ExpoMamba: Exploiting Frequency SSM Blocks for Efficient and Effective Image Enhancement

Eashan Adhikarla [1] **Kai Zhang** [1] **John Nicholson** [2] **Brian D. Davison** [1]

Abstract

Low-light image enhancement remains a challenging task in computer vision, with existing state-of-the-art models often limited by hardware constraints and computational inefficiencies, particularly in handling high-resolution images. Recent foundation models, such as transformers and diffusion models, despite their efficacy in various domains, are limited in use on edge devices due to their computational complexity and slow inference times. We introduce ***ExpoMamba***, a novel architecture that integrates components of the frequency state space within a modified U-Net, offering a blend of efficiency and effectiveness. This model is specifically optimized to address mixed exposure challenges—a common issue in low-light image enhancement—while ensuring computational efficiency. Our experiments demonstrate that ExpoMamba enhances low-light images up to **2-3x** faster than traditional models with an inference time of **36.6 ms** and achieves a PSNR improvement of approximately **15-20%** over competing models, making it highly suitable for real-time image processing applications. Model code is open sourced at: github.com/eashanadhikarla/ExpoMamba.

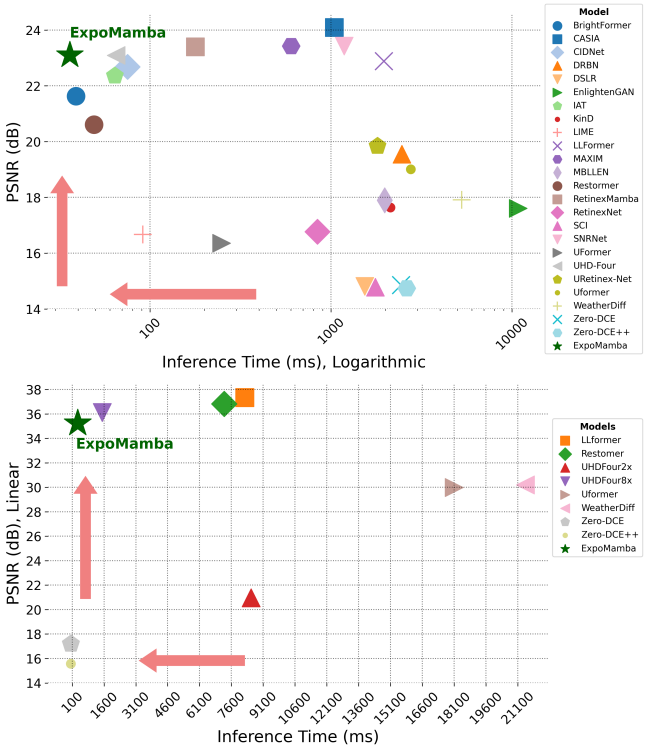

Figure 1. [**top:** 400x600; **bottom:** 3840x2160] Scatter plot of model inference time vs. PSNR. Baselines that used ground-truth mean information to produce metrics were reproduced without such information for fairness.

1. Introduction

Enhancing low-light images is crucial for applications ranging from consumer gadgets like phone cameras (Liba et al., 2019; Liu et al., 2024) to sophisticated surveillance systems (Xian et al., 2024; Guo et al., 2024; Shrivastav, 2024). Traditional techniques (Dale-Jones & Tjahjadi, 1993; Singh et al., 2015; Khan et al., 2014; Land & McCann, 1971; Ren et al., 2020) often struggle to balance processing speed and high-quality results, particularly with high-resolution images, leading to issues like noise and color distortion in scenarios requiring quick processing such as mobile photography and real-time video streaming.

Limitations of Current Approaches. Foundation models have revolutionized computer vision, including low-light image enhancement, by introducing advanced architectures that model complex relationships within image data. In particular, transformer-based (Wang et al., 2023b; Chen et al., 2021a; Zhou et al., 2023b; Adhikarla et al., 2024) and diffusion-based (Wang et al., 2023c;a; Zhou et al., 2023a) low-light techniques have made significant strides. However,

[1]Department of Computer Science, Lehigh University, Bethlehem, PA, USA [2]Lenovo Research, Raleigh, NC, USA. Correspondence to: Eashan Adhikarla <eaa418@lehigh.edu>, Brian D. Davison <bdd3@lehigh.edu>.

