CACRNET: COMPRESSION-AWARE CONTEXT ROUTING FOR JPEG ARTIFACT REMOVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Blind JPEG image restoration is a practical yet challenging task, as compression artifacts are spatially non-uniform and highly entangled with genuine image structures. Existing methods typically rely on either local restoration or global prompts alone. As a result, heavily degraded regions may receive insufficient contextual compensation, while artifact-contaminated shallow features can still leak into the decoder through skip connections. To address these limitations, we propose a Compression-Aware Context Routing Network, a restoration framework that uses local degradation evidence to coordinate both contextual compensation and skip-feature selection. Specifically, we proposed a Spatial-Channel Interaction (SCI) module derives local degradation cues and uses them to retrieve location-specific contextual prompts. These prompts are drawn from a Contextual Codebook Bank, which recursively aggregates multiscale encoder features into a compact set of image-adaptive tokens, and are injected into the decoder through efficient factorized spatial attention. In parallel, we design a Compression-Aware Gated Fusion (CAGF) module to recalibrate skip features before fusion, suppressing artifacts while preserving reliable structural details. Extensive experiments on blind, non-blind, and double JPEG restoration benchmarks demonstrate that CACRNet achieves superior quantitative performance and faithfully restores structural details across different compression settings.

1 INTRODUCTION

JPEG remains pervasive owing to its compact representation and broad hardware support, yet its quantization step discards high-frequency coefficients, introducing blocking artifacts, ringing, and texture attenuation. Early convolutional methods learn a direct mapping from compressed to clean pixels Dong et al. (2015); Zhang et al. (2017), while later networks exploit residual, dense, and attention mechanisms to improve restoration fidelity Zhang et al. (2018a); Liang et al. (2021); Zamir et al. (2022). However, the challenge is not simply the loss of information. JPEG divides an image into transform blocks and quantizes each frequency component differently: block boundaries become visible in flat regions, strong edges give rise to ringing, and stochastic textures lose energy. Consequently, a high-frequency response may constitute a valid detail in one neighborhood and an artifact in another. A restoration model must therefore identify where its reliable evidence lies.

Existing approaches address this problem from different perspectives. Quality-aware methods estimate quality factors, degradation ranks, or compression-aware features to guide restoration Jiang et al. (2021); Li et al. (2020); Liu et al. (2026). Codec-aware methods exploit JPEG priors, including DCT coefficients, quantization tables, and chroma information, to compensate for quantization and color distortions Zhang et al. (2018a); Fu et al. (2021b); Yang et al. (2023); Liu et al. (2025). Prompt-based methods such as PromptIR and PromptCIR encode degradation information into learned prompts Potlapalli et al. (2023); Li et al. (2024). As shown in Figure 1(a), these methods generate input-adaptive guidance by dynamically combining learned prompt components from image features. This improves adaptability, but their context comes primarily from learned prompt bases rather than structural references retrieved from the current image using local compression evidence. Recent sequence and universal restoration models employ frequency-aware Mamba scanning or codec-aware conditioning across compression formats Wang et al. (2025); Han et al. (2026). Despite these advances, degradation guidance is often represented by global or codec-level conditions, or remains implicit in feature-conditioned prompt generation. Consequently, local com-

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.

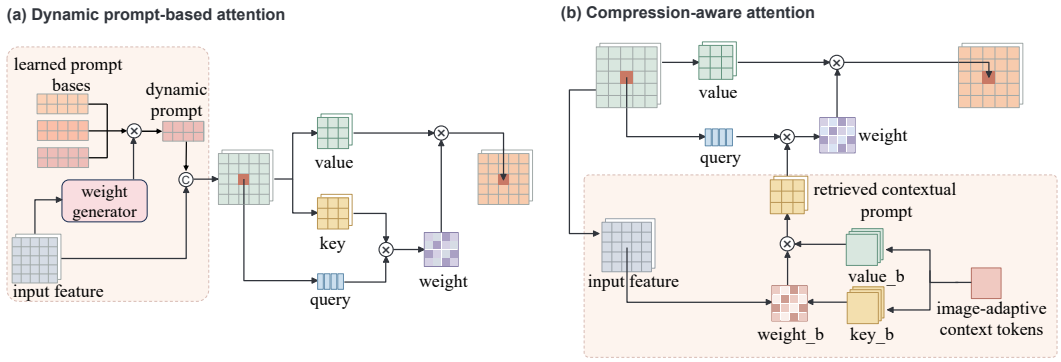


Figure 1: Comparison of (a) Dynamic prompt-based attention, and (b) our compression-aware attention, which retrieves image-specific context according to local degradation cues.

pression evidence is not explicitly associated with non-local references. This limitation is relevant to JPEG restoration, where flat regions, edges, and textures require different restoration behavior. Effective restoration should therefore consider both the compression level and evidence reliability at each spatial location. Uniform global interaction may disturb reliable regions, whereas strictly local operators cannot exploit similar structures elsewhere to recover destroyed patterns. Skip connections introduce a related problem: shallow features contain both genuine details and compression artifacts, which may be passed to the decoder without appropriate filtering.

To address these limitations, we introduce the Compression-Aware Context Routing Network, which uses local degradation evidence to guide both contextual interaction and skip-feature transmission. We introduce Spatial-Channel Interaction to extract location-dependent cues from gradient responses, local variance, and channel-wise feature disagreement. These cues retrieve a distinct contextual prompt for each decoder location from a Contextual Codebook Bank that aggregates multiscale encoder features into image-adaptive tokens. As shown in Figure 1(b), the retrieved prompt provides contextual keys according to the local degradation state, while the decoder feature retains the queries and values. This design introduces reliable non-local information without replacing the restoration content or indiscriminately propagating corrupted responses. In parallel, We design a Compression-Aware Gated Fusion module to recalibrate skip features before fusion, suppressing artifacts while preserving reliable structural details. CACRNet therefore adapts both contextual retrieval and detail transmission to the degradation observed at each location. Our contributions are:

- We propose a Spatial-Channel Interaction module, which replaces global degradation conditioning with local compression cues to retrieve location-specific references from an image-adaptive Contextual Codebook Bank.
- We introduce a Compression-Aware Gated Fusion module to gate spatial and channel-wise cues separately, using local feature disagreement to suppress artifact-contaminated responses while preserving reliable structural detail.
- Extensive experiments on blind, non-blind, and double JPEG restoration demonstrate that CACRNet achieves superior quantitative performance and effectively preserves structural details while suppressing compression artifacts.

