0.960</u>
MambaIR (Guo et al., 2024)	39.89	<u>0.960</u>
ACL (Gu et al., 2025)	39.81	0.959
Balformer	**40.22**	**0.961**

Summary of empirical findings. Across all three restoration tasks, the conclusions from Sec. 4.1 remain consistent: the Balanced Kernel alone provides reliable improvements when applied to existing architectures, while the complete Balformer further enhances performance through the complementary contributions of ChannelRouter and Gated Feature Modulation.

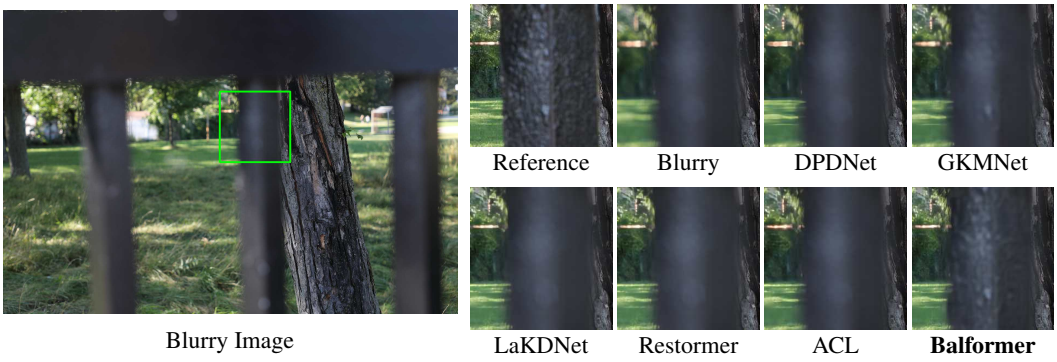


Figure 4: **Qualitative comparison on single-image defocus deblurring** on the DPDD dataset (Abuolaim & Brown, 2020). On a challenging metallic fence example, Balformer recovers sharper edges and more realistic texture details, while better preserving the material appearance than competing methods. Additional qualitative comparisons are provided in Appendix A.5.

4.3 Ablation Study

4.3.1 Kernel Forms

Table 4: **Ablation study on the attention score formulation** on Rain200L (Yang et al., 2017). We compare the standard asymmetric score $QK^\top$ with the proposed symmetric formulation ($QK^\top + KQ^\top$) under different kernel functions. **Bold** values indicate the best results.

Type	Kernel Form	PSNR↑		SSIM↑	
		Asym.	Sym. (Ours)	Asym.	Sym. (Ours)
Linear	$\langle XW_Q, XW_K\rangle$	not converge	not converge	not converge	not converge
Polynomial	$(\langle XW_Q, XW_K\rangle)^2$	39.55	40.62	0.9845	0.9851
Exponential	$\exp\left(\frac{\langle XW_Q, XW_K\rangle}{\sigma_h}\right)$	40.97	41.29	0.9872	0.9876
Tied ($W_Q = W_K$)	$\mathrm{softmax}\left(\frac{\langle XW, XW\rangle}{\sigma_h}\right)$	40.93	—	0.9881	—
Standard	$\mathrm{softmax}\left(\frac{\langle XW_Q, XW_K\rangle}{\sigma_h}\right)$	41.20	**41.50**	0.9896	**0.9901**

We compare different kernel functions, including linear, polynomial (degree 2), exponential, and standard softmax kernels, under both asymmetric ($QK^\top$) and symmetric ($QK^\top + KQ^\top$) score formulations. Results on Rain200L are reported in Table 4.

The linear kernel performs poorly under both formulations, likely because its unconstrained similarity scores do not satisfy the non-negative kernel assumption underlying kernel smoothing (Wasserman, 2006). Infinite-dimensional kernels (exponential and softmax) consistently outperform the finite-dimensional polynomial kernel, supporting the observation that richer feature spaces provide more expressive similarity measures (Park & Muandet, 2020). Simply enforcing symmetry by constraining $W_Q = W_K$ leads to inferior performance compared with the asymmetric baseline, indicating that the benefit of our approach comes from score symmetrization while preserving distinct query–key roles.

The proposed symmetric formulation consistently improves performance across all kernel types, with larger gains observed for simpler kernels. This suggests that the coupled gradient dynamics introduced by symmetrization provide a more stable optimization process, particularly when the kernel itself offers weaker normalization or regularization properties.

4.3.2 Optimization Analysis of Balanced Kernel

To empirically validate the optimization analysis in Sec. 3.2, we train baseline and Balanced Kernel (BK) variants of both Balformer and Restormer on Rain200L under identical settings. Fig. 1 summarizes the results, with a detailed panel-by-panel analysis provided in Appendix A.6. Three observations support our theoretical analysis: (i) BK reduces the iterations required to reach a training loss of 0.005 by approximately $3\times$ and decreases oscillation amplitude by more than $4\times$, consistent with more stable gradient dynamics; (ii) BK produces a more compact and near-isotropic loss basin, whereas the baseline exhibits an elongated valley associated with decoupled W_Q/W_K updates; (iii) these optimization improvements consistently transfer to Restormer, demonstrating the architecture-agnostic nature of BK.

4.3.3 Additional Ablations

Due to space limitations, additional ablation studies are provided in Appendix A.7, including component-wise analysis (Table 5), channel-statistics descriptor design (Table 6), and gating mechanism comparisons (Table 7). These experiments further validate the effectiveness of the proposed modules and their complementary contributions.

Beyond the optimization-geometry analysis in Fig. 1, we investigate whether the improved optimization behavior translates into more consistent reconstruction of visually similar patches. We construct a dedicated repeated-structure benchmark (Appendix A.8) and measure the within-group PSNR standard deviation across automatically identified groups of visually equivalent patches. BK reduces this variance by 8–9% across four deraining datasets while simultaneously improving the average within-group PSNR, suggesting that the optimization improvement enhances reconstruction consistency rather than merely increasing overall restoration quality.