Work presented at the ES-FoMo II Workshop at ICML 2024, Vienna, Austria.

the sampling process requires a computationally intensive iterative procedure, and the quadratic runtime of self-attention in transformers make them unsuitable for real-time use on edge devices where limited processing power and battery constraints pose significant challenges. Innovations such as linear attention (Katharopoulos et al., 2020; Shen et al., 2018; Wang et al., 2020), self-attention approximation, windowing, striding (Kitaev et al., 2020; Zaheer et al., 2020), attention score sparsification (Liu et al., 2021b), hashing (Chen et al., 2021c), and self-attention operation kernelization (Katharopoulos et al., 2020; Lu et al., 2021; Chen et al., 2021b) have aimed to address these complexities, but often at the cost of increased computation errors compared to simple self-attention (Duman Keles et al., 2023; Dosovitskiy et al., 2021). (More details can be found in Appendix A)

Purpose. With rising need for better images, advanced small camera sensors in edge devices have made it more common for customers to capture high quality images, and use them in real-time applications like mobile, laptop and tablet cameras (Morikawa et al., 2021). However, they all struggle with non-ideal and low lighting conditions in the real world. Our goal is to develop an approach that has high image quality (e.g., like CIDNet (Feng et al., 2024)) for enhancement but also at high speed (e.g., such as that of IAT (Cui et al., 2022) and Zero-DCE++ (Li et al., 2021)).

Contributions. Our contributions are summarized as:

- We introduce the use of Mamba for efficient low-light image enhancement (LLIE), specifically focusing on mixed exposure challenges, where underlit (insufficient brightness) and overlit (excessive brightness) exist in the same image frame.
- We propose a novel Frequency State Space Block (FSSB) that combines two distinct 2D-Mamba blocks, enabling the model to capture and enhance subtle textural details often lost in low-light images.
- We describe a novel dynamic batch training scheme to improve robustness of multi-resolution inference in our proposed model.
- We implement dynamic processing of the amplitude component to highlight distortion (noise, illumination) and the phase component for image smoothing and noise reduction.

2. Exposure Mamba

Along the lines of recent efficient sequence modeling approaches (Gu & Dao, 2023; Zhu et al., 2024a; Wang et al., 2024), we introduce *ExpoMamba*, a model combining frequency state-space blocks with spatial convolutional blocks (Fig. 2). This combination leverages the advantages of frequency domain analysis to manipulate features at different scales and frequencies, crucial for isolating and enhancing patterns challenging to detect in the spatial domain, like subtle textural details in low-light images or managing noise in overexposed areas. Additionally, by integrating these insights with the linear-time complexity benefits of the Mamba architecture, our model efficiently manages the spatial sequencing of image data, allowing rapid processing without the computational overhead of transformer models.

Our proposed architecture utilizes a 2D scanning approach to tackle mixed-exposure challenges in low-light conditions. This model incorporates a combination of $U^2 - Net$ (Qin et al., 2020) and M-Net (Mehta & Sivaswamy, 2017), supporting 2D `sRGB` images with each block performing operations using a convolutional and encoder-style SSM ($x(t) \in \mathbf{R} \rightarrow y(t) \in \mathbf{R}$)[1]. The subsequent section provides detailed information about our overall pipeline.

2.1. Frequency State Space Block (FSSB)

We utilize the frequency state space block (FSSB) to address the computational inefficiencies of transformer architectures especially when processing high-resolution image or long-sequence data. The FSSB's motivation is in two parts; first, towards enhancing the intricacies that are unaddressed/missed by the spatial domain alone; and second, to speed deep feature extraction using the frequency domain. The FSS block (as in Fig. 3) initiates its processing by transforming the input image I into the frequency domain using the Fourier transform:

$$\mathbf{F}(u, v) = \iint \mathbf{I}(x, y) e^{-i2\pi(ux+vy)} dx \cdot dy \qquad (1)$$

where, $\mathbf{F}(u, v)$ denotes the frequency domain representation of the image, and (u, v) are the frequency components corresponding to the spatial coordinates (x, y). This transformation allows for the isolation and manipulation of specific frequency components, which is particularly beneficial for enhancing details and managing noise in low-light images. By decomposing the image into its frequency components, we can selectively enhance high-frequency components to improve edge and detail clarity while suppressing low-frequency components that typically contain noise (Lazzarini, 2017; Zhou et al., 2022). This selective enhancement and suppression improve the overall image quality.