2 Related Work

2.1 JPEG Artifact Removal.

Deep learning-based artifact removal has benefited from the strong representation capacity of neural networks. ARCNN first formulated JPEG artifact reduction as an end-to-end convolutional learning problem Dong et al. (2015), while subsequent CNN restorers improved feature representation through persistent memory Tai et al. (2017), residual dense aggregation Zhang et al. (2018b), and residual non-local attention Zhang et al. (2019). However, these architectures do not explicitly model the JPEG formation process. Codec-aware methods therefore incorporate JPEG-specific structures: DDCN, DMCNN, and D3-style approaches exploit DCT-domain priors Zhang et al. (2017; 2018a),

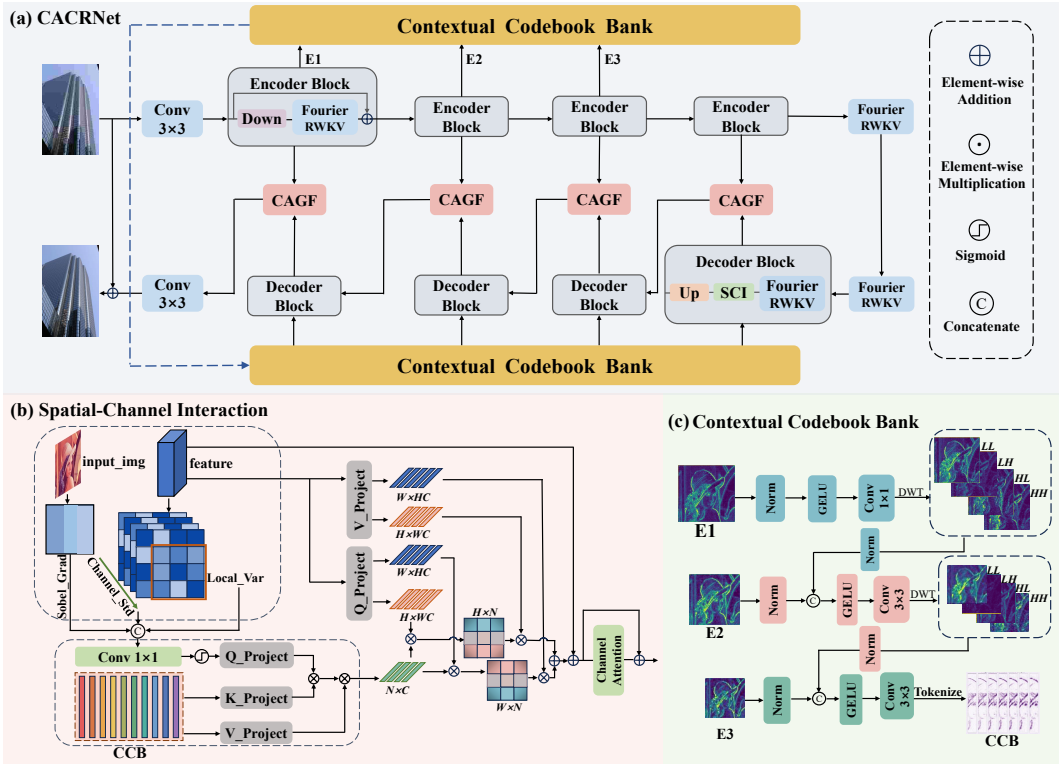


Figure 2: Overall architecture. Encoder features form a multiscale wavelet contextual bank. Local degradation evidence queries the bank, while compression-aware gated fusion regulates encoder–decoder transfer.

followed by dual-prior learning Fu et al. (2021b), model-driven unfolding Fu et al. (2021a), contrastive representation learning Wang et al. (2022a), and dual-domain formulations combining pixel and transform representations Yang et al. (2023). Although these studies demonstrate the value of JPEG priors, quantized coefficients mainly describe local degradation and do not identify cross-scale features that can compensate for missing content. Moreover, methods trained for specific quality factors often generalize poorly to unseen compression settings. Blind restoration methods address this limitation through quality-ranker guidance Wang et al. (2020), wide-QF training Li et al. (2020), and flexible quality-factor prediction Jiang et al. (2021). JDEC estimates a continuous DCT spectrum for joint dequantization and upsampling Han et al. (2024), while CODiff incorporates JPEG priors into diffusion-based restoration Guo et al. (2025b). Nevertheless, prior extraction and restoration may involve conflicting objectives, and most methods still condition restoration on a single image-level descriptor, assigning the same degradation estimate to structurally different regions. In contrast, we estimate degradation evidence at each decoder location to guide location-specific restoration.

2.2 Global Context in Image Restoration

Global context is important for restoration because many degradations cannot be resolved locally. Transformer-based methods capture long-range dependencies through shifted-window attention in SwinIR Liang et al. (2021), hierarchical self-attention in Uformer Wang et al. (2022b), and channel-wise transposed attention in Restormer Zamir et al. (2022). Other approaches employ axial attention Wang et al. (2020), multi-axis aggregation Tu et al. (2022), or stripe and window attention Li et al. (2023), while multiscale networks such as MPRNet Zamir et al. (2021), MIRNet Zamir et al. (2020), HINet Chen et al. (2021), and NAFNet Chen et al. (2022) exploit progressive refinement or cross-scale interaction. Despite their effectiveness, these methods generally derive context from the current feature map without assessing its reliability under compression. This is problematic for JPEG restoration, where genuine structures and artifact-induced responses may be propagated together by global operators. Degradation-conditioned methods offer another solution: AirNet learns degradation representations Li et al. (2022), while PromptIR introduces learnable prompts for unknown degradations Potlapalli et al. (2023). However, such prompts are typically image- or stage-level vari-

ables with limited correspondence to local artifacts. Since flat regions, edges, and textures require different restoration strategies, contextual guidance should remain responsive to spatially varying degradation. We therefore build an image-adaptive Contextual Codebook Bank from multiscale features and retrieve location-specific prompts using local degradation cues. This routes reliable global context while preserving decoder content.

3 METHODOLOGY

In this section, we present the proposed Compression-Aware Context Routing Network for blind JPEG image restoration, consisting of three main components: the Contextual Codebook Bank, the Spatial-Channel Interaction module, and the Compression-Aware Gated Fusion module. We first describe how the CCB recursively aggregates multiscale encoder features into contextual tokens, then present the SCI mechanism for retrieving and injecting location-specific prompts, and finally introduce the CAGF module for degradation-aware skip fusion.