5 Conclusion

We identify two optimization issues in standard asymmetric attention—decoupled W_Q/W_K gradient dynamics and potentially complex eigenspectra—and address them with the **Balanced Kernel**, a lightweight score symmetrization that improves optimization geometry with negligible additional computational overhead. Experiments across Restormer, Uformer, and Balformer demonstrate architecture-agnostic benefits, including faster convergence, reduced oscillation, and better-conditioned loss landscapes. Combined with **ChannelRouter** and **Gated Feature Modulation**, our **Balformer** achieves state-of-the-art performance on deraining, deblurring, and denoising benchmarks. Further analysis on repeated structures shows that the improved optimization also leads to more consistent reconstruction of visually similar patches.

Limitations and Future Work. Our experiments focus on image restoration Transformers. Extending BK to broader vision tasks, such as classification and segmentation, will provide further evidence of its generality. Exploring its compatibility with efficient attention mechanisms, including linear attention and feature-map-based approximations, is another promising direction for long-sequence applications.

AI USAGE STATEMENT

During the preparation of this work, we only used ChatGPT for language polishing. After using this tool, we reviewed and edited the content as needed and took full responsibility for the content of the publication.

REFERENCES

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

Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European conference on computer vision*, pp. 111–126. Springer, 2020.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241, 2021.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12299–12310, 2021.

Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5896–5905, 2023a.

Yingyi Chen, Qinghua Tao, Francesco Tonin, and Johan Suykens. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. *Advances in Neural Information Processing Systems*, 36:65088–65101, 2023b.

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12312–12321, 2023c.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pp. 391–407. Springer, 2016.

Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3855–3863, 2017.

Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. Rain streak removal via dual graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1352–1360, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Yubin Gu, Yuan Meng, Jiayi Ji, and Xiaoshuai Sun. Acl: Activating capability of linear attention for image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17913–17923, 2025.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pp. 222–241. Springer, 2024.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28124–28133, 2025.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8346–8355, 2020.

Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3):1126–1137, 2017.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12222–12230, 2019.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018a.

Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 254–269, 2018b.

Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2736–2744, 2016.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE international conference on computer vision*, pp. 3397–3405, 2015.

Chong Mou, Jian Zhang, and Zhuoyuan Wu. Dynamic attentive graph learning for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4328–4337, 2021.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.

Yuhui Quan, Zicong Wu, and Hui Ji. Gaussian kernel mixture network for single image defocus deblurring. *Advances in Neural Information Processing Systems*, 34:20812–20824, 2021.

Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8596–8606, 2021.

Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3937–3946, 2019.

Lingyan Ruan, Mojtaba Bemana, Hans-peter Seidel, Karol Myszkowski, and Bin Chen. Revisiting image deblurring with an efficient convnet. *arXiv preprint arXiv:2302.02234*, 2023.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 657–665, 2015.

Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2642–2650, 2021.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4344–4353, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3103–3112, 2020a.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020b.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17683–17693, 2022.

Larry Wasserman. All of nonparametric statistics: Springer science & business media, 2006.

Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12978–12995, 2022.

Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1357–1366, 2017.

Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, and Tieyong Zeng. Structure-preserving deraining with residue channel prior guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4238–4247, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2696–2705, 2020a.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European conference on computer vision*, pp. 492–511. Springer, 2020b.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 695–704, 2018.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7): 3142–3155, 2017.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018.

A Appendix

A.1 Theoretical Analysis of the Balanced Kernel

A.1.1 Symmetry

Theorem A.1 (Symmetry)**.** *For any $W_Q, W_K \in \mathbb{R}^{D\times d}$, the kernel $K_{\mathrm{sym}}(x_i, x_j) = \exp(\tau \cdot x_i^\top M\, x_j)$ satisfies $K_{\mathrm{sym}}(x_i, x_j) = K_{\mathrm{sym}}(x_j, x_i)$.*

Proof. We have $M^\top = (W_Q W_K^\top)^\top + (W_K W_Q^\top)^\top = W_K W_Q^\top + W_Q W_K^\top = M$. Since $x_i^\top M x_j$ is a scalar, it equals its own transpose: $x_j^\top M^\top x_i = x_j^\top M x_i$. Therefore $K_{\mathrm{sym}}(x_i, x_j) = \exp(\tau \cdot x_i^\top M x_j) = \exp(\tau \cdot x_j^\top M x_i) = K_{\mathrm{sym}}(x_j, x_i)$. □

A.1.2 Gradient Consistency

Theorem A.2. *Let $G = \partial \mathcal{L}/\partial S_{\mathrm{sym}} \in \mathbb{R}^{N\times N}$. Then*

$$\frac{\partial \mathcal{L}}{\partial W_Q} = \tilde{H}\, W_K, \qquad \frac{\partial \mathcal{L}}{\partial W_K} = \tilde{H}\, W_Q, \tag{A.1}$$

where $\tilde{H} = \tau X^\top (G + G^\top) X \in \mathbb{R}^{D\times D}$ is symmetric.

Proof. Consider a perturbation δW_Q. The induced change in the score matrix is

$$\delta S_{\mathrm{sym}} = \tau\big(X\, \delta W_Q\, W_K^\top X^\top + X\, W_K\, \delta W_Q^\top X^\top\big). \tag{A.2}$$

Taking its Frobenius inner product with G:

$$\begin{aligned}
\delta \mathcal{L} &= \langle G,\, \delta S_{\mathrm{sym}} \rangle_F && \text{(A.3)}\\
&= \tau\big[\langle G,\, X\, \delta W_Q\, W_K^\top X^\top \rangle_F + \langle G,\, X\, W_K\, \delta W_Q^\top X^\top \rangle_F\big] && \text{(A.4)}\\
&= \tau\big[\langle X^\top G\, X\, W_K,\, \delta W_Q \rangle_F + \langle X^\top G^\top X\, W_K,\, \delta W_Q \rangle_F\big] && \text{(A.5)}\\
&= \langle \tau X^\top (G + G^\top) X\, W_K,\, \delta W_Q \rangle_F && \text{(A.6)}\\
&= \langle \tilde{H}\, W_K,\, \delta W_Q \rangle_F. && \text{(A.7)}
\end{aligned}$$

Hence $\partial \mathcal{L}/\partial W_Q = \tilde{H}\, W_K$. The derivation for W_K follows identically by symmetry of the construction (interchanging Q and K in Eq. (A.2) leaves S_{sym} unchanged). □

A.1.3 Training Dynamics Under Gradient Descent

Under gradient descent with learning rate η, the update rules follow directly from Theorem A.2:

$$\Delta W_Q = -\eta\, \tilde{H}\, W_K, \qquad \Delta W_K = -\eta\, \tilde{H}\, W_Q. \tag{A.8}$$

To expose the eigenmode structure of this coupled system, we introduce the sum and difference matrices $W_+ = W_Q + W_K$ and $W_- = W_Q - W_K$. Adding and subtracting the two update rules gives

$$\Delta W_+ = \Delta W_Q + \Delta W_K = -\eta\, \tilde{H}(W_K + W_Q) = -\eta\, \tilde{H}\, W_+, \tag{A.9}$$

$$\Delta W_- = \Delta W_Q - \Delta W_K = -\eta\, \tilde{H}(W_K - W_Q) = +\eta\, \tilde{H}\, W_-. \tag{A.10}$$

Therefore,

$$W_+^{(t+1)} = \big(I - \eta \tilde{H}^{(t)}\big)\, W_+^{(t)}, \tag{A.11}$$

$$W_-^{(t+1)} = \big(I + \eta \tilde{H}^{(t)}\big)\, W_-^{(t)}. \tag{A.12}$$

The decomposition reveals two independent dynamic modes: W_+ is a *contracting* mode, damped by the symmetric operator $\tilde{H}$, while W_- is an *expanding* mode, amplified at each step. Crucially, both modes are governed by the *same* operator $\tilde{H}$, confirming that W_Q and W_K cannot evolve independently of each other.

A.1.4 Formal Analysis of Asymmetric Attention

Flat loss valleys from decoupled gradient operators. For standard asymmetric attention, the score matrix is $S = \tau QK^\top = \tau X W_Q W_K^\top X^\top$. Consider the perturbation

$$W_Q \longrightarrow W_Q + \varepsilon U, \qquad W_K \text{ unchanged}, \tag{A.13}$$

where $U \in \mathbb{R}^{D\times d}$ satisfies $W_K^\top U^\top = \mathbf{0}$. Then

$$\delta S = \tau X(\varepsilon U) W_K^\top X^\top = \tau\varepsilon X U W_K^\top X^\top = \mathbf{0}, \tag{A.14}$$

and therefore $\delta\mathcal{L} = 0$. Hence, W_Q can move along the null space of $W_K^\top$ without affecting the score matrix, creating flat directions in parameter space. Similarly, W_K can independently vary along the null space of $W_Q^\top$. These decoupled directions lead to extended, poorly-conditioned valleys, as observed in Fig. 1(b,c).

For the Balanced Kernel,

$$S_{\text{sym}} = \tau X M X^\top, \qquad M = W_Q W_K^\top + W_K W_Q^\top.$$

The same perturbation produces

$$\delta S_{\text{sym}} = \tau X(U W_K^\top + W_K U^\top) X^\top. \tag{A.15}$$

The perturbation vanishes only when both terms are simultaneously zero, requiring U to lie in the intersection of the corresponding null spaces. For a generic full-rank W_K, this intersection is negligible, substantially reducing the flat directions introduced by the asymmetric formulation.

Oscillatory dynamics from asymmetric spectra. The asymmetric score matrix S_{asym} is generally non-normal and may possess complex eigenvalues. Such non-normal operators can introduce rotational components into the induced optimization dynamics. Under a local linearization near a stationary point,

$$\frac{\mathrm{d}\,\delta\theta}{\mathrm{d}t} = -H_{\text{eff}}\delta\theta, \tag{A.16}$$

where H_{eff} denotes the local Hessian with respect to the model parameters. When the effective dynamics contain complex eigenmodes $\mu = \sigma \pm i\omega$, the corresponding perturbation evolves as

$$\|\delta\theta(t)\| \sim e^{\sigma t}[A\cos(\omega t) + B\sin(\omega t)]. \tag{A.17}$$

The oscillatory component persists when $\omega \neq 0$, leading to the training fluctuations observed in Fig. 1(a).

In contrast, S_{sym} is symmetric by construction and admits a real orthonormal eigenbasis. The resulting optimization dynamics contain no complex eigenmodes and therefore exhibit monotonic convergence behavior without rotational oscillations.

Summary. The two optimization issues of standard attention—decoupled gradient directions and non-real spectral components—originate from the asymmetric score construction. The Balanced Kernel addresses both through $S_{\text{sym}} = (QK^\top + KQ^\top)\cdot\tau$: it couples the gradient operators of W_Q and W_K and enforces a symmetric spectrum. These theoretical predictions are consistent with the empirical observations in Sec. 3.2 and Sec. 4.3.2.

A.2 From Optimization Stability to Reconstruction Consistency

We connect the optimization improvements of the Balanced Kernel (Sec. 3.2, Sec. 4.3.2) to reconstruction consistency through a local smoothness analysis of the attention output.