The core of the FSSB comprises two 2D-Mamba (VisualSSM) blocks to process the `amplitude` and `phase` components separately in the frequency domain. These blocks model state-space transformations as follows:

$$\mathbf{h}[t+1] = \mathbf{A}[t] \cdot \mathbf{h}[t] + \mathbf{B}[t].x[t] \qquad (2)$$

$$\mathbf{y}[t] = \mathbf{C}[t] \cdot \mathbf{h}[t] \qquad (3)$$

Here, $\mathbf{A}[t]$, $\mathbf{B}[t]$, and $\mathbf{C}[t]$ are the state matrices that adapt dynamically based on the input features, and $h[t]$ represents

[1]A state space model is a type of sequence model that transforms a one-dimensional sequence via an implicit hidden state.

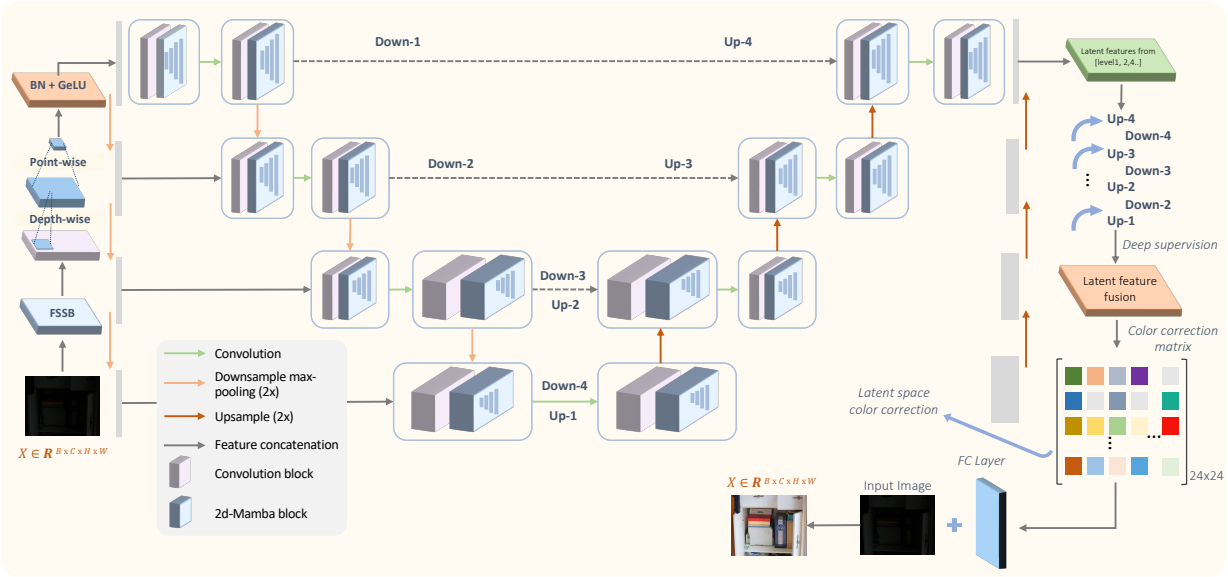

Figure 2. **Overview of the ExpoMamba Architecture.** The diagram illustrates the information flow through the *ExpoMamba* model. The architecture efficiently processes sRGB images by integrating convolutional layers, 2D-Mamba blocks, and deep supervision mechanisms to enhance image reconstruction, particularly in low-light conditions.