3.1 OVERVIEW

Given a compressed image $I_c \in \mathbb{R}^{3\times H\times W}$, CACRNet follows a U-shaped encoder–decoder architecture and uses Fourier-RWKV blocks Zheng et al. (2025) as the restoration backbone. The encoder extracts multiscale features $\{S_i\}_{i=1}^{4}$. At each decoder stage, the Spatial-Channel Interaction module estimates compression-related cues from gradient responses, local variance, and channel-wise feature discrepancy, and transforms them into a degradation prompt. The Contextual Codebook Bank (CCB) builds an image-specific token bank $B \in \mathbb{R}^{N\times C}$ from the multiscale encoder features, serving as a context repository for decoder-side restoration. SCI uses the degradation prompt to query B, and the retrieved contextual keys are injected into parallel axial attention to regulate the decoder feature in horizontal and vertical directions. At the same time, the Compression-Aware Fusion Gate uses the local variance discrepancy between decoder and encoder features to select and reconstruct skip information before fusion. The network finally predicts a residual image that is added to the compressed input I_c. The CACRNet pipeline is illustrated in Figure 2.

3.2 SPATIAL CHANNEL INTERACTION

At decoder stage l, SCI constructs a degradation-aware query to retrieve contextual prompts from an image-specific Contextual Codebook Bank, whose construction is described in the following subsection. Let $X_l \in \mathbb{R}^{C\times H\times W}$ denote the decoder feature and Y the luminance component of the compressed input. Three complementary cues are extracted:

$$\begin{aligned} G &= \mathrm{Norm}\left(|\nabla_x Y| + |\nabla_y Y|\right), \\ V_{\mathrm{loc}} &= \mathrm{Mean}_c\left(\mathbb{E}_{3\times3}[\bar{X}_l^2] - \mathbb{E}_{3\times3}[\bar{X}_l]^2\right), \\ V_c &= \mathrm{Std}_c(X_l), \end{aligned} \tag{1}$$

where $\bar{X}_l$ is the group-normalized feature and $\mathbb{E}_{3\times3}[\cdot]$ denotes local averaging. These cues describe input edges, local feature activity, and channel disagreement, respectively.

After per-sample normalization, the three maps are concatenated and processed by a lightweight gating network ϕ_l:

$$M_l = \sigma(\phi_l([G, V_{\mathrm{loc}}, V_c])), \qquad Q_l = \mathrm{LN}(X_l \odot M_l), \tag{2}$$

where $M_l \in [0,1]^{C\times H\times W}$ is a channel-wise spatial gate.

Given the contextual bank $B \in \mathbb{R}^{N\times C_B}$, its keys and values are obtained as $K_B = W_K B$ and $V_B = W_V B$. SCI retrieves a location-dependent prompt using cosine-similarity attention:

$$A_l = \mathrm{softmax}\left(\widehat{Q}_l \widehat{K}_B^{\top}\right), \qquad Z_l = A_l V_B, \tag{3}$$

where $N = 256$ and C_B equals the network width. Thus, each spatial location selects a weighted combination of image-specific bank tokens.

The retrieved prompt is injected through parallel horizontal and vertical axial interactions:

$$\Delta_l = \frac{1}{2}\left[\mathcal{A}_h(X_l, Z_l, X_l) + \mathcal{A}_v(X_l, Z_l, X_l)\right]. \tag{4}$$

Here, X_l provides the query and value, while Z_l serves as the contextual key. SCI therefore separates contextual retrieval from feature injection, routing non-local information according to local degradation evidence while retaining the decoder feature as the primary carrier of restored content.

3.3 CONTEXTUAL CODEBOOK BANK

To provide SCI with reliable context, CCB recursively aggregates adjacent encoder stages. Directly pooling each encoder stage into separate tokens may discard fine spatial information before cross-scale alignment, weakening the correspondence between local JPEG artifacts in shallow features and structural cues at coarser scales. We therefore construct the Contextual Codebook Bank by recursively aligning and fusing adjacent encoder stages before tokenization. Starting from the first encoder feature S_1, we apply normalization, nonlinearity, and a projection to reduce the channel dimension to $C/4$. A fixed Haar transform then decomposes each projected channel into four subbands and reduces the spatial resolution by a factor of two:

$$W_1 = \mathcal{R}_1(\mathcal{H}(\phi_1(S_1))), \tag{5}$$

where $\mathcal{H}$ denotes grouped stride-two Haar filtering, $\phi_1(\cdot)$ denotes the normalization–activation–projection operation, and $\mathcal{R}_1(\cdot)$ compresses the expanded subband channels. The output W_1 has the same spatial resolution as the next encoder feature. We concatenate it with the normalized S_2, fuse the two branches with a 3×3 convolution, and apply the same decomposition:

$$W_2 = \mathcal{R}_2(\mathcal{H}(\phi_2([W_1, \mathrm{GN}(S_2)]))), \tag{6}$$

The propagated representation W_2 is then fused with S_3 to obtain the bank feature F_B. This recursive design keeps cross-scale frequency information available. We finally pool F_B to 16×16 and flatten it into a sequence of contextual tokens. The bank is produced separately for each input image. It is thus an image-conditioned contextual bank rather than a learned discrete codebook shared across the training set.

We use fixed Haar filtering to separate the propagated feature into complementary low- and high-frequency subbands. This operation reduces the spatial resolution while retaining explicit high-frequency responses, which are important for describing JPEG artifacts. Group normalization mitigates the scale mismatch between the propagated wavelet branch and the current encoder feature. We exclude the deepest feature S4 from bank construction, since at that resolution local compression artifacts are already entangled with high-level semantics, making the resulting tokens less suitable as frequency-aware contextual references. Since all tokens are generated from the current compressed image, CCB stores no external prototypes; retrieval is self-contained within the input itself.

3.4 COMPRESSION-AWARE GATED FUSION

Skip connections preserve spatial details but may also transmit blocking and ringing artifacts from shallow encoder features. We therefore introduce Compression-Aware Gated Fusion (CAGF) to recalibrate encoder features before skip fusion. The CAGF module is illustrated in Figure 3. Given encoder and decoder features E_l and D_l, we first align and normalize the two branches:

$$\bar{E}_l = \mathrm{GN}(\mathrm{Conv}_{1\times1}(E_l)), \qquad \bar{D}_l = \mathrm{GN}(D_l). \tag{7}$$

We then compute their local variance within a 3×3 window:

$$\begin{aligned} V_E &= \mathbb{E}_{3\times3}[\bar{E}_l^2] - \mathbb{E}_{3\times3}[\bar{E}_l]^2, \\ V_D &= \mathbb{E}_{3\times3}[\bar{D}_l^2] - \mathbb{E}_{3\times3}[\bar{D}_l]^2. \end{aligned} \tag{8}$$