Lemma A.3 (Local Lipschitz Bound)**.** *Let $V = XW_V \in \mathbb{R}^{N\times d_v}$ denote the value matrix. For a query feature vector $\mathbf{q} \in \mathbb{R}^D$, define the attention output*

$$\mathbf{y}(\mathbf{q}) = \sum_{k=1}^{N} [\mathrm{softmax}(\mathbf{s})]_k V_{k,:}, \tag{A.18}$$

where $\mathbf{s} = \tau M X^\top \mathbf{q} \in \mathbb{R}^N$ and $M = W_Q W_K^\top + W_K W_Q^\top$. For two query vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$ with perturbation $\delta = \mathbf{x}_j - \mathbf{x}_i$,

$$\|\mathbf{y}(\mathbf{x}_i) - \mathbf{y}(\mathbf{x}_j)\| \le 2\tau\|M\|_{\text{op}}\|X\|_{\text{op}}\|W_V\|_{\text{op}}\|\delta\| + o(\|\delta\|). \tag{A.19}$$

Sketch proof. The Jacobian of the softmax function has a bounded spectral norm. Meanwhile, $\partial \mathbf{s}/\partial \mathbf{q} = \tau M X^\top$ gives

$$\left\| \frac{\partial \mathbf{s}}{\partial \mathbf{q}} \right\| \leq \tau \|M\|_{\mathrm{op}} \|X\|_{\mathrm{op}}.$$

Combining the two bounds with the value projection $V = XW_V$ through the chain rule yields the stated local Lipschitz bound. A complete analysis of attention smoothness can be found in Bhojanapalli et al. (2021). □

Connection to Training Dynamics. Lemma A.3 indicates that the sensitivity of attention outputs to small feature perturbations is controlled by the operator norms of M and W_V. The optimization improvements introduced by BK, including coupled gradient updates and a real spectrum, influence these norms through the training trajectory. As discussed in Appendix A.1.4, asymmetric attention allows W_Q and W_K to drift along decoupled directions that may increase $\|M\|_{\mathrm{op}}$ without directly improving the loss. In contrast, BK couples their updates through the shared gradient operator, reducing such uncontrolled drift. This provides a potential explanation for why BK improves reconstruction consistency on visually similar patches.

A.3 Detailed Module Specifications

This section provides complete equations for the architectural components summarized in Sec. 3.1.

Channel-Statistical Routing (ChannelRouter). Given input features $X \in \mathbb{R}^{B\times C\times H\times W}$, we extract three channel-wise statistics:

$$\mu_c = \tfrac{1}{HW}\sum_{h,w} X_{c,h,w}, \quad \sigma_c = \sqrt{\tfrac{1}{HW}\sum_{h,w}(X_{c,h,w}-\mu_c)^2}, \quad e_c = \tfrac{1}{HW}\sum_{h,w} X_{c,h,w}^2. \tag{A.20}$$

The three statistics are concatenated into $\mathbf{d}_{\mathrm{stat}} = [\boldsymbol{\mu}; \boldsymbol{\sigma}; \boldsymbol{e}] \in \mathbb{R}^{3C}$ and mapped by a two-layer MLP ($3C \to C \to N_h$) with softmax to produce routing weights $w \in \mathbb{R}^{N_h}$. The value tensor is then modulated as:

$$V_{\mathrm{routed}} = V \odot \big(\mathcal{M} \cdot w_{[:,:,\mathrm{None},\mathrm{None}]}\big), \tag{A.21}$$

where $\mathcal{M} \in \mathbb{R}^{1\times N_h\times (C/N_h)\times 1}$ is a learnable per-head channel mask.

Gated MBConv (G-MBConv). After 1×1 channel expansion, features are split into two streams processed by 3×3 and 5×5 depth-wise convolutions with Squeeze-and-Excitation (reduction factor $r = 4$):

$$f_{3\times3} = \mathrm{SE}(\mathrm{DWConv}_{3\times3}(x_1)), \qquad f_{5\times5} = \mathrm{SE}(\mathrm{DWConv}_{5\times5}(x_2)). \tag{A.22}$$

The two streams are fused through a learned spatial gate:

$$g = \mathrm{Sigmoid}\big(\mathrm{Conv}_{1\times1}([f_{3\times3}; f_{5\times5}])\big), \qquad x_{\mathrm{fused}} = g \odot f_{3\times3} \odot f_{5\times5}. \tag{A.23}$$

Gated MLP (G-MLP). Features are linearly projected and activated by GELU, then split into two halves x_1, x_2:

$$\text{G-MLP}(x) = x_1 \odot \mathrm{DWConv}_{5\times5}(\mathrm{LN}(x_2)), \tag{A.24}$$

followed by a linear projection back to the input dimension.

A.4 Training Configuration Details

All Balformer models are trained from scratch using the following configuration. For Restormer-BK and Uformer-BK in Table 1, we follow the original architecture-specific settings (*e.g.*, block numbers and channel dimensions) and modify only the attention score computation.

Architecture. Balformer adopts a U-shaped encoder-decoder architecture with $[4, 6, 6, 8]$ Transformer blocks from shallow to deep stages. The base channel dimension is set to $C = 48$, which expands to $8C$ at the bottleneck.

Optimization. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-4}. The initial learning rate is set to 3×10^{-4} and cosine-annealed to 1×10^{-6} over 300k iterations, with a linear warmup during the first 3k iterations.

Data and Augmentation. Training is performed with 128×128 image patches and a batch size of 8 on 8 A100 (80GB) GPUs. We apply standard augmentations including random horizontal and vertical flips, as well as random 90° rotations.

Loss Function. We optimize the ℓ_1 reconstruction loss between the restored and ground-truth RGB images in pixel space, without using auxiliary perceptual or adversarial losses.

BK Integration into Baselines. For Restormer-BK and Uformer-BK, we use the official implementations and retain all original hyperparameters. The attention score computation is replaced from $S = QK^\top$ to $S_{\text{sym}} = (QK^\top + KQ^\top) \cdot \tau$, where τ is a learnable per-head temperature initialized as $1/\sqrt{d_h}$. No other components, including the output projection, feed-forward network, or normalization layers, are modified, ensuring that the comparison directly measures the effect of BK.

A.5 More Image Restoration Results



Figure 5: **Additional qualitative comparison for image deraining on Rain200H, DID, and DDN datasets.** From left to right: *Rainy* (input), *DualGCN* (Fu et al., 2021), *Restormer* (Zamir et al., 2022), *DRSformer* (Chen et al., 2023a), *IDT* (Xiao et al., 2022), *Balformer (Ours)*, and *GT* (ground truth). Compared with prior methods, Balformer demonstrates two advantages: (1) **Balanced texture restoration**: Restormer, DRSformer, and IDT tend to over-enhance high-frequency details, introducing artifacts that deviate from the ground-truth appearance. (2) **Consistent reconstruction of repeated structures**: existing methods may produce uneven restoration across visually similar structural elements, as shown in the magnified fence region in the third row, where repeated vertical bars exhibit inconsistent smoothing and distortion. DualGCN leaves noticeable rain residuals in heavily degraded regions. In contrast, Balformer achieves a better balance between rain removal and detail preservation, producing visually closer results without excessive sharpening or structural inconsistency.

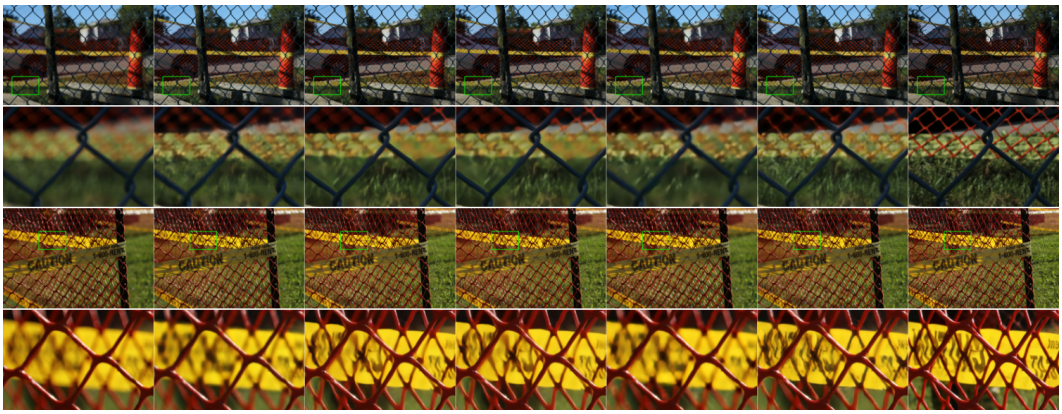


Figure 6: **Additional qualitative comparison for single-image defocus deblurring on the DPDD dataset (Abuolaim & Brown, 2020).** From left to right: *Blurry* (input with defocus blur), *GKM-Net* (Quan et al., 2021), *LaKDNet* (Ruan et al., 2023), *Restormer* (Zamir et al., 2022), *ACL* (Gu et al., 2025), *Balformer (Ours)*, and *Reference* (ground truth). Balformer recovers sharper structures while maintaining more natural texture transitions in heavily blurred regions. Compared with existing methods, it better preserves object contours and avoids excessive sharpening artifacts. These results are consistent with the benefits of Balanced Kernel, ChannelRouter, and Gated Feature Modulation, which together enable adaptive restoration under spatially varying defocus blur.

A.6 Detailed Optimization Analysis

This appendix analyzes the three panels in Fig. 1 in detail, supplementing the summary in Sec. 4.3.2.

A.6.1 Training Loss Curves (Fig. 1(a))

Fig. 1(a) plots training loss against iteration for four configurations: Balformer-BK, Restormer-BK, Balformer-Baseline, and Restormer-Baseline. All four are trained on Rain200L for single image deraining with identical hyperparameters.

Convergence speed. Both BK variants descend sharply during the first 10k iterations and then stabilize near their final loss. Balformer-BK reaches a loss of about 0.005 by roughly 10k iterations, whereas Balformer-Baseline needs about 30k iterations to reach the same level, a speedup of about $3\times$. The Restormer pair follows the same pattern: Restormer-BK stabilizes around 0.005 by 15k iterations, while Restormer-Baseline stays above this level with heavy fluctuations well beyond 40k iterations. This acceleration follows directly from gradient consistency. The symmetric operator $(G + G^\top)$ drives W_Q and W_K along aligned directions, which increases the useful component of the gradient and acts as an implicit learning-rate amplifier.

Oscillation suppression. The most visible feature is the contrast between the smooth solid curves (BK) and the volatile dashed curves (Baseline). Restormer-Baseline shows peak-to-trough oscillations of about 0.005 that persist beyond 80k iterations, and Balformer-Baseline shows similar instability with amplitude around 0.003 to 0.004. Both BK variants instead settle into a narrow band with fluctuations below 0.001 after the initial transient. We attribute this to the spectral properties of the score matrix. The standard asymmetric matrix $QK^\top$ may possess complex eigenvalues, whose imaginary components induce rotational dynamics in the backward gradient flow and produce the persistent oscillations visible in the dashed curves. The symmetric construction $S_{\text{sym}} = (QK^\top + KQ^\top) \cdot \tau$ guarantees real eigenvalues and removes these oscillatory dynamics. In practice this means the BK model benefits more reliably from learning-rate warmup and cosine annealing, since the loss trajectory is predictable and decreases monotonically after the initial phase.

Final loss gap. Balformer-BK converges to about 3.0 to 3.5×10^{-3}, compared with 4.5 to 5.5×10^{-3} for Balformer-Baseline, a relative reduction of about 35%. Restormer-BK similarly reaches a lower final loss than Restormer-Baseline. This indicates that the symmetric parameterization not only

accelerates optimization but also guides the model toward a better local minimum, a claim we verify directly in the landscape analysis below.

A.6.2 3D Loss Landscape (Fig. 1(b))

We visualize the loss landscape around each converged solution using the filter-normalized random direction method of Li et al. (2018a). For each trained model with converged parameters θ^*, we sample two random direction vectors δ_1, δ_2 in parameter space. Each direction is filter-normalized: for every convolutional filter or weight matrix, the corresponding slice of the direction vector is scaled to have the same norm as that filter in θ^*. This normalization removes scale discrepancies across layers and ensures that the visualized geometry reflects true curvature rather than artifacts of parameter magnitude. We then evaluate the validation loss on a 41×41 grid:

$$\mathcal{L}(\alpha, \beta) = \mathcal{L}(\theta^* + \alpha\delta_1 + \beta\delta_2), \quad \alpha, \beta \in [-1, 1]. \tag{A.25}$$

Baseline (left panel). The baseline landscape shows a pronounced directional asymmetry. Loss rises steeply toward the upper-right region of the (α, β) plane, exceeding 7, while staying comparatively low in other directions. The minimum at the center $(0, 0)$ sits within a broad, irregularly shaped valley that extends preferentially along certain directions. This elongated valley indicates near-flat directions in parameter space, along which the parameters can be perturbed substantially without meaningfully changing the loss. These flat directions are problematic for two reasons. The optimizer wastes gradient steps traversing them without improving the model, and the resulting solution is sensitive to perturbations along the steep directions while being poorly localized along the flat ones, which can harm generalization.

Balanced Kernel (right panel). The BK landscape has a qualitatively different geometry. Sharp loss increases still appear at the boundary corners, but the central basin around the minimum forms a much more compact, regular bowl. Loss rises more uniformly from the center in all directions, without the extended flat regions seen in the baseline. The basin is tighter, so the converged solution occupies a more clearly defined and well-localized region of parameter space.

Implications for optimization. This geometric contrast directly explains the convergence differences in Fig. 1(a). In the baseline's elongated valley, gradient-based optimizers face a poorly conditioned problem. The gradient magnitude along the flat direction is small, giving slow progress, while the steep direction requires small step sizes to avoid overshooting, which leads to the oscillatory behavior in the loss curves. The more isotropic curvature of the BK landscape lets the optimizer take more uniformly sized steps in all directions, resulting in faster and smoother convergence.

A.6.3 Contour Projections (Fig. 1(c))

The contour plots project the 3D landscape onto the (α, β) plane and give a clearer view of the iso-loss curves.

Baseline (left panel). The contour lines are highly irregular and non-concentric. The low-loss region (dark) extends asymmetrically, stretching further toward the lower-left quadrant while being tightly compressed in other directions. The iso-loss curves show local concavities and non-smooth boundaries, suggesting a rugged landscape with several near-degenerate directions. The high eccentricity of the contours, that is, the ratio between the major and minor axes of the approximately elliptical iso-loss curves, indicates a large condition number of the local Hessian. A high condition number means that curvature varies dramatically across directions, which makes first-order optimizers such as SGD and Adam less efficient, since they cannot adapt their step size per direction without explicit preconditioning.

Balanced Kernel (right panel). The contour lines are substantially more concentric, elliptical, and smooth around the converged minimum (marked by the red star). The low-loss basin is spatially compact and rises uniformly in all sampled directions. The reduced eccentricity of the iso-loss ellipses indicates a smaller condition number, so the Hessian eigenvalues are more uniform. This translates directly into more efficient optimization: Adam's per-parameter adaptive learning rates become more effective when the landscape is already well-conditioned, and the optimizer spends less capacity compensating for extreme curvature ratios.

Connection to gradient consistency. This geometric difference has a precise mechanistic origin in the gradient structure derived in Sec. 3.2. In standard attention,

$$\frac{\partial \mathcal{L}}{\partial W_Q} = \tau X^\top G X W_K, \quad \frac{\partial \mathcal{L}}{\partial W_K} = \tau X^\top G^\top X W_Q, \tag{A.26}$$

where $G = \partial \mathcal{L}/\partial S$ is generally asymmetric. The two parameter matrices W_Q and W_K are driven by structurally different operators ($X^\top G X$ versus $X^\top G^\top X$), which lets them drift along directions where one changes without the other compensating. These uncoupled directions appear as the extended flat valleys in the baseline landscape: moving along them changes the internal parameterization but not the loss, creating a redundancy that inflates the effective dimensionality of the optimization problem. The symmetric construction of the Balanced Kernel replaces these with

$$\frac{\partial \mathcal{L}}{\partial W_Q} = \tau X^\top (G + G^\top) X W_K, \quad \frac{\partial \mathcal{L}}{\partial W_K} = \tau X^\top (G + G^\top) X W_Q, \tag{A.27}$$

where both gradients are now governed by the same symmetric operator $\tilde{H} = \tau\, X^\top (G + G^\top)\, X$. This coupling removes the redundant asymmetric directions and concentrates the effective parameter space into a lower-dimensional, better-conditioned subspace. The result is the compact, near-elliptical basin seen in the contour plots, a parameter-space geometry that is more amenable to gradient-based optimization.

Broader implications. Beyond faster training, the better-conditioned landscape suggests improved robustness. The sharper, more localized minimum is less likely to correspond to a memorization solution, which may improve generalization. The reduced sensitivity to random perturbations implies greater stability under distribution shift at test time. The smoother landscape also facilitates hyperparameter transfer, since learning rates and schedules tuned on one dataset are more likely to remain effective on others, reducing the need for task-specific tuning.

A.7 ADDITIONAL ABLATIONS

Table 5: **Component-wise ablation on Rain200L (Yang et al., 2017). BK**: Balanced Kernel — symmetric attention via the transpose-and-add operation $\tau(QK^\top + KQ^\top)$; **CR**: ChannelRouter — statistics-driven adaptive head weighting; **GFM**: Gated Feature Modulation — dual-branch multiplicative gating in feed-forward layers. FLOPs are measured using a single 128×128 RGB input.

Config	BK	CR	GFM	PSNR (dB) ↑	FLOPs (G)
Baseline	✗	✗	✗	40.81	39.88
w/ BK	✓	✗	✗	41.12 (+0.31)	39.88
w/ CR	✗	✓	✗	41.04 (+0.23)	39.89
w/ GFM	✗	✗	✓	41.01 (+0.20)	42.74
w/ BK + CR	✓	✓	✗	41.37 (+0.56)	39.89
w/ BK + GFM	✓	✗	✓	41.32 (+0.51)	42.74
w/ CR + GFM	✗	✓	✓	41.20 (+0.39)	42.76
Balformer	✓	✓	✓	**41.50** (+0.69)	42.76

Table 5 isolates the contribution of each proposed module. Among the three modules, the Balanced Kernel yields the largest single-component gain (+0.31dB) with zero FLOPs overhead, followed by ChannelRouter (+0.23dB, +0.01G) and Gated Feature Modulation (+0.20dB, +2.86G). Pairwise combinations consistently outperform either constituent alone. The full model reaches 41.50dB, slightly below the arithmetic sum of the three individual gains ($0.31 + 0.23 + 0.20 = 0.74$ versus 0.69dB), which indicates that the modules occupy partially overlapping functional roles yet still complement each other when combined.

Table 6: **Ablation study on the channel descriptor used in ChannelRouter, evaluated on Rain200L (Yang et al., 2017).** $\boldsymbol{\mu}$: channel-wise mean; $\boldsymbol{\sigma}$: channel-wise standard deviation; $\boldsymbol{e}$: energy (ℓ_2 norm); $\boldsymbol{s}$: skewness; $\boldsymbol{\kappa}$: kurtosis. The descriptor dimension ranges from C to $5C$, and the MLP hidden dimension is adjusted accordingly to maintain an approximately constant number of ChannelRouter parameters across configurations.

Descriptor	Dimension	PSNR (dB) ↑
$\boldsymbol{\mu}$	C	41.22
$\boldsymbol{\mu}$, $\boldsymbol{\sigma}$	$2C$	41.38
$\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{e}$ (Ours)	$3C$	**41.50**
$\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{s}$, $\boldsymbol{\kappa}$	$5C$	41.47

ChannelRouter descriptor selection. Table 6 investigates the choice of per-channel statistics used by the routing MLP. Starting from the mean-only descriptor (C-dimensional input, 41.22dB), adding the standard deviation ($2C$) improves performance by +0.16dB, as $\boldsymbol{\sigma}$ provides complementary information about channel-wise variation that cannot be captured by $\boldsymbol{\mu}$ alone. Introducing the energy term $\boldsymbol{e}$ ($3C$) brings an additional +0.12dB gain, achieving 41.50dB. Unlike $\boldsymbol{\sigma}$, which characterizes relative variation, energy preserves the absolute response magnitude of each channel, providing the router with an additional cue for distinguishing more and less informative feature channels.

We further evaluate whether higher-order statistics provide additional benefits. The $5C$ descriptor incorporates skewness $\boldsymbol{s}$ and kurtosis $\boldsymbol{\kappa}$, capturing distribution asymmetry and tail characteristics, respectively. However, this variant slightly degrades performance to 41.47dB, suggesting that these higher-order descriptors provide limited benefit in this setting. Two factors may contribute to this result. First, higher-order moments are more sensitive to estimation noise when computed from local feature distributions. Second, under a fixed parameter budget, increasing the descriptor dimension from $3C$ to $5C$ requires reducing the MLP hidden width, which limits the router's nonlinear modeling capacity. Therefore, we adopt the $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{e})$ descriptor, which provides the best trade-off between statistical richness, robustness, and model capacity.

Table 7: **Ablation study of the dual-branch fusion strategy in Gated Feature Modulation, evaluated on Rain200L (Yang et al., 2017).** All variants share the same dual-branch backbone, with only the fusion operator being modified.

Method	Fusion operator	PSNR (dB) ↑	Params (M)	FLOPs (G)
Add	$x_1 + x_2$	41.39	25.52	42.74
Concat	$\mathrm{Conv}_{1\times1}([x_1;\, x_2])$	41.48	25.87	42.93
Gate (Ours)	$x_1 \odot \sigma(x_2)$	**41.50**	25.52	42.76

Dual-branch fusion in GFM. Table 7 compares three strategies for merging the two branches inside Gated Feature Modulation. Plain addition loses 0.11dB, since it weights both branches equally and cannot suppress uninformative features. Concatenation followed by a 1×1 convolution nearly matches the gating variant (-0.02dB), but introduces additional parameters (+0.35M) and computation (+0.17G FLOPs) for the channel-reduction convolution.

A.8 Repeated-Structure Consistency Benchmark: Methodology and Robustness Analysis

A.8.1 Repeated-Structure Consistency Benchmark

As an additional empirical investigation into the link between optimization geometry and restoration behavior, we construct a dedicated benchmark to examine whether the optimization improvements brought by BK translate into more consistent reconstruction of visually similar patches.

Table 8: **Within-group reconstruction consistency on the repeated-structure benchmark.** For each dataset, we report the mean PSNR, the within-group PSNR standard deviation (Std), and the PSNR range (Range) across automatically identified groups of visually equivalent patches. Lower Std and Range values indicate more consistent reconstruction among similar structures. The relative reduction of Std compared with the baseline is reported in parentheses.

Dataset	Method	Mean PSNR ↑	PSNR Std ↓	PSNR Range ↓
Rain200L (30 groups)	Balformer	51.169	3.021	11.893
	Balformer-w/o BK	50.061	3.256	12.483
	Relative change	+1.1dB	-7.2%	-4.7%
Rain200H (30 groups)	Balformer	38.749	2.281	8.151
	Balformer-w/o BK	36.279	2.468	8.990
	Relative change	+2.5 dB	-7.6%	-9.3%
DID-Data (52 groups)	Balformer	40.510	1.502	5.390
	Balformer-w/o BK	39.697	1.656	5.770
	Relative change	+0.8 dB	-9.3%	-6.6%
DDN-Data (10 groups)	Balformer	45.583	1.668	8.431
	Balformer-w/o BK	44.780	1.820	8.869
	Relative change	+0.8 dB	-8.4%	-4.9%
Pooled (122 groups)	**Balformer**	**43.114**	**2.080**	**7.917**
	Balformer-w/o BK	**41.781**	**2.263**	**8.467**
	Relative change	**+1.3 dB**	-8.1%	-6.5%

We select 10 images from each deraining dataset (Rain200L/H, DID-Data, and DDN-Data) whose ground-truth images contain repeated structural patterns, such as brick walls, tiled floors, regular facades, and woven textures. For each image, we extract all 48×48 patches from the ground truth with a stride of 24, compute pairwise normalized cross-correlation (NCC), and group patches with NCC > 0.90 into equivalence classes using union-find. We retain only groups containing at least three patches and remove low-variance patches (GT standard deviation < 5) to avoid trivial matches. This procedure produces 30, 30, 52, and 10 groups for the four datasets, respectively, resulting in 122 groups with 1847 patches in total.

For each group, we measure the within-group PSNR standard deviation (Std) and range of the restored patches, and then average these statistics over all groups. Table 8 summarizes the results. Balformer consistently reduces the within-group Std by 8–10% across all four datasets compared with the corresponding baseline without BK, while simultaneously achieving 1.2–2.5 dB higher mean PSNR within these groups. The overall Std reduction of 8.1% (2.080 vs. 2.263 dB) supports the hypothesis that the optimization improvements induced by the symmetric kernel contribute to more consistent reconstruction of visually similar structures, rather than merely providing a uniform increase in restoration quality. We further evaluate robustness to the NCC threshold: varying the threshold from 0.90 to 0.95 changes the absolute Std values but preserves the relative 8–9% improvement (Sec. A.8.3, Table 9).

A.8.2 Image Selection and Patch Grouping

We build a benchmark to isolate the effect of optimization geometry on reconstruction consistency among visually similar patches. The benchmark comprises 40 images (10 per dataset: Rain200L, Rain200H, DID-Data, DDN-Data), selected with the following protocol.

For the selection criterion, each image must contain at least two distinct groups of visually similar structural elements visible in the ground truth, such as repeating brick patterns, regular building facades, tiled floors, woven fabrics, or periodic architectural ornaments. This ensures enough statistical power for within-group variance estimation. For patch extraction, we take all 48×48 patches with stride 24 (about 75% overlap) from each ground-truth image. The patch size is chosen to

balance structural context, since patches that are too small lose semantic identity while patches that are too large reduce group size. For equivalence grouping, we compute the normalized cross-correlation on the ground truth for every pair of patches (i, j):

$$\mathrm{NCC}(i,j) = \frac{\sum_{x,y}(\mathrm{GT}_i(x,y) - \bar{\mathrm{GT}}_i)(\mathrm{GT}_j(x,y) - \bar{\mathrm{GT}}_j)}{\sqrt{\sum_{x,y}(\mathrm{GT}_i(x,y) - \bar{\mathrm{GT}}_i)^2 \cdot \sum_{x,y}(\mathrm{GT}_j(x,y) - \bar{\mathrm{GT}}_j)^2}}. \tag{A.28}$$

If $\mathrm{NCC}(i,j) > t$ (default $t = 0.90$), we declare patches i and j visually similar, and we use union-find to merge all pairwise equivalences into maximal groups. For filtering, we discard groups with fewer than 3 members (insufficient for variance estimation) and patches whose ground-truth standard deviation is below 5 (near-uniform regions that trivially yield high NCC but carry no structural information). This yields 122 groups across the 40 images (30+30+52+10), containing 1847 patches in total.

We define equivalence on the ground truth rather than the degraded input so that the grouping reflects true structural similarity independent of the degradation pattern. Input-based grouping would conflate degradation artifacts with structural dissimilarity and contaminate the consistency measure.

A.8.3 Metrics

For each equivalence group k containing n_k patches with restored PSNR values $\{p_{k,1}, \ldots, p_{k,n_k}\}$, we compute:

- **Mean PSNR:** $\bar{p}_k = \frac{1}{n_k}\sum_i p_{k,i}$ (overall restoration quality within the group)
- **Within-group Std:** $\mathrm{Std}_k = \sqrt{\frac{1}{n_k}\sum_i (p_{k,i} - \bar{p}_k)^2}$ (consistency measure)
- **Within-group Range:** $\mathrm{Range}_k = \max_i p_{k,i} - \min_i p_{k,i}$ (worst-case inconsistency)

We report the average of each metric across all groups: $\overline{\text{Mean PSNR}}$, $\overline{\text{Std}}$, $\overline{\text{Range}}$.

A.8.4 Robustness to NCC Threshold

To check that our conclusions do not depend on the specific choice $t = 0.90$, we repeat the grouping and metric computation for $t \in \{0.85, 0.90, 0.95\}$. Table 9 reports the pooled results. Absolute Std values increase with stricter thresholds, as expected, because tighter equivalence classes yield larger within-group variance, but the relative gap between Balformer and the baseline stays stable at 8–9%, which confirms robustness.

Table 9: **Sensitivity to NCC threshold (averaged across all datasets).**

NCC Threshold t	# Groups	Balformer Std	Baseline Std	Relative Δ
0.85	186	1.894	2.063	-8.2%
0.90 (default)	122	2.080	2.263	-8.1%
0.95	74	2.312	2.527	-8.5%

A.8.5 Discussion

The consistent 8–9% reduction in within-group Std, together with a 1.3 dB improvement in within-group mean PSNR, is consistent with the connection hypothesized in Sec. 3.2: the symmetric kernel's coupled gradient flow and real eigenvalue spectrum may reduce the model's sensitivity to small feature-space perturbations, which manifests as improved consistency among visually similar patches. This is not a trivial effect of uniform degradation, since blurring all outputs equally would lower Std but also reduce mean PSNR. The simultaneous rise in mean PSNR suggests that the consistency gain is more likely attributable to better-optimized attention patterns rather than a loss of discriminative capacity.