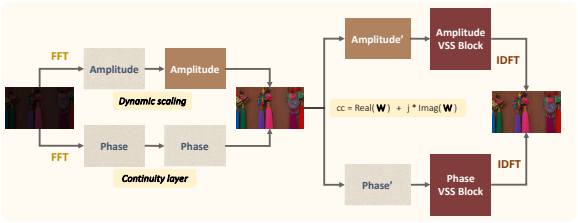

Figure 3. **Frequency State-Space Block (FSSB) Processing.** The FSSB module is detailed within the ExpoMamba architecture.

the state vector at time t. $\mathbf{y}[t]$ represents processed feature at time t, capturing the transformed information from the input features. This dual-pathway setup within the FSSB processes amplitude and phase in parallel.

After processing through each of the VSS blocks, the modified amplitude and phase components are recombined and transformed back to the spatial domain using the inverse Fourier transform:

$$\hat{\mathbf{I}}(x,y) = \iint \hat{\mathbf{F}}(u,v) e^{-i2\pi(ux+vy)} du \cdot dv \qquad (4)$$

where, $\hat{\mathbf{F}}(u,v)$ is the processed frequency domain representation in the latent space of each M-Net block. This method preserves the structural integrity of the image while enhancing textural details that are typically lost in low-light conditions, removing the need of self-attention mechanisms that are widely seen in transformer-based pipelines (Tay et al., 2022). The FSSB also integrates hardware-optimized strategies similar to those employed in the Vision-Mamba architecture (Gu & Dao, 2023; Zhu et al., 2024a) such as scan operations and kernel fusion reducing amount of memory IOs, facilitating efficient data flow between the GPU's memory hierarchies. This optimization significantly reduces computational overhead by a factor of $O(N)$ speeding the operation by $20-40$ times (Gu & Dao, 2023), enhancing processing speed for real-time applications. This can be evidently seen through our Fig. 1, where increasing the resolution size/input length increases the inference time gap tremendously due to which is more for transformer based models due to $O(N^2)$.

Within the FSS Block, the amplitude $\mathbf{A}(u,v)$ and phase $\mathbf{P}(u,v)$ components extracted from $\mathbf{F}(u,v)$ are processed through dedicated state-space models. These models, adapted from the Mamba framework, are particularly tailored (dynamic adaptation of state matrices ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$) based on spectral properties and the dual processing of amplitude and phase components.[2]) to enhance information across frequencies, effectively addressing the typical loss of detail in low-light conditions.

Amplitude and Phase Component Modeling. Each component $\mathbf{A}(u,v)$ and $\mathbf{P}(u,v)$ undergoes separate but parallel processing paths, modeled by:

$$s_{t+1} = \mathbf{A}(s_t) + \mathbf{B}\mathbf{F}(\mathbf{x_t}), \qquad y_t = \mathbf{C}\mathbf{s_t} \qquad (5)$$

where s_t denotes the state at time t, $\mathbf{F}(\mathbf{x_t})$ represents the frequency-domain input at time t (either amplitude or phase),

[2]refer to FSSB module in Appx E

and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are the state-space matrices that dynamically adapt during training.

Frequency-Dependent Dynamic Adaptation. The matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are not only time-dependent but also frequency-adaptive, allowing the model to respond to varying frequency components effectively. This adaptation is crucial for enhancing specific frequencies more affected by noise and low-light conditions. Specifically, these matrices evolve based on the spectral properties of the input: $\mathbf{A}(\mathbf{u}, \mathbf{v}, \mathbf{t}), \mathbf{B}(\mathbf{u}, \mathbf{v}, \mathbf{t}), \mathbf{C}(\mathbf{u}, \mathbf{v}, \mathbf{t})$ adjust dynamically during the processing. This means that $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ change their values according to both the time step t and the frequency components (u, v), enabling targeted enhancement of the amplitude and phase components in the frequency domain. By evolving to match the spectral characteristics of the input, these matrices optimize the enhancement process.

After separate processing through the state-space models, the modified amplitude $\mathbf{A}''(\mathbf{u}, \mathbf{v})$ and phase $\mathbf{P}''(\mathbf{u}, \mathbf{v})$ are recombined and transformed back into the spatial domain to reconstruct the enhanced image:

$$\mathbf{I}'(\mathbf{x}, \mathbf{y}) = \mathbf{F}^{-1}(\mathbf{A}''(\mathbf{u}.\mathbf{v}) + i \cdot \mathbf{P}''(\mathbf{u}.\mathbf{v})) \tag{6}$$

where, $\mathbf{I}^{-1}$ denotes the inverse Fourier transform.

Feature Recovery in FSSB. The HDR (High Dynamic Range) tone mapping process within the Frequency State Space Block (FSSB) is designed to enhance visibility and detail in low-light conditions by selectively normalizing brightness in overexposed areas. Feature recovery in FSSB aims to address the challenges of high dynamic range scenes, where standard methods often fail to maintain natural aesthetics and details. By implementing a thresholding mechanism set above 0.90, the HDR layer selectively applies tone mapping to overexposed areas, effectively normalizing brightness without compromising detail or causing unnatural halos often seen in standard HDR processes (Fig. 4). This selective approach is crucial as it maintains the natural aesthetic of the image while enhancing visibility and detail. The HDR layer is consistently applied as the final layer within each FSSB block and culminates as the ultimate layer in the ExpoMamba model, providing a coherent enhancement across all processed images.

We leverage the ComplexConv function from complex networks as introduced by Trabelsi et al. (Trabelsi et al., 2018). This function is incorporated into our model to capture and process additional information beyond traditional real-valued convolutions. Specifically, the ComplexConv function allows the simultaneous manipulation of amplitude and phase information in the frequency domain, which is essential to preserve the integrity of textural details in low-light images. The dual processing of amplitude and phase ensures that each component to be optimized separately. Tone mapping and ComplexConv have proven to be effective

Figure 4. Representing the effectiveness of HDR tone mapping layer inside FSS block. Using CSRNet with shrinked conditional blocks and dilated convolutions to remove overexposed artifacts.

in overcoming limitations of traditional image processing techniques (Hu et al., 2022; Liu, 2024). We integrate these methods into our FSS design to address adverse lighting conditions in low light environments.

The input components in the frequency representation are processed through dynamic amplitude scaling and phase continuity layer, as shown in Fig. 3. As claimed by Fourmer (Zhou et al., 2023b), we have determined that the primary source of image degradation is indeed amplitude, specifically in the area between the amplitude and phase division within the image. Moreover, we found that the amplitude component primarily contains information about the brightness of the image, which directly impacts the visibility and the sharpness of the features within the image. However, the phase component encodes the positional information of these features, defining the structure and the layout of the image. Previously, it has been found that phase component of the image has a close relation with perceptual analysis (Xiao & Hou, 2004). Along those lines, we show that the human visual system is more sensitive to changes in phase rather than amplitude (proof in Appx C.1).

2.2. Multi-modal Feature Learning

The inherent complexity of low-light images, where both underexposed and overexposed elements coexist, necessitates a versatile approach to image processing. Traditional methods, which typically focus either on spatial details or frequency-based features, fail to adequately address the full spectrum of distortions encountered in such environments. By contrast, the hybrid modeling approach of "ExpoMamba" leverages the strengths of both the spatial and frequency domains, facilitating a more comprehensive and nuanced enhancement of image quality.

Operations in the frequency domain, such as the Fourier transform, can isolate and address specific types of distortion, such as noise and fine details, which are often exacerbated in low-light conditions. This domain provides a global view of the image data, allowing for the manipulation of features that are not easily discernible in the spatial layout. Simultaneously, the spatial domain is critical to maintaining the local coherence of image features, ensuring that enhancements do not introduce unnatural artifacts.

Table 1. Comparing four popular metrics such that every column showcases the top three methods; Red , Green , and Blue representing the best, second best, and third best models among the proposed and all popular SOTA models from 2011–2024.

Methods	Reference	LOLv1				LOLv2 (Real Captured)				Inference time (ms)
		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	
NPE†	TIP'13	16.970	0.484	0.400	104.05	17.333	0.464	0.396	100.02	-
SRIE†	CVPR'16	11.855	0.495	0.353	088.72	14.451	0.524	0.332	078.83	-
BIMEF†	arXiv'17	13.875	0.595	0.326	-	17.200	0.713	0.307	-	-
FEA†	ICME'11	16.716	0.478	0.384	120.05	17.283	0.701	0.398	119.28	-
MF†	Signal Process'16	16.966	0.507	0.379	-	17.500	0.751	-	-	-
LIME†	TIP'16	17.546	0.531	0.387	117.89	17.483	0.505	0.428	118.17	91.12
Retinex‡	BMVC'18	16.774	0.462	0.417	126.26	17.715	0.652	0.436	133.91	4493
DSLR‡	TMM'20	14.816	0.572	0.375	104.43	17.000	0.596	0.408	114.31	1537
KinD‡	ACM MM'19	17.647	0.771	0.175	-	-	-	-	-	2130
DRBN‡	CVPR'20	16.677	0.73	0.345	098.73	18.466	0.768	0.352	089.09	2462
Zero-DCE	CVPR'20	14.861	0.562	0.372	087.24	18.059	0.58	0.352	080.45	2436
Zero-DCE++	TPAMI'21	14.748	0.517	0.328	-	-	-	-	-	2618
MIRNet	ECCV'20	24.138	0.830	0.250	069.18	20.020	0.82	0.233	049.11	1795
EnlightenGAN‡	TIP'21	17.606	0.653	0.372	094.70	18.676	0.678	0.364	084.04	-
ReLLIE‡	ACM MM'21	11.437	0.482	0.375	095.51	14.400	0.536	0.334	079.84	3.500
RUAS‡	CVPR'21	16.405	0.503	0.364	101.97	15.351	0.495	0.395	094.16	15.51
DDIM	ICLR'21	16.521	0.776	0.376	084.07	15.280	0.788	0.387	076.39	1213
CDEF	TMM'22	16.335	0.585	0.407	090.62	19.757	0.63	0.349	074.06	-
SCI	CVPR'22	14.784	0.525	0.366	078.60	17.304	0.54	0.345	067.62	1755
URetinex-Net	CVPR'22	19.842	0.824	0.237	052.38	21.093	0.858	0.208	049.84	1804
SNRNet‡	CVPR'22	23.432	0.843	0.234	055.12	21.480	0.849	0.237	054.53	72.16
Uformer*	CVPR'22	19.001	0.741	0.354	109.35	18.442	0.759	0.347	098.14	901.2
Restormer*	CVPR'22	20.614	0.797	0.288	073.10	24.910	0.851	0.264	058.65	513.1
Palette♣	SIGGRAPH'22	11.771	0.561	0.498	108.29	14.703	0.692	0.333	083.94	168.5
UHDFour‡	ICLR'23	23.093	0.821	0.259	056.91	21.785	0.854	0.292	060.84	64.92
WeatherDiff♣	TPAMI'23	17.913	0.811	0.272	073.90	20.009	0.829	0.253	059.67	5271
GDP♣	CVPR'23	15.896	0.542	0.421	117.46	14.290	0.493	0.435	102.41	-
DiffLL♣	ACM ToG'23	26.336	0.845	0.217	048.11	28.857	0.876	0.207	045.36	157.9
CIDNet‡	arXiv'24	23.090	0.851	0.085	-	23.220	0.863	0.103	-	-
LLformer*	AAAI'23	22.890	0.816	0.202	-	23.128	0.855	0.153	-	1956
ExpoMamba		22.870	0.845	0.215	097.65	23.000	0.860	0.203	094.27	36.00
ExpoMamba$_{da}$	-	23.092	0.847	0.214	092.17	23.131	0.868	0.224	090.22	38.00
ExpoMamba$_{gt}$		25.770	0.860	0.212	089.21	28.040	0.885	0.232	085.92	36.00

"**da**" - Dynamic adjustment. (refer Appendix-C.3) / "**gt**" - With ground-truth mean.

Finally, the hybrid-modeled features pass through deep supervision, where we combine ExpoMamba's intermediate layer outputs, apