

CogSENet: Blind Image Deblurring with Blur-Conditioned Semantic Routing and Explicit Frequency Fusion

Pan Wang¹^{*}, Yihao Hu²^{*},[†], and Xiujin Liu³^{*}

¹ University of Science and Technology of China, Hefei, China

diogenescask@gmail.com

² Westlake University, Hangzhou, China

mistletoehyh@gmail.com

³ University of Michigan, Ann Arbor, MI, USA

jeanliu@umich.edu

Abstract. Blind image deblurring demands the recovery of high-fidelity details and coherent structures from complex, unknown degradations. Current blind image deblurring methods struggle with real-world, spatially varying degradations, and lack the semantic awareness necessary to reliably differentiate valid textures from artifacts. To bridge this gap, we propose **CogSENet**, a dynamic, semantic-aligned reconstruction framework inspired by the eagle’s visual system. By mimicking the eagle’s active saccadic scanning, we devise a Semantic-Driven State Space Module (SDSSM) with semantic-aware token regrouping via differentiable routing, enabling prompt-conditioned long-range dependency modeling. To ensure physically interpretable recovery of textures and structures, a BiFreqFusionBlock (BFFB) mirrors functional differentiation of the eagle’s retina by decomposing features into high and low frequencies using wavelet transforms. Finally, we estimate a continuous Blur Field (CBF) from blur image and fuse it with CLIP semantic priors to modulate the deepest latent features, emulating focal adaptation and enabling adaptive restoration under spatially non-uniform blur. Extensive experiments demonstrate that **CogSENet** outperforms state-of-the-art deblurring methods in both visual quality and structural fidelity with fewer parameters, while also performing favorably on dehazing, deraining, and denoising tasks.

Keywords: Blind image deblurring · Semantic-Driven · Continuous Blur Field

1 Introduction

Blind image deblurring (BID) is a fundamental yet challenging low-level vision task, aiming to recover high-fidelity details and coherent structures from images

^{*} Equal contribution.

[†] Corresponding author.

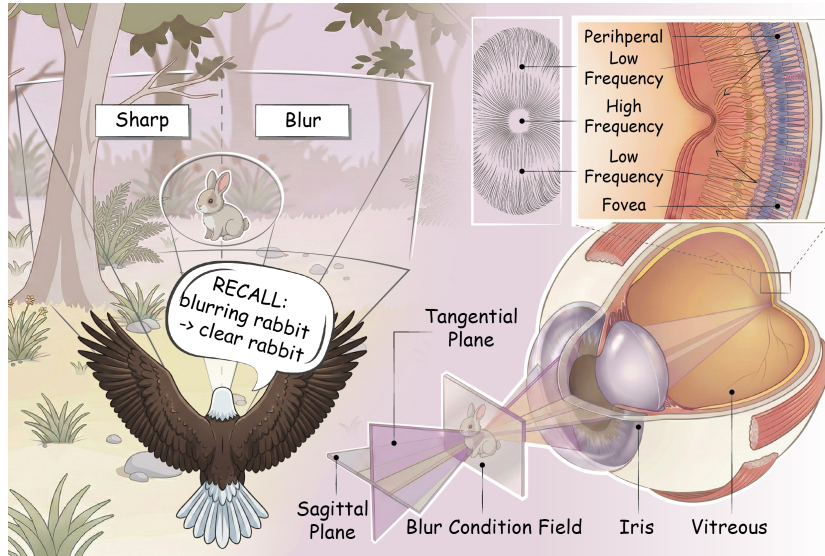


Fig. 1: Biological inspiration of CogSENet. Emulating the eagle’s visual system: (1) Active saccadic scanning for semantic targets inspires our SDSSM to perform semantic-grouped scanning; (2) Retinal differentiation motivates our BFFB to decouple high/low frequencies; (3) Continuous focal adaptation inspires our Continuous Blur Field to model spatially varying blur; (4) Memory recall inspires a frozen CLIP prior to provide high-level semantic guidance at the deepest features.

degraded by unknown and complex blur. Unlike non-blind settings where blur kernels are known, BID requires the network to implicitly or explicitly estimate the degradation while reconstructing the clear image. The field has evolved significantly, shifting from Convolutional Neural Networks (CNNs) [1], which prioritize local extraction but miss global context, to Vision Transformers [35], which capture long-range dependencies at high computational costs. Recently, State Space Models (SSMs) [15, 23], have emerged as a powerful alternative, offering linear complexity and effective global modeling.

However, a fundamental gap persists: most existing methods treat restoration as a static pixel-mapping regression. They often overlook the intrinsic link between high-level semantic understanding and low-level physical degradation. Real-world degradations are inherently non-uniform and spatially varying. This gap becomes more pronounced under complex real-world blurs. Current restoration methods fail because they process image features indiscriminately, lacking mechanisms to explicitly model semantic groupings or separate structural frequencies. Standard SSMs and frequency filters often process image content indiscriminately, failing to distinguish semantic boundaries or decouple complex textures from structural frequencies. Moreover, by relying on static, discrete estimations, these models cannot capture the continuous, spatially-varying trajectory of actual degradation. Thus, distinguishing valid textures from arti-

facts requires the model to have semantic awareness. This motivates our shift from static regression to a dynamic, semantically aligned reconstruction process, where the network understands *what* it is restoring to guide *how* it restores it.

To bridge the gap, as shown in Fig. 1, we draw inspiration from the highly adaptive visual system of an eagle, which maintains extraordinary clarity across diverse environments through four specialized mechanisms. In this work, we reframe blind image restoration from a static pixel-mapping regression to a dynamic, semantic-aligned reconstruction process, and propose **CogSENet**, a unified framework that elevates blind deblurring from a passive pixel-regression to an active, semantic-aligned reconstruction. By shifting the focus to semantic-driven state modeling and explicit frequency fusion, we aim to bridge the gap between semantic perception and physical texture recovery, enabling the network to understand *what* to restore based on semantic context and *how* to restore it via interpretable frequency analysis. To address these challenges, **CogSENet** harnesses the eagle’s natural biological strengths through three core modules working in synergy:

First, mirroring the eagle’s continuous focal adaptation, we estimate a Continuous Blur Field (CBF) and inject it into the deepest latent features before bottleneck processing. Rather than relying on discrete kernel assumptions, CBF spatially varying degradation in a dense form. It is then jointly used with frozen CLIP semantic priors to modulate deep features in a content-adaptive manner, improving restoration under complex non-uniform motion blur. Second, to emulate the functional differentiation of the eagle’s retina, we design BiFreqFusionBlocks (BFFB). By explicitly disentangling features into high-acuity foveal textures and broad-field structural cues via wavelet transforms, BFFB processes these components through specialized transformer and dense convolution branches, ensuring a physically interpretable recovery. Finally, we internalize the eagle’s active saccadic scanning via Semantic-Driven State Space Modules (SDSSM). Instead of uniform 2D selective scan, SDSSM performs differentiable hard routing into semantic groups and scans the semantically regrouped sequence. This design enables a prompt-conditioned, non-causal scanning mechanism to capture the global semantic dependencies beyond fixed causal order.

The main contributions of this paper are summarized as follows:

- We transform long-range dependency modeling from uniform scanning to semantic-driven routing by non-causal token allocation within Semantic-Driven State Space Modules (SDSSM), enabling prompt-modulated selective scan over semantically regrouped tokens.
- We reframe feature processing from implicit spatial convolution to explicit frequency disentanglement by employing wavelet-based decomposition and specialized high/low frequency branches (BFFB), ensuring the physically interpretable recovery of textures and structures.
- We elevate degradation modeling from discrete kernel prediction to continuous blur-field perception, and jointly leverage frozen CLIP semantic priors for deep feature modulation, providing robust conditioning for complex non-uniform blur.

Through this bio-mimetic integration, **CogSENet** fundamentally departs from prior restoration paradigms by unifying semantic routing, frequency-aware reconstruction, and continuous degradation conditioning. Extensive experiments show that, compared with existing SSM-based methods, our approach restores blind images with fewer parameters while outperforming state-of-the-art methods.

2 Related works

CNN-based image restoration methods. In recent years, convolutional neural networks (CNNs) have been the dominant paradigm for low-level image restoration due to their strong locality bias and efficient computation [4, 5, 37]. Early approaches typically built encoder-decoder or multi-scale architectures to enlarge receptive fields, with later designs introducing residual learning and attention to improve texture recovery. While these CNN-based models achieve strong performance across various tasks, their structural rigidity limits their capacity to handle spatially varying degradations where global context is essential. To overcome this local bias, our framework introduces a CLIP-conditioned Continuous Blur Field (CBF). By explicitly injecting global semantic priors, our approach enables continuous, spatially adaptive degradation modeling far beyond the reach of conventional discrete convolutions.

Transformer-based restoration methods. To better model global interactions, Transformers adapt self-attention to dense prediction, capturing long-range correlations that are difficult for purely convolutional designs [3, 22]. To make attention practical for high-resolution inputs, works often adopt windowed, hierarchical, or hybrid CNN-Transformer blocks. Despite their effectiveness, attention-based global modeling incurs high computational costs and can indiscriminately propagate noise or erroneous details when the degradation is ill-posed. In contrast, our Semantic-Driven State Space Module (SDSSM) replaces quadratic self-attention with non-causal dynamic routing. This ensures that only contextually relevant tokens interact, effectively suppressing the propagation of artifacts while maintaining linear complexity.

State-space models for image restoration. More recently, state-space models (SSMs) have emerged as an alternative for global modeling with linear complexity [9, 14, 29]. By using structured recurrent-style propagation (e.g., selective scanning [16]), SSMs offer a favorable trade-off between receptive field and computational efficiency. Nevertheless, without explicit control, global propagation in SSMs can diffuse errors across distant regions and amplify high-frequency artifacts. Furthermore, existing constrained SSMs largely treat image features as homogeneous signals. To address this unexplored bottleneck, our Bi-Frequency Fusion Block (BFFB) explicitly disentangles high-frequency foveal textures from broad low-frequency structures via wavelet transforms. This explicit frequency decoupling ensures physically interpretable and stable state-space propagation, setting our method apart from existing generic SSM designs.

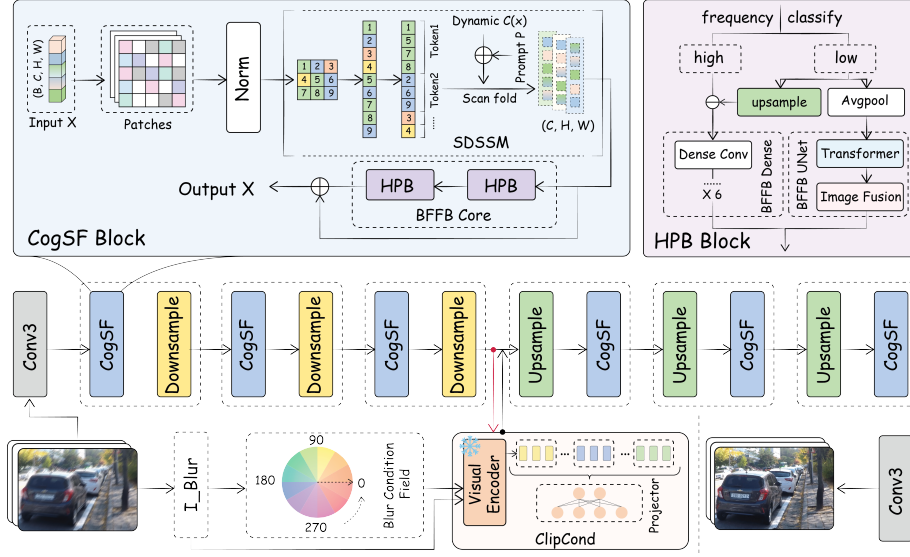


Fig. 2: Our proposed **CogSENet**. A U-shaped backbone is built with stacked CogSF blocks across multi-scale encoder-decoder stages. Each CogSF block integrates an SDSSM module for prompt-guided dynamic token routing and a BFFB core composed of HPB units for explicit high/low-frequency decoupling and fusion. At the deepest bottleneck, we inject a continuous blur field conditioned on frozen CLIP embeddings to modulate bottleneck features.

3 Method

3.1 Overall Architecture

Our goal is to develop a semantically guided and frequency-aware restoration network that recovers sharp structures while remaining computationally practical. To this end, we propose **CogSENet**, a hierarchical encoder-decoder architecture built upon stacked CogSF blocks.

As shown in Fig. 2, given a degraded image $\mathbf{I}_{in} \in \mathbb{R}^{H \times W \times 3}$, we first apply a 3×3 overlap convolution to extract shallow features $\mathbf{F}_s \in \mathbb{R}^{H \times W \times C}$, where C is the base channel width. The shallow feature is then fed into a three-level symmetric encoder-decoder backbone. At level l ($l \in \{1, 2, 3\}$), features have spatial resolution $\frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}}$ and channel dimension $2^{l-1}C$. Each level consists of multiple CogSF blocks. A CogSF block is composed of two complementary modules: an SDSSM module for semantic context aggregation and a BFFB module for explicit bi-frequency refinement (see Secs. 3.2).

For transitions across scales, we use bilinear resizing for both downsampling and upsampling, followed by a 3×3 convolution. Specifically, downsampling reduces the spatial resolution by half while doubling the number of channels, whereas upsampling restores the resolution and correspondingly halves the chan-

nel dimension. We further introduce lateral skip connections between encoder and decoder features at the same scale to preserve fine spatial details.

Different from a standard U-shaped design, we inject top-down guidance only at the bottleneck. After the level-3 encoder, we predict a continuous blur field, implemented as a dense 2D displacement field, from the input using a lightweight convolutional head. We embed the blur field and fuse it with the bottleneck feature map via a lightweight feature fusion layer, producing blur-aware descriptors.

Meanwhile, a frozen CLIP image encoder extracts a global semantic embedding from the input. We then compute a spatial attention map by cosine matching between the semantic embedding and the blur-aware descriptors, and use this attention to modulate the *original* bottleneck features with a single learnable scalar. The resulting bottleneck representation, enriched with both semantic cues and blur awareness, is then decoded and progressively integrated with higher-resolution skip connections, enabling a coarse-to-fine restoration process.

Finally, a lightweight 3×3 reconstruction head predicts a residual image, which is added to the input via an identity shortcut to stabilize optimization and preserve low-frequency content.

3.2 CogSF Blocks

CogSF is the basic building block of **CogSENet**. Each CogSF sequentially applies a Semantic-Driven State-Space Module (SDSSM) to aggregate long-range context in a structured manner, followed by a Bi-Frequency Fusion Block (BFFB) to explicitly refine complementary frequency components. In practice, BFFB has two instantiations: FullBFFB, which follows Eq. (4), and LightBFFB, an efficient approximation using average-pooling-based high/low splitting and lightweight implicit frequency branches.

Semantic-Driven State-Space Module (SDSSM) Generic self-attention lacks explicit semantic organization, while small-kernel convolutions are strictly local. Because degradations are spatially non-uniform, SDSSM couples Gumbel-Softmax semantic routing with efficient regrouping for selective-scan state-space modeling, enabling interaction among distant but semantically related regions.

Semantic routing and prompt construction. Given a feature $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$ flattened into tokens $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^L \in \mathbb{R}^{L \times C}$, we sample an assignment vector $\mathbf{p}_i \in \mathbb{R}^K$ over K clusters via Gumbel-Softmax relaxation. Concurrently, we compute its semantic prompt $\mathbf{z}_i \in \mathbb{R}^{d_s}$ by taking the expectation over a learnable codebook $\mathcal{E} \in \mathbb{R}^{K \times d_s}$:

$$p_{i,k} = \frac{\exp((\mathbf{w}_{\theta,k}^\top \mathbf{t}_i + g_k)/\tau)}{\sum_{j=1}^K \exp((\mathbf{w}_{\theta,j}^\top \mathbf{t}_i + g_j)/\tau)}, \quad \mathbf{z}_i = \sum_{k=1}^K p_{i,k} \mathbf{e}_k \quad (1)$$

where $g_k \sim \text{Gumbel}(0, 1)$ and τ is the temperature. This generates a prompt sequence $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^L$. The continuous $\mathbf{p}_i$ preserves end-to-end differentiability during backpropagation.

Semantic regrouping and prompt-modulated selective scan. We derive a permutation operator $\mathcal{P}_\pi$ by sorting tokens based on their dominant cluster indices ($k_i^* = \arg \max_k p_{i,k}$). To prevent gradient truncation from the non-differentiable $\arg \max$, we explicitly detach the sorting indices and route gradients backward exclusively through the continuous prompt $\mathbf{Z}$. We apply $\mathcal{P}_\pi$ to form semantically contiguous sequences: $\tilde{\mathbf{T}} = \mathcal{P}_\pi(\mathbf{T})$ and $\tilde{\mathbf{Z}} = \mathcal{P}_\pi(\mathbf{Z})$. Scanning along this semantic sequence—combined with alternating spatial flips across network blocks—allows the native 1D causal scan to inherently achieve non-causal global perception. For the regrouped sequence $\tilde{\mathbf{t}}_j \in \mathbb{R}^{d_{inner}}$, the scan operates via discretized matrices $\bar{\mathbf{A}}_j$ and $\bar{\mathbf{B}}_j$. We inject context-aware dynamics by adding $\tilde{\mathbf{z}}_j$ to the emission matrix $\mathbf{C}_j$:

$$\mathbf{h}_j = \bar{\mathbf{A}}_j \mathbf{h}_{j-1} + \bar{\mathbf{B}}_j \tilde{\mathbf{t}}_j, \quad \tilde{\mathbf{y}}_j = (\mathbf{C}_j + \tilde{\mathbf{z}}_j) \mathbf{h}_j + \mathbf{D} \tilde{\mathbf{t}}_j \quad (2)$$

Here, $\mathbf{h}_j \in \mathbb{R}^{d_s}$. The state parameters $\bar{\mathbf{B}}_j \in \mathbb{R}^{d_s}$, $\mathbf{C}_j \in \mathbb{R}^{d_s}$, and the step Δ_j (used for $\bar{\mathbf{A}}_j$) are dynamically projected from $\tilde{\mathbf{t}}_j$. To maintain efficiency, these projection weights and the skip parameter $\mathbf{D} \in \mathbb{R}^{d_{inner}}$ are strictly shared across all semantic groups.

Output restoration. Finally, we apply the inverse permutation $\mathcal{P}_\pi^{-1}$ and sequence-to-grid folding $\mathcal{R}(\cdot)$ to form the residual update:

$$\mathbf{X}_{out} = \mathbf{X} + \mathcal{R}(\mathcal{P}_\pi^{-1}(\tilde{\mathbf{Y}})) \quad (3)$$

Bi-Frequency Fusion Block (BFFB) While SDSSM provides global context, high-quality restoration also requires explicitly handling frequency components: low-frequency structures govern global appearance, whereas high-frequency details determine sharpness. BFFB addresses this by combining an explicit high/low decomposition pathway via wavelet transforms with an efficient spectral refinement branch, adaptively fusing their outputs.

Explicit high/low modeling. To achieve physically interpretable frequency disentanglement, we employ the orthogonal wavelet projection $\mathcal{W}$ (e.g., Haar wavelet) to decompose the input feature map $\mathbf{X}$ into low-frequency approximation coefficients $\mathbf{X}_{LL}$ and high-frequency detail coefficients $\mathbf{X}_H$. The explicit high/low modeling branch, denoted as $\mathcal{H}(\cdot)$, processes these decomposed sub-bands independently. Specifically, a low-frequency branch Φ_{low} models global structures using a lightweight U-shaped module, while a high-frequency branch Φ_{high} enhances local details using dense convolutions.

Frequency-domain refinement. In parallel, the frequency-domain refinement branch $\mathcal{F}(\cdot)$ maps the signal to the Fourier domain for adaptive spectral filtering. By reweighting the frequency responses via a learned spectral filter $\mathbf{M}$ and projecting back to the spatial domain, this branch explicitly suppresses artifacts and recovers fine textures. This operation leverages the discrete Fourier transform $\mathcal{F}$ and its inverse $\mathcal{F}^{-1}$ to ensure rigorous spectral modulation.

Adaptive fusion. The final restored feature $\mathbf{X}_{\text{out}}$ is obtained via an adaptive, parameterized residual fusion that balances the physical structure from the wavelet domain and the spectral texture from the Fourier domain:

$$\mathbf{X}_{\text{out}} = \mathbf{X} + \beta \mathcal{W}^{-1}(\Phi_{\text{low}}(\mathbf{X}_{LL}), \Phi_{\text{high}}(\mathbf{X}_H)) + (1 - \beta) \mathcal{F}^{-1}(\mathbf{M} \odot \mathcal{F}(\mathbf{X})), \quad (4)$$

where $\mathcal{W}^{-1}$ is the inverse wavelet transform, $\odot$ denotes element-wise multiplication, and β is a learnable scalar dynamically adapting to the content to weight the two branches. This design enables content-aware restoration that jointly preserves robust global structures and sharp fine details.

3.3 Continuous Blur Field and Semantic Joint Modulation

To emulate focal adaptation under spatially non-uniform degradations, we propose to jointly modulate the deepest latent features using both physical and high-level cognitive priors. As shown in our network bottleneck, this is achieved through a Continuous Blur Field (CBF) coupled with CLIP semantics.

Concretely, a lightweight CNN extracts a degradation prior from the input image $\mathbf{I}_{in}$ and outputs a continuous flow field $\mathbf{u} \in \mathbb{R}^{B \times 2 \times H \times W}$. To stabilize optimization and constrain blur magnitude, we bound the field with $\tanh(\cdot)$ and scale it by a maximum displacement d_{max} :

$$\mathbf{u} = d_{\text{max}} \cdot \tanh(\text{CNN}_{\text{blur}}(\mathbf{I}_{in})). \quad (5)$$

We apply CBF only at the deepest bottleneck to inject degradation awareness into global-context features. Given bottleneck features $\mathbf{F} \in \mathbb{R}^{B \times C \times h \times w}$, we first resize $\mathbf{u}$ to the bottleneck resolution and encode it with a convolutional embedding. The encoded degradation feature is then concatenated with $\mathbf{F}$ and fused by a lightweight 1×1 convolution to obtain blur-aware descriptors $\tilde{\mathbf{F}}$.

In parallel, a frozen CLIP image encoder extracts a global semantic token $\mathbf{c} \in \mathbb{R}^d$. To make cosine similarity well-defined across modalities, we map CLIP semantics and blur-aware descriptors into the same channel dimension before ℓ_2 normalization. Then, for each sample, we compute a spatial attention map by pixel-wise cosine similarity followed by spatial softmax:

$$A_{i,j}^{(b)} = \frac{\exp(\langle \bar{\mathbf{c}}^{(b)}, \bar{\mathbf{F}}_{i,j}^{(b)} \rangle)}{\sum_{x=1}^h \sum_{y=1}^w \exp(\langle \bar{\mathbf{c}}^{(b)}, \bar{\mathbf{F}}_{x,y}^{(b)} \rangle)}, \quad (6)$$

where $\langle \cdot, \cdot \rangle$ denotes inner product, and the bar notation represents ℓ_2 -normalized features.

Finally, we modulate the *original* bottleneck feature map $\mathbf{F}$ using residual attention gating:

$$\mathbf{F}'_{i,j} = \mathbf{F}_{i,j} + \gamma(A_{i,j} \cdot \mathbf{F}_{i,j}), \quad (7)$$

where γ is a learnable scalar initialized to zero. This preserves identity mapping at early training and gradually introduces semantic-physical guidance as training proceeds. The guided representation $\mathbf{F}'$ is then fed into the decoder for coarse-to-fine restoration.

Table 1: Quantitative comparisons on GoPro, HIDE, RealBlur-R and RealBlur-J datasets. PSNR and SSIM are reported separately. **Boldface** and underlined indicate the best and second-best results, respectively.

Methods	Params (M)↓	GoPro		HIDE		RealBlur-R		RealBlur-J	
		PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑
Restormer [35]	26.1	33.57	0.9656	31.22	0.9420	36.19	0.9570	28.96	0.8790
MPRNet [36]	20.1	32.66	0.9590	30.96	0.9390	39.31	0.9720	31.76	0.9220
NAFNet [1]	67.9	33.71	0.9668	31.31	0.9427	39.80	0.9731	32.40	0.9249
ALGNet [12]	54.6	33.49	0.9640	31.64	0.9470	36.35	0.961	29.12	0.8860
SRN [30]	6.8	30.26	0.9342	28.36	0.9040	38.65	0.9652	31.38	0.9091
MIMO-Unet+ [2]	16.1	32.45	0.9567	29.99	0.9304	38.55	0.9650	31.92	0.9190
Stripformer [32]	19.7	33.08	0.9624	31.03	0.9395	39.84	0.9737	32.48	0.9290
AdaRevD [24]	152.4	34.50	0.9710	32.26	0.9520	36.56	0.957	30.12	0.8940
EVSSM [19]	17.1	34.51	0.9713	31.99	0.9503	41.27	0.9776	34.34	0.9456
MLWNet [13]	24.1	33.8	0.9676	31.06	0.9320	40.69	0.9760	33.84	0.9410
BANet [33]	18.0	32.54	0.9570	30.16	0.9300	39.55	0.971	32.00	0.9230
MRLPFNet [6]	20.6	33.50	0.9650	31.63	0.9465	40.92	0.9753	33.19	0.9361
BANet+ [33]	40.0	33.03	0.9610	30.58	0.9350	39.90	0.972	32.42	0.9290
FFTformer [18]	16.6	34.21	0.9692	31.62	0.9455	40.11	0.9753	32.62	0.9326
Ours	<u>8.9</u>	34.72	<u>0.9744</u>	32.18	<u>0.9514</u>	41.83	<u>0.9799</u>	<u>34.54</u>	<u>0.9473</u>
Ours+	19.4	34.91	0.9757	32.42	0.9533	41.91	0.9810	34.72	0.9488

4 Experiments

4.1 Experimental Settings

Dataset and Evaluation. Following recent state-of-the-art image restoration methods, we evaluate our approach on benchmark datasets spanning image deblurring, deraining, dehazing, and low-light enhancement. For image deblurring, we use the widely adopted GoPro [26], HIDE [28], and RealBlur datasets [27]. RealBlur contains two subsets, RealBlur-J and RealBlur-R, constructed with different post-processing pipelines, and comprises 3,758 training images and 980 testing images in total. For cross-task evaluation, we use Rain100L [10] for deraining, RESIDE-SOTS [20] for dehazing, and BSD68 [25] for denoising task. For fair comparison, we strictly follow the standard training and testing protocols of each dataset. Unless otherwise stated, quantitative performance is reported using PSNR and SSIM.

Training Details. The shallow feature dimension is set to 48, and the numbers of backbone blocks at the three hierarchical levels are [6, 6, 12]. We train the model using AdamW and apply standard data augmentation, including random horizontal/vertical flips and rotations. Training follows a three-stage progressive strategy. In Stage 1, we train with a patch size of 128×128 and a total batch size of 32 for 300,000 iterations, with the learning rate annealed from 1×10^{-3} to 1×10^{-7} . In Stage 2, the patch size is increased to 256×256 and the total batch size is reduced to 8 for another 300,000 iterations; the learning rate is reset to 5×10^{-4} and annealed to 1×10^{-7} . In Stage 3, we further increase the patch size to 320×320 and reduce the total batch size to 4 for 150,000 iterations, initializing the learning rate at 1×10^{-4} and gradually decaying it to

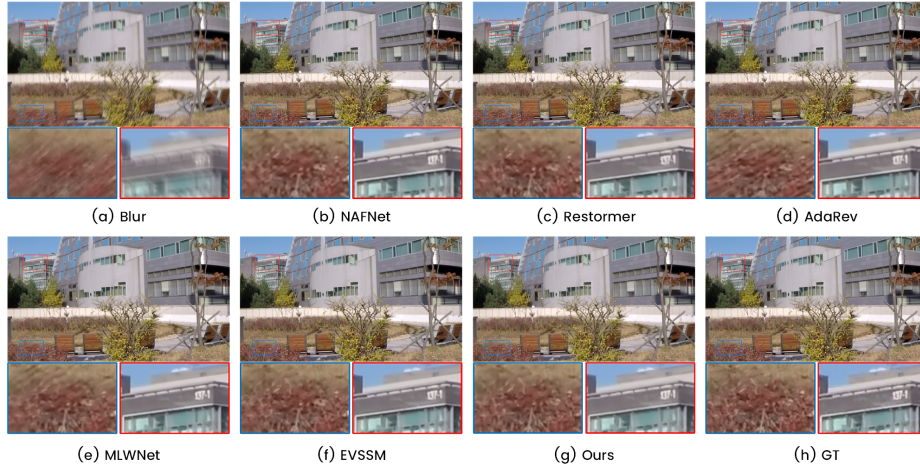


Fig. 3: Visual comparison on the GoPro dataset. Compared with the methods in (b)–(f), the proposed method produces a deblurred image with clearer structures and finer details in (g), yielding results that are more faithful to the ground truth in (h).

1×10^{-7} . We use cosine annealing for learning-rate scheduling throughout. For cross-task training, we disable the blur-field branch. In our implementation, this branch predicts a dense 2D continuous blur field from the input image and injects it into the deepest features via a conditional fusion module for blur-specific, motion-aware modulation. Because these blur-specific cues are not required for non-deblurring tasks, disabling this branch avoids introducing task-mismatched blur priors. Unless otherwise stated, all experiments are implemented in PyTorch and conducted on four NVIDIA L40 GPUs. The training code and models are available at CogSENet.

Blind Image Motion Deblurring. Our denotes the lightweight configuration, which uses LightBFFB throughout the backbone for efficiency. Our+ denotes a hybrid configuration, which keeps LightBFFB in the backbone while introducing a FullBFFB at the bottleneck to enhance representational capacity. Furthermore, the CLIP model is strictly frozen and acts solely as an external semantic prior. Since it requires no gradient updates, its parameters are excluded from the total model complexity.

Tab. 1 reports quantitative comparisons on GoPro, HIDE, RealBlur-R, and RealBlur-J. Our base model achieves 34.72 dB PSNR and 0.9744 SSIM on GoPro with only 8.9M parameters, yielding a favorable accuracy-efficiency trade-off. Compared with representative efficient methods, Our improves PSNR by 0.51 dB over FFTformer (34.21 dB) and by 0.21 dB over EVSSM (34.51 dB), while using substantially fewer parameters (8.9M vs. 16.6M/17.1M). Building on this lightweight baseline, Our+ further improves restoration quality across all deblurring benchmarks, validating the benefit of stronger bottleneck frequency fusion.

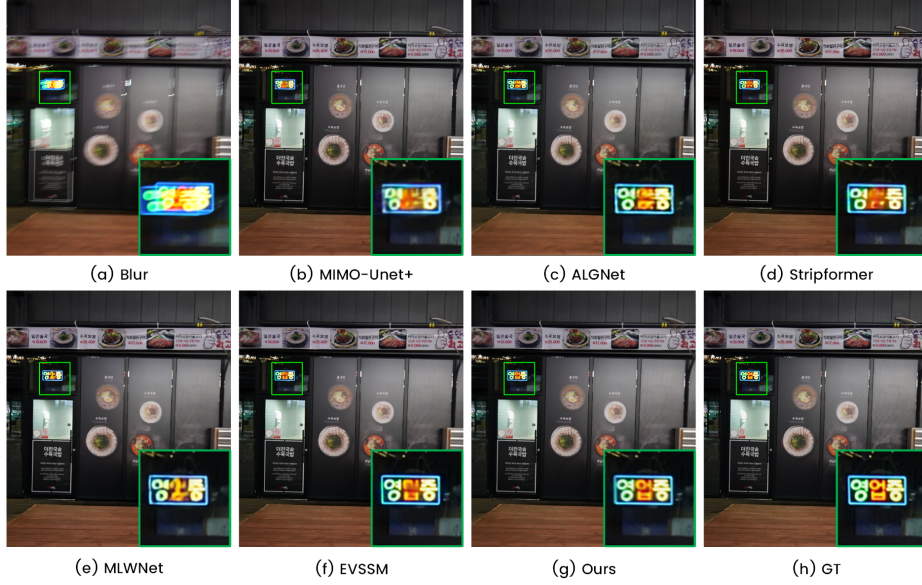


Fig. 4: Deblurred results on the RealBlur dataset. The characters and structural details in (b)–(f) are still blurry, while our method restores sharper characters and clearer structural details.

Table 2: Quantitative comparison on the image deraining task.

Dataset	MSPFN [17]		LPNet [11]		MPRNet [36]		Restormer [35]		SwinIR [22]		IDT [34]		Ours	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Rain100L	32.40	0.9330	34.26	0.9500	36.40	0.9650	38.99	0.9780	<u>39.02</u>	<u>0.9788</u>	39.05	0.9792	<u>39.02</u>	0.9785
Rain100H	28.66	0.8600	23.73	0.8100	30.41	0.8900	31.46	0.9040	31.75	0.9050	32.17	<u>0.9080</u>	<u>32.15</u>	0.9099

For cross-dataset generalization, we follow the common protocol of training on GoPro and testing on HIDE. Under this setting, Our reaches 32.18 dB PSNR and 0.9514 SSIM, and Our+ provides further gains under the same protocol. On real-world benchmarks, Our obtains 41.83/0.9799 (PSNR/SSIM) on RealBlur-R and 34.54/0.9473 on RealBlur-J; Our+ remains consistently stronger, indicating improved robustness to diverse real blur patterns and imaging conditions.

Qualitative comparisons are shown in Fig. 3 and Fig. 4. On GoPro, previous methods tend to leave residual blur around fine structures and high-frequency regions, while our method recovers cleaner edges and sharper local textures. On RealBlur, especially in character and contour regions, competing methods still produce over-smoothed details, whereas our reconstruction preserves clearer boundaries and more faithful structures. These visual observations are consistent with the quantitative improvements in Tab. 1.

Single Image Deraining. Tab. 2 reports the quantitative comparison on the Rain100L and Rain100H benchmarks. While most state-of-the-art methods sat-

Table 3: Quantitative comparison on image denoising and dehazing tasks.

Task	Dataset	BRDNet [31]		LPNet [11]		FDGAN [7]		MPRNet [36]		DL [8]		AirNet [21]		Ours	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Denoise	BSD68 ($\sigma = 15$)	<u>34.10</u>	0.9291	32.31	0.9236	31.11	0.9147	34.01	0.9334	33.25	0.9225	34.14	0.9356	34.02	<u>0.9336</u>
	BSD68 ($\sigma = 25$)	31.43	0.8847	27.87	0.8674	29.57	0.8770	31.34	0.8892	30.38	0.8679	<u>31.48</u>	<u>0.8928</u>	31.50	0.8931
	BSD68 ($\sigma = 50$)	28.16	0.7942	25.71	0.7656	27.12	0.7895	28.10	0.8014	26.68	0.7415	<u>28.23</u>	<u>0.8057</u>	28.29	0.8062
Dehaze	SOTS	23.31	0.9116	21.43	0.8631	23.15	0.9207	<u>28.21</u>	<u>0.9672</u>	24.68	0.9243	23.18	0.9000	28.72	0.9692

urate on the light rain dataset (Rain100L), our method demonstrates remarkable robustness on the significantly more challenging Rain100H dataset. Specifically, our model achieves the highest SSIM and highly competitive PSNR, trailing the best PSNR by merely 0.02 dB. Notably, compared to the prominent Restormer, our approach yields a substantial +0.69 dB PSNR improvement on Rain100H. This performance leap under severe degradation confirms that our frequency-aware design (BFFB) effectively decouples high-frequency dense rain streaks from the underlying background, while the SDSSM maintains global structural integrity.

Single Image Denoising. As shown in Tab. 3, our framework delivers leading performance on the BSD68 dataset across varying noise levels. For mild noise ($\sigma = 15$), our method performs comparably to top-tier approaches. More importantly, our architectural superiority becomes increasingly pronounced as the noise severity escalates. At higher noise levels ($\sigma = 25$ and $\sigma = 50$), our model consistently establishes the new state-of-the-art in both PSNR and SSIM. This trend indicates that explicitly modeling the high/low-frequency components allows the network to accurately isolate and suppress high-frequency noise without compromising true image textures in heavily corrupted regimes.

Single Image Dehazing. Tab. 3 further details the evaluation on the standard SOTS dehazing benchmark. Our model secures the top rank, achieving 28.72 dB PSNR and 0.9692 SSIM. Impressively, it outperforms the strong baseline MPRNet by a significant margin of +0.51 dB. Atmospheric haze typically causes global contrast degradation coupled with local detail attenuation. Our superior performance demonstrates that the content-adaptive routing in SDSSM effectively captures long-range dependencies for global contrast recovery, while BFFB refines the attenuated high-frequency details, leading to artifact-free dehazing.

4.2 Analysis and Discussion

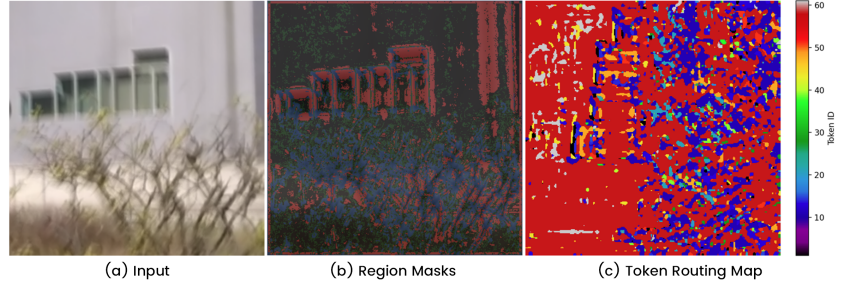
We have shown that **CogSENet** achieves a favorable accuracy–efficiency trade-off against prior deblurring methods. In this section, we present ablation studies to analyze the contributions of its key components. Unless otherwise specified, all ablation variants are trained on the GoPro dataset for 300K iterations using 128×128 patches, the AdamW optimizer (initial learning rate 1×10^{-3} , $\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay 5×10^{-4}), and a True Cosine Annealing schedule with 5K warm-up iterations and a minimum learning rate of 1×10^{-7} . The total batch size is 64 on 4 GPUs.

Table 4: SDSSM ablations on GoPro.

Method	Par. (M)	Mem. (G)	Time (ms)	PSNR /SSIM
w/o prompt	8.9	22	92.11	34.15/0.9685
w/o regroup	8.9	22	92.19	34.21/0.9691
w/o SDSSM	9.1	22	92.54	33.95/0.9677

Table 5: BFFB ablations on GoPro.

Method	Par. (M)	Mem. (G)	Time (ms)	PSNR /SSIM
w/o H/L Split	8.7	20	92.17	34.01/0.9680
w/o Imp. Freq	8.8	21	92.18	34.15/0.9689
w/o BFFB	7.9	20	90.02	33.92/0.9677

**Fig. 5:** Region-aware token routing in SDSSM.

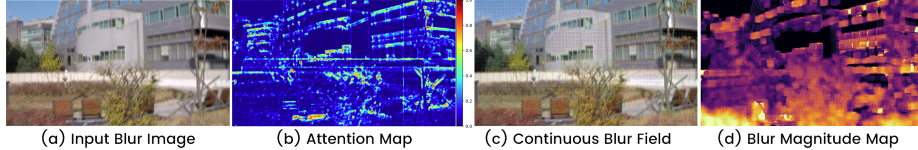
Effects of SDSSM. Table 4 investigates SDSSM on the GoPro dataset. Removing SDSSM entirely degrades performance (33.95 dB) despite incurring higher parameter and latency costs, confirming the superior efficiency of our state-space design. Inside SDSSM, discarding the semantic prompt (34.15 dB) deprives state transitions of essential context-aware dynamics. Furthermore, omitting semantic regrouping (34.21 dB) restricts the scan to rigid spatial axes, failing to associate distant but semantically related regions. By coupling regrouping with prompt modulation, the full SDSSM achieves peak quality while maintaining highly favorable computational footprints.

Fig. 5 further provides qualitative evidence. The routing map is semantically consistent with the degradation masks: large smooth/motion-dominated areas are assigned to dominant token groups, while texture-rich regions (e.g., shrubs and structural edges) activate diverse token IDs. This behavior confirms that SDSSM learns content-adaptive specialization and allocates representation capacity to spatially varying blur patterns.

Effects of BFFB. We ablate the frequency-aware designs within BFFB in Tab. 5. Discarding the explicit high/low-frequency split severely degrades performance, confirming that decoupling low-frequency contents from high-frequency details is essential for accurate restoration. Furthermore, omitting the implicit frequency branch restricts the network’s capacity for dynamic frequency modulation, dropping the PSNR to 34.15 dB. Removing BFFB entirely yields the lowest performance. Overall, BFFB delivers consistent and substantial quality improvements with only a trivial architectural footprint, proving its efficacy as a foundational component.

Table 6: Ablation on Continuous Blur Field and Semantic Priors.

Dataset	CogSENet (Ours)	w/ only CBF	w/ only CLIP	w/ only CBF&CLIP
GoPro	34.31/0.9719	33.80/0.9668	33.78/0.9667	33.88/0.9673
HIDE	31.68/0.9470	31.32/0.9431	31.29/0.9430	31.40/0.9433

**Fig. 6:** Visualization of Continuous Blur Field and Semantic Priors. (a) Input blurry image. (b) CLIP-guided attention map, highlighting semantically informative structures. (c) Estimated continuous blur field. (d) Corresponding blur magnitude map.

Effectiveness of Continuous Blur Field and Semantic Priors. As quantitatively verified in Tab. 6, applying either the physical blur field or cognitive CLIP semantics individually yields limited gains, whereas their joint modulation substantially boosts performance, proving them highly complementary. Fig. 6 visualizes the behavior of our joint modulation module. First, the network directly estimates a spatially continuous blur field from the input to explicitly capture the non-uniform degradation. As shown in (c) and (d), the blur field and its magnitude map are structure-aligned, naturally assigning higher blur intensity to heavily degraded regions while preserving smooth transitions.

By integrating the physical degradation prior with deep latent features and querying it via high-level CLIP semantics, we obtain the attention map visualized in (b). This semantic-driven modulation effectively isolates informative high-frequency structures while suppressing feature responses in non-informative flat regions. Consequently, our joint modulation emulates focal adaptation: it dynamically steers the network to prioritize high-fidelity restoration in severely degraded semantic regions, effectively averting the over-amplification of noise or artifacts in weakly blurred areas.

Limitations. Our physical-semantic joint modulation relies on frozen CLIP priors. As shown in Sec. 4.2, our model occasionally struggles with extreme motion blur that destroys distinct semantic cues. In such failure cases, the CLIP encoder fails to extract valid semantics, causing our focal adaptation to degrade to nearly uniform attention (b-c) and the blur field to miss complex local motions (d-e), ultimately leaving residual blur and vertical artifacts in the restored image (f). Future work includes integrating degradation-robust semantic extractors and extending this architecture to video restoration.

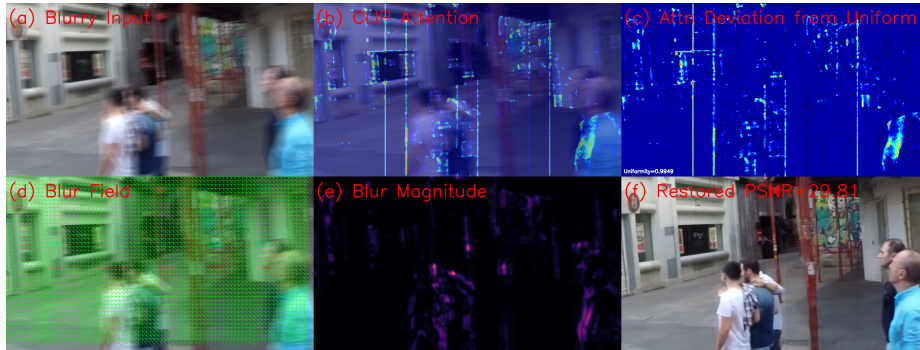


Fig. 7: Failure analysis.

5 Conclusion

In this paper, we propose an efficient region- and frequency-aware framework for image deblurring. We introduce the Semantic-Driven State-Space Module (SDSSM) for linear-complexity, content-adaptive long-range modeling, and the Bi-Frequency Fusion Block (BFFB) to explicitly decouple and refine high/low-frequency components. To tackle spatially non-uniform degradations, we present a physical-semantic joint modulation mechanism coupling a continuous blur field with frozen CLIP priors. This emulates focal adaptation by dynamically prioritizing restoration in heavily blurred, semantic regions. Extensive experiments show our method achieves state-of-the-art restoration quality with highly favorable memory and runtime efficiency.

References

1. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV. pp. 17–33. Springer (2022)
2. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4641–4650 (2021)
3. Conde, M.V., Choi, U.J., Burchi, M., Timofte, R.: Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In: European Conference on Computer Vision. pp. 669–687. Springer (2022)
4. Dong, C., Deng, Y., Loy, C.C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: Proceedings of the IEEE international conference on computer vision. pp. 576–584 (2015)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. pp. 184–199. Springer (2014)
6. Dong, J., Pan, J., Yang, Z., Tang, J.: Multi-scale residual low-pass filter network for image deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12345–12354 (2023)

7. Dong, Y., Liu, Y., Zhang, H., Chen, S., Qiao, Y.: Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 10729–10736 (2020)
8. Fan, Q., Chen, D., Yuan, L., Hua, G., Yu, N., Chen, B.: A general decoupled learning framework for parameterized image operators. *IEEE transactions on pattern analysis and machine intelligence* **43**(1), 33–47 (2019)
9. Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C.: Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052* (2022)
10. Fu, X., Huang, J., Ding, X., Liao, Y., Paisley, J.: Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing* **26**(6), 2944–2956 (2017)
11. Fu, X., Liang, B., Huang, Y., Ding, X., Paisley, J.: Lightweight pyramid networks for image deraining. *IEEE transactions on neural networks and learning systems* **31**(6), 1794–1807 (2019)
12. Gao, H., Ma, B., Zhang, Y., Yang, J., Yang, J., Dang, D.: Learning enriched features via selective state spaces model for efficient image deblurring. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 710–718 (2024)
13. Gao, X., Qiu, T., Zhang, X., Bai, H., Liu, K., Huang, X., Wei, H., Zhang, G., Liu, H.: Efficient multi-scale network with learnable discrete wavelet transform for blind motion deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2733–2742 (2024)
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021)
15. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. In: European conference on computer vision. pp. 222–241. Springer (2024)
16. Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: Localmamba: Visual state space model with windowed selective scan. In: European Conference on Computer Vision. pp. 12–22. Springer (2024)
17. Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., Jiang, J.: 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)
18. Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5886–5895 (2023)
19. Kong, L., Dong, J., Tang, J., Yang, M.H., Pan, J.: Efficient visual state space model for image deblurring. In: Proceedings of the computer vision and pattern recognition conference. pp. 12710–12719 (2025)
20. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing* **28**(1), 492–505 (2019). <https://doi.org/10.1109/TIP.2018.2867951>
21. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: CVPR. pp. 17452–17462 (2022)
22. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1833–1844 (October 2021)
23. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. *Advances in neural information processing systems* **37**, 103031–103063 (2024)

24. Mao, X., Li, Q., Wang, Y.: Adarevd: Adaptive patch exiting reversible decoder pushes the limit of image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25681–25690 (June 2024)
25. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings eighth IEEE international conference on computer vision. ICCV 2001. vol. 2, pp. 416–423. Ieee (2001)
26. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017)
27. Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: European conference on computer vision. pp. 184–201. Springer (2020)
28. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5572–5581 (2019)
29. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933 (2022)
30. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8174–8182 (2018)
31. Tian, C., Xu, Y., Zuo, W.: Image denoising using deep cnn with batch renormalization. Neural Networks **121**, 461–473 (2020)
32. Tsai, F.J., Peng, Y.T., Lin, Y.Y., Tsai, C.C., Lin, C.W.: Stripformer: Strip transformer for fast image deblurring. In: European conference on computer vision. pp. 146–162. Springer (2022)
33. Tsai, F.J., Peng, Y.T., Tsai, C.C., Lin, Y.Y., Lin, C.W.: Banet: A blur-aware attention network for dynamic scene deblurring. IEEE Transactions on Image Processing **31**, 6789–6799 (2022)
34. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 12978–12995 (2023). <https://doi.org/10.1109/TPAMI.2022.3183612>
35. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
36. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021)
37. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017)