Their normalized signed contrast is

$$\Delta V = \mathrm{clip}\left(\frac{V_E - V_D}{V_E + V_D + \epsilon}, -1, 1\right). \tag{9}$$

This contrast indicates whether the encoder has stronger or weaker local fluctuations than the decoder, providing a cue for skip-feature recalibration. We obtain spatial and channel descriptors by channel averaging and global average pooling, respectively:

$$S_l = \mathrm{Mean}_c(\Delta V), \qquad C_l = \mathrm{GAP}(\Delta V). \tag{10}$$

After the corresponding heads produce logits A_s and A_c, we form an identity-centered gate:

$$W_l = \exp[\gamma \tanh(A_s + A_c)], \tag{11}$$

where γ is a learnable scaling factor. Finally, we fuse the recalibrated encoder feature with the decoder feature:

$$F_l = \mathrm{GELU}(\mathrm{Conv}_{1\times1}([E_l \odot W_l, D_l])). \tag{12}$$

The final layers of both gating heads are zero-initialized, yielding $W_l = 1$ at initialization. CAGF therefore starts from standard skip fusion and progressively learns to preserve reliable details while attenuating artifact-contaminated encoder responses

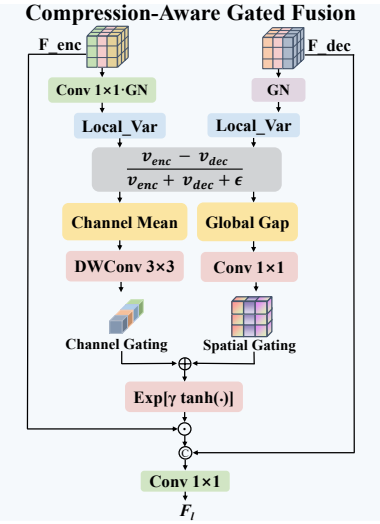


Figure 3: Compression-Aware Gated Fusion module.

3.5 Losses

We optimize the proposed network using a combination of pixel-, frequency-, and gradient-domain supervision. Specifically, the multiscale $\mathcal{L}_{\mathrm{PSNR}}$ is applied to the final prediction and auxiliary decoder outputs as the primary reconstruction objective. $\mathcal{L}_{\mathrm{FFT}}$ minimizes the ℓ_1 distance between the real and imaginary components of the predicted and ground-truth Fourier spectra, while $\mathcal{L}_{\mathrm{grad}}$ further encourages structural consistency by comparing horizontal and vertical gradient responses. The overall objective is defined as

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{PSNR}} + \lambda_f \mathcal{L}_{\mathrm{FFT}} + \lambda_g \mathcal{L}_{\mathrm{grad}}, \tag{13}$$

where λ_f and λ_g control the contributions of the frequency- and gradient-domain constraints, respectively.

4 Experiments

4.1 Experimental Setup

Following the previous work Li et al. (2024), we use the DIV2K and Flickr2K datasets to construct our training set and apply the PIL toolkit to generate JPEG images. During training, the quality factor is randomly sampled from $[10, 70]$, covering a broad range of compression levels. For evaluation, we use the LIVE1, BSDS500, ICB, and DIV2K datasets. In the blind setting, the compression quality is unknown to the network, whereas the non-blind evaluation is conducted at fixed quality factors of 10, 20, 30, and 40. Following the evaluation protocol of Li et al. (2024), we report PSNR, SSIM, and PSNR-B, where PSNR-B is specifically designed to measure the reduction of blocking artifacts.

We additionally evaluate robustness to double JPEG compression following the protocol of FBCNN Jiang et al. (2021). Each clean image is first compressed using quality factor QF_1 and then compressed again using QF_2, where $QF_1, QF_2 \in \{30, 50\}$. We consider both aligned and misaligned double compression. For the aligned setting, the two JPEG block grids share the same origin, corresponding to a spatial shift of 0 pixels. For the misaligned setting, the decoded image is shifted by 4 pixels before the second compression, producing inconsistent block boundaries between the two compression stages.

Similar to the setting of Li et al. (2024), training proceeds in two stages: the network is first pre-trained and subsequently fine-tuned, with quality factors randomly sampled from $[10, 70]$ in both stages. Training images are randomly cropped into paired 128×128 patches, with random flipping and rotation applied for data augmentation. We adopt the Adam optimizer Kingma & Ba (2014), with the learning rate initialized at 2×10^{-4} and decayed to 2×10^{-5} following a cosine annealing schedule over 2000 epochs. We train CACRNet using PyTorch with a batch size of 16. All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24 GB of memory.

4.2 Quantitative Comparison

To comprehensively evaluate the effectiveness of CACRNet, we compare it with seven representative restoration models, including classical CNN-based restorers, quality-aware JPEG-specific methods,

codec-aware dual-domain approaches, and recent general-purpose restoration backbones. All reproducible methods are retrained using their official implementations and our data protocol. Tables 1 and 2 report the results under blind and non-blind settings, respectively.

Table 1: Quantitative comparison for blind CIR. Results are reported in terms of PSNR↑/SSIM↑/PSNR-B↑. Params denotes the number of model parameters in millions. The best performances are **bolded**.

Methods	Venue	Params (M)	LIVE1			BSDS500			ICB			DIV2K		
			PSNR	SSIM	PSNR-B	PSNR	SSIM	PSNR-B	PSNR	SSIM	PSNR-B	PSNR	SSIM	PSNR-B
JPEG	–	–	30.02	0.872	28.18	30.58	0.876	28.29	33.87	0.842	33.05	31.71	0.885	30.02
FBCNN Jiang et al. (2021)	ICCV 2021	71.92	32.08	0.904	31.44	32.41	0.903	31.45	35.90	0.865	36.33	33.96	0.918	33.41
CRLM Wang et al. (2022a)	ECCV 2022	13.90	32.26	0.907	31.70	32.75	0.908	31.78	35.93	0.865	36.35	34.13	0.919	33.62
PromptIR Potlapalli et al. (2023)	NeurIPS 2023	35.59	32.34	0.907	31.75	32.67	0.907	31.67	36.35	0.868	36.75	34.38	0.922	33.81
PromptCIR Li et al. (2024)	CVPR 2024	34.75	32.45	0.908	31.87	32.74	0.907	31.76	36.41	0.870	36.79	34.45	0.922	33.88
MaIR Li et al. (2025)	CVPR 2025	22.67	32.55	0.909	31.96	32.63	0.906	31.65	36.34	0.868	36.71	34.33	0.921	33.77
MambaIRv2 Guo et al. (2025a)	CVPR 2025	36.25	32.63	0.910	32.04	32.62	0.906	31.74	36.24	0.866	36.63	34.48	0.923	33.90
Q-MambaIR Chen et al. (2025)	CVPR 2026	24.08	32.69	0.910	32.07	32.74	0.907	31.77	36.49	0.871	36.85	34.60	0.924	34.01
CACRNet	Ours	23.45	**33.03**	**0.917**	**32.43**	**32.76**	**0.908**	**31.84**	**37.51**	**0.883**	**37.41**	**34.83**	**0.926**	**34.28**

Blind JPEG restoration. Results under the blind setting are reported in Table 1. CACRNet achieves the best overall performance across all four benchmarks. Compared with Q-MambaIR, it improves PSNR by 0.34, 1.02, and 0.23 dB on LIVE1, ICB, and DIV2K, with corresponding PSNR-B gains of 0.36, 0.56, and 0.27 dB. On BSDS500, CACRNet achieves 32.76 dB PSNR and 31.84 dB PSNR-B while matching the best SSIM. Blind restoration requires the model to handle unknown compression levels without an explicit QF input. Instead of relying on a fixed global degradation descriptor, SCI estimates compression evidence at each location and uses it to retrieve suitable image-specific context from CCB. This allows CACRNet to adapt its restoration behavior to both the unknown compression strength and spatially varying artifacts. The pronounced improvement on ICB further suggests that multiscale contextual retrieval is particularly effective for complex structures and textures.

Table 2: Quantitative comparison for non-blind JPEG restoration. The first row of each method reports PSNR↑ | PSNR-B↑, while the second row reports SSIM↑. The best and second-best results are **bolded** and underlined, respectively.

Method	LIVE1				ICB			
	QF10	QF20	QF30	QF40	QF10	QF20	QF30	QF40
JPEG	25.69 | 24.20 0.743	28.06 | 26.49 0.826	29.37 | 27.84 0.861	30.28 | 28.84 0.882	29.44 | 28.53 0.757	32.01 | 31.11 0.806	33.20 | 32.35 0.831	33.95 | 33.14 0.840
FBCNN Jiang et al. (2021)	27.77 | 27.51 0.803	30.11 | 29.70 0.868	31.43 | 30.92 0.897	32.34 | 31.80 0.913	32.18 | 32.15 0.815	34.38 | 34.34 0.844	35.41 | 35.35 0.857	36.02 | 35.95 0.866
CRLM Wang et al. (2022a)	27.80 | 27.57 0.805	30.23 | 29.85 0.872	31.58 | 31.13 0.900	32.53 | 32.04 0.916	32.05 | 32.04 0.813	34.32 | 34.31 0.842	35.37 | 35.35 0.856	35.99 | 35.97 0.860
PromptIR Potlapalli et al. (2023)	27.98 | 27.66 0.806	30.31 | 29.82 0.872	31.66 | 31.07 0.900	32.58 | 31.96 0.915	32.70 | 33.01 0.821	34.84 | 35.21 0.847	35.85 | 36.24 0.860	36.45 | 36.85 0.869
PromptCIR Li et al. (2024)	28.00 | 27.73 0.806	30.32 | 29.86 0.872	31.69 | 31.10 0.901	32.64 | 32.12 0.917	32.73 | **33.12** 0.822	34.89 | 35.24 0.848	35.89 | 36.26 0.861	36.52 | 36.92 0.871
MaIR Li et al. (2025)	28.05 | 27.72 0.802	30.41 | 29.93 0.873	31.73 | 31.14 0.902	32.72 | 32.20 0.917	32.71 | 33.07 0.821	34.87 | 35.17 0.844	35.81 | 36.25 0.860	36.53 | 36.92 0.871
MambaIRv2 Guo et al. (2025a)	28.09 | 27.84 0.804	30.47 | 30.07 0.872	31.83 | 31.23 0.901	32.80 | 32.29 0.918	32.75 | 33.09 **0.822**	34.91 | 35.28 0.847	35.99 | 36.32 0.862	36.67 | 37.02 0.870
Q-MambaIR Chen et al. (2025)	28.11 | 27.77 0.807	30.51 | 30.09 0.874	31.86 | 31.30 0.901	32.78 | 32.29 0.917	32.92 | 33.10 0.820	35.03 | 35.22 0.848	36.23 | 36.41 0.867	36.98 | 37.04 0.877
CACRNet	**28.16 | 27.87** **0.808**	**30.56 | 30.12** **0.874**	**31.94 | 31.41** **0.902**	**32.94 | 32.32** **0.919**	**33.09** | 33.05 0.821	**35.45 | 35.40** **0.854**	**36.62 | 36.54** **0.874**	**37.32 | 37.22** **0.883**

Non-blind JPEG restoration. Results at fixed quality factors are presented in Table 2. CACRNet achieves the highest PSNR across all evaluated settings. On LIVE1, it surpasses the strongest competitors by 0.05, 0.05, 0.08, and 0.14 dB from QF10 to QF40; the corresponding gains on ICB are 0.17, 0.42, 0.39, and 0.34 dB. It also obtains the best SSIM and PSNR-B in most cases. Although PromptCIR performs slightly better in SSIM and PSNR-B on ICB at QF10, CACRNet retains the highest PSNR. Under a fixed QF, the overall compression strength is known, but blocking, ringing, and texture attenuation still vary across spatial regions. SCI accounts for these local differences when selecting contextual references, while CAGF suppresses compressed responses in shallow features without discarding reliable edges and textures. This explains the consistent gains across quality factors and shows that CACRNet is not specialized for a particular compression strength.

Table 3: Double JPEG compression restoration results on LIVE1 and ICB. Each entry reports RGB PSNR (dB)/SSIM. Column headers denote $(\mathrm{QF}_1, \mathrm{QF}_2, \mathrm{Shift})$. Best results at the displayed precision are shown in bold.

Method	$(30, 30, 0)$	$(30, 30, 4)$	$(30, 50, 0)$	$(30, 50, 4)$	$(50, 30, 0)$	$(50, 30, 4)$	$(50, 50, 0)$	$(50, 50, 4)$	Average
LIVE1									
FBCNN Jiang et al. (2021)	30.95/0.891	29.64/0.863	30.66/0.891	30.27/0.877	30.55/0.901	30.33/0.879	32.59/0.918	31.23/0.897	30.78/0.890
CRLM Wang et al. (2022a)	31.10/0.894	29.79/0.866	30.81/0.894	30.42/0.880	30.70/0.904	30.48/0.882	32.79/0.921	31.42/0.900	30.94/0.893
PromptIR Potlapalli et al. (2023)	31.18/0.894	29.87/0.866	30.89/0.894	30.50/0.880	30.78/0.904	30.56/0.882	32.84/0.920	31.47/0.899	31.01/0.892
PromptCIR Li et al. (2024)	31.21/0.895	29.90/0.867	30.92/0.895	30.53/0.881	30.80/0.905	30.59/0.883	32.90/0.922	31.53/0.901	31.05/0.894
MaIR Li et al. (2025)	31.23/0.894	29.91/0.865	30.89/0.892	30.52/0.880	30.74/0.902	30.58/0.880	32.87/0.921	31.55/0.899	31.04/0.892
MambaIRv2 Guo et al. (2025a)	31.34/0.895	30.07/0.867	31.09/0.895	30.71/0.882	30.97/0.905	30.76/0.883	**33.07**/0.923	31.67/0.901	31.21/0.894
Q-MambaIR Chen et al. (2025)	31.39/0.896	30.04/0.867	31.07/0.894	30.66/0.881	30.95/0.905	30.72/0.882	33.02/0.9227	31.69/0.901	31.19/0.894
CACRNet	**31.42/0.897**	**30.10/0.868**	**31.11/0.896**	**30.72/0.883**	**31.31/0.908**	**30.78/0.884**	33.03/**0.924**	**31.73/0.902**	**31.28/0.895**
ICB									
FBCNN Jiang et al. (2021)	34.76/0.850	33.82/0.836	34.35/0.847	34.21/0.843	34.12/0.853	34.25/0.843	35.65/0.866	34.86/0.853	34.50/0.849
CRLM Wang et al. (2022a)	34.72/0.849	33.78/0.835	34.31/0.846	34.17/0.842	34.08/0.852	34.21/0.842	35.62/0.860	34.83/0.847	34.46/0.846
PromptIR Potlapalli et al. (2023)	35.20/0.853	34.26/0.839	34.79/0.850	34.65/0.846	34.56/0.856	34.69/0.846	36.08/0.869	35.29/0.856	34.94/0.852
PromptCIR Li et al. (2024)	35.24/0.854	34.30/0.840	34.83/0.851	34.69/0.847	34.60/0.857	34.73/0.847	36.15/0.871	35.36/0.858	34.99/0.853
MaIR Li et al. (2025)	35.30/0.856	34.39/0.843	34.87/0.853	34.72/0.849	34.63/0.858	34.79/0.849	36.32/0.872	35.48/0.860	35.06/0.855
MambaIRv2 Guo et al. (2025a)	35.48/0.857	34.54/0.843	35.07/0.854	34.93/0.850	34.86/0.860	34.97/0.850	36.48/0.873	35.70/0.860	35.25/0.856
Q-MambaIR Chen et al. (2025)	35.44/0.857	34.50/0.843	35.03/0.855	34.89/0.850	34.78/0.860	34.93/0.850	36.43/0.873	35.63/0.861	35.20/0.856
CACRNet	**35.64/0.858**	**34.71/0.844**	**35.18/0.855**	**35.10/0.851**	**35.19/0.863**	**35.13/0.851**	**36.60/0.873**	**35.78/0.861**	**35.42/0.857**

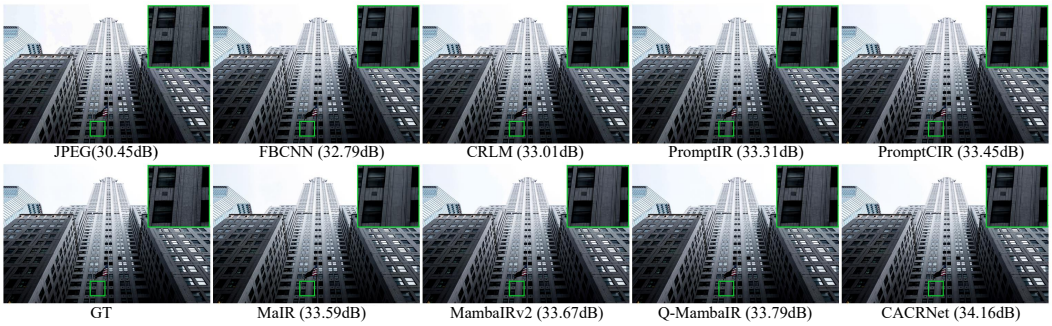


Figure 4: Qualitative comparison on representative images under the same JPEG inputs. The green boxes indicate selected regions, with enlarged patches provided for detailed comparison.

Double JPEG restoration. Results on LIVE1 and ICB are reported in Table 3. On LIVE1, CACRNet achieves the best PSNR in seven of eight settings and the best SSIM throughout, obtaining average scores of 31.28 dB and 0.895. It outperforms MambaIRv2 by 0.07 dB on average, with the largest PSNR gain of 0.34 dB at $(50, 30, 0)$. At $(50, 50, 0)$, MambaIRv2 has a 0.04 dB PSNR advantage, whereas CACRNet achieves higher SSIM. The improvement is more pronounced on ICB, where CACRNet obtains the best PSNR in all eight settings and the best or tied-best SSIM throughout. Its average PSNR reaches 35.42 dB, exceeding MambaIRv2 by 0.17 dB, with a maximum gain of 0.33 dB at $(50, 30, 0)$. Double compression combines artifacts from two quantization stages, while grid shifting introduces spatially misaligned blocking patterns that cannot be represented reliably by a single quality factor or fixed block prior. CCB preserves multiscale frequency and structural information from the input, while SCI retrieves suitable context according to the compound degradation at each location. CAGF further prevents overlapping compression artifacts in shallow features from being transmitted directly to the decoder. The consistent gains on both datasets, particularly on the more structurally complex ICB images, demonstrate the robustness of CACRNet to compression order and grid misalignment.

4.3 Qualitative Comparison

We present a qualitative comparison on a representative building image in Figure 4. JPEG compression weakens high-frequency facade textures and introduces blocking and ringing artifacts around window boundaries. Compared with existing methods, CACRNet reconstructs clearer window edges and more consistent facade patterns while suppressing visible compression artifacts. Additional qualitative comparisons are provided in Appendix A.

Table 4: Ablation study of the architectural components at JPEG quality factor 40.

Method	SCI	CAGF	Axial	LIVE1		BSDS500		ICB	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline (A0)				32.74	0.917	35.08	0.944	37.03	0.880
Model-1 (A1)	✓			32.89	0.918	35.32	0.945	37.21	0.882
Model-2 (A2)		✓		32.81	0.917	35.24	0.944	37.08	0.881
Model-3 (A3)	✓	✓		32.92	0.919	35.36	0.946	37.26	0.882
Model-4 (A4)	✓		✓	32.91	0.918	35.35	0.945	37.27	0.883
Full	✓	✓	✓	**32.94**	**0.919**	**35.39**	**0.946**	**37.32**	**0.883**

4.4 Impact of Architectural Components

To validate the effectiveness of each proposed component, we conduct a series of ablation experiments on LIVE1, BSDS500, and ICB at JPEG quality factor 40. All variants follow the same training and degradation protocol, with the Fourier-RWKV Zheng et al. (2025) block and conventional skip fusion serving as the baseline model.

Effectiveness of Spatial-Channel Interaction. As shown in Table 4, introducing SCI in A1 improves the baseline PSNR by 0.15, 0.24, and 0.18 dB on LIVE1, BSDS500, and ICB, respectively. By replacing global degradation conditioning with local compression evidence, SCI retrieves location-specific references from the image-adaptive contextual bank. The consistent gains demonstrate the effectiveness of local degradation-guided context retrieval.

Effectiveness of Compression-Aware Gated Fusion. Introducing CAGF alone in A2 improves PSNR over the baseline by 0.07, 0.16, and 0.05 dB on the three datasets, demonstrating that filtering artifact-contaminated skip responses helps preserve reliable structural information. When CAGF is combined with SCI, A3 further improves upon A2 by 0.11, 0.12, and 0.18 dB, confirming that contextual retrieval and skip refinement provide complementary benefits.

Effectiveness of Contextual Interaction. Building upon SCI, A4 introduces axial interaction to inject the retrieved context into decoder features. This produces further PSNR gains of 0.02, 0.03, and 0.06 dB over A1. The larger improvement on ICB suggests that non-local contextual interaction is particularly beneficial for complex structures and textures.

4.5 Impact of losses.

As shown in Table 5, using only $\mathcal{L}_{\mathrm{PSNR}}$ provides a stable reconstruction baseline. Adding $\mathcal{L}_{\mathrm{FFT}}$ improves PSNR by 0.05 dB by constraining compression-induced spectral discrepancies, while $\mathcal{L}_{\mathrm{grad}}$ brings a 0.04 dB gain by preserving local edges and structural transitions. Combining all three objectives achieves the best performance of 32.94 dB and 0.919 SSIM. These results demonstrate that frequency correction and gradient preservation provide complementary supervision for JPEG artifact removal.

Table 5: Ablation study of the training objectives on LIVE1 at JPEG quality factor 40.

Setting	$\mathcal{L}_{\mathrm{PSNR}}$	$\mathcal{L}_{\mathrm{FFT}}$	$\mathcal{L}_{\mathrm{grad}}$	PSNR/SSIM
L1	✓			32.8209 / 0.917983
L2	✓	✓		32.8655 / 0.918497
L3	✓		✓	32.8589 / 0.918417
Full	✓	✓	✓	**32.9348 / 0.919320**

5 Conclusion

We introduced a JPEG restoration network in which local degradation evidence controls two routes into the decoder. An image-specific multiscale wavelet bank supplies non-local references, while identity-centered gated fusion regulates the transfer of shallow encoder detail. Gradient, local-variance, and channel-disagreement cues connect these routes by guiding contextual retrieval at each location. This formulation makes global context and local skip transfer parts of the same restoration decision, and the accompanying ablations separate that mechanism from gains due to model capacity.

REFERENCES

Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 182–192, 2021.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pp. 17–33. Springer, 2022.

Yujie Chen, Haotong Qin, Zhang Zhang, Michelo Magno, Luca Benini, and Yawei Li. Q-mambair: Accurate quantized mamba for efficient image restoration. *arXiv preprint arXiv:2503.21970*, 2025.

Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE international conference on computer vision*, pp. 576–584, 2015.

Xueyang Fu, Menglu Wang, Xiangyong Cao, Xinghao Ding, and Zheng-Jun Zha. A model-driven deep unfolding method for jpeg artifacts removal. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6802–6816, 2021a.

Xueyang Fu, Xi Wang, Aiping Liu, Junwei Han, and Zheng-Jun Zha. Learning dual priors for jpeg compression artifacts removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4086–4095, 2021b.

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, 2025a.

Jinpei Guo, Zheng Chen, Wenbo Li, Yong Guo, and Yulun Zhang. Compression-aware one-step diffusion model for jpeg artifact removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14930–14939, 2025b.

Changwoo Han, Hongil Kim, Donghyun Kim, Sung-Chang Lim, and Seung-Won Jung. Universal compressed image restoration via codec-aware conditioning with reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4529–4537, 2026.

Woo Kyoung Han, Sunghoon Im, Jaedeok Kim, and Kyong Hwan Jin. Jdec: Jpeg decoding via enhanced continuous cosine coefficients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2784–2793, 2024.

Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4997–5006, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bingchen Li, Xin Li, Yiting Lu, Ruoyu Feng, Mengxi Guo, Shijie Zhao, Li Zhang, and Zhibo Chen. Promptcir: Blind compressed image restoration with prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6442–6452, 2024.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 17452–17462, 2022.

Boyun Li, Haiyu Zhao, Wenxin Wang, Peng Hu, Yuanbiao Gou, and Xi Peng. Mair: A locality-and continuity-preserving mamba for image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7491–7501, 2025.

Jianwei Li, Yongtao Wang, Haihua Xie, and Kai-Kuang Ma. Learning a single model with a wide range of quality factors for jpeg image artifacts removal. *IEEE Transactions on Image Processing*, 29:8842–8854, 2020.

Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 18278–18289, 2023.

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.

Shuai Liu, Binqiang Liu, Qingyu Mao, Jiacong Chen, Fanyang Meng, Yonghong Tian, and Yongsheng Liang. Blind jpeg artifacts removal via inverse jpeg compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Shuai Liu, Yuqing Zhao, Qingyu Mao, Binqiang Liu, Fanyang Meng, Shuangyan Yi, and Yongsheng Liang. Hierarchical quality-aware guidance for blind jpeg artifacts removal. *Journal of Visual Communication and Image Representation*, pp. 104818, 2026.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pp. 4539–4547, 2017.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 5769–5780, 2022.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pp. 108–126. Springer, 2020.

Xi Wang, Xueyang Fu, Yurui Zhu, and Zheng-Jun Zha. Jpeg artifacts removal via contrastive representation learning. In *European Conference on Computer Vision*, pp. 615–631. Springer, 2022a.

Xi Wang, Xueyang Fu, Liang Li, and Zheng-Jun Zha. Dctmamba: Advancing jpeg image restoration through long-sequence modeling and adaptive frequency strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7925–7933, 2025.

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, 2022b.

Guang Yang, Lu Lin, Chen Wu, and Feng Wang. Dual-domain learning for jpeg artifacts removal. In *International Conference on Neural Information Processing*, pp. 556–568. Springer, 2023.

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, 2020.

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.

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.

Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal. In *2018 25th IEEE international conference on image processing (icip)*, pp. 390–394. IEEE, 2018a.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, 2018b.

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *International Conference on Learning Representations*, 2019.

Lirong Zheng, Yanshan Li, Rui Yu, and Kaihao Zhang. Fourier-rwkv: A multi-state perception network for efficient image dehazing. *arXiv preprint arXiv:2512.08161*, 2025.

A Additional Qualitative Comparisons

This appendix provides additional qualitative comparisons for blind, Non-blind, and double JPEG restoration. All methods are evaluated using the same compressed inputs, and identical regions are enlarged for direct comparison. No additional post-processing is applied to the restored images.

A.1 Blind JPEG Restoration

Figure 5 presents a qualitative comparison under an unknown compression level. The enlarged region contains fine fur textures and high-contrast facial contours, which are susceptible to texture attenuation and ringing artifacts. Compared with competing methods, CACRNet better preserves the eye and mouth boundaries while recovering more coherent local textures with fewer visible compression artifacts. This visual advantage is also consistent with its higher PSNR.

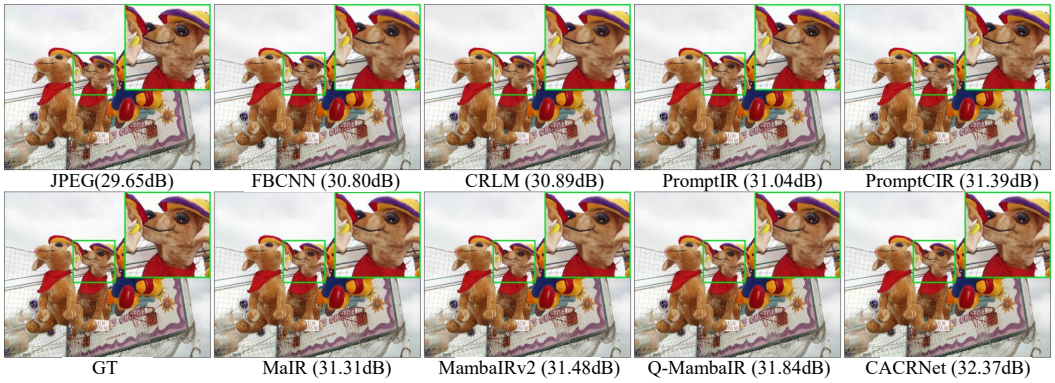


Figure 5: Additional qualitative comparisons for blind JPEG restoration. The quality factors are unknown to the evaluated models. Green boxes indicate enlarged regions.

A.2 Fixed-QF JPEG Restoration

Figure 6 presents a qualitative comparison at the fixed quality factor of 10. The heavily compressed input exhibits pronounced blocking, ringing, and loss of fine structures around the motorcycle components. Compared with competing methods, CACRNet more effectively restores the curved boundaries and repeated wheel structures while producing fewer residual artifacts in the enlarged region. The result demonstrates its ability to recover structural details under severe JPEG compression.

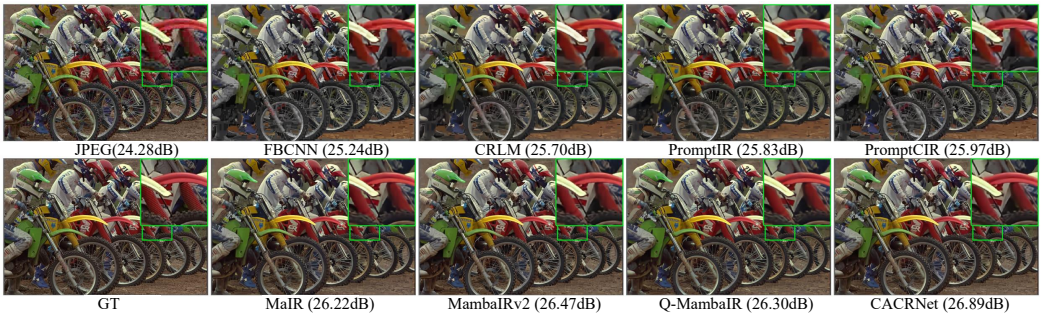


Figure 6: Additional qualitative comparison for JPEG restoration at QF=10. All methods are evaluated on the same compressed input, and the green boxes indicate the enlarged region used for detailed comparison.

A.3 Double JPEG Restoration

Figure 7 presents a qualitative comparison under double JPEG compression with $(30, 30, 4)$. The spatial shift produces misaligned blocking patterns across the two compression stages, leading to

compound artifacts around structural boundaries and repeated textures. As shown in the enlarged roof region, CACRNet more effectively suppresses residual blocking and ringing while preserving clearer roof tiles, window boundaries, and diagonal structures than competing methods.

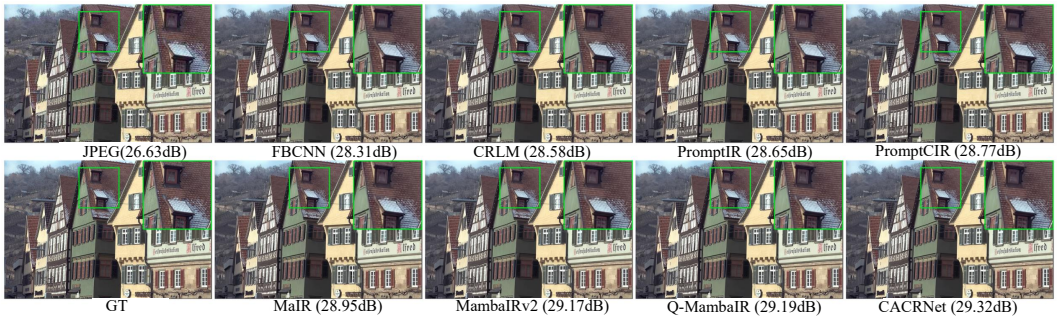


Figure 7: Qualitative comparison for double JPEG restoration under $(\mathrm{QF}_1, \mathrm{QF}_2, \mathrm{Shift}) = (30, 30, 4)$. Green boxes indicate the enlarged region used for detailed comparison.

B AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation, LaTeX formatting, scientific-figure preparation, code development, and feedback on experimental presentation and result interpretation. All AI-assisted text, code, figures, and analyses were reviewed and verified by the authors. All reported numerical results were obtained from actual experiments and independently checked by the authors. The authors take full responsibility for the final content of this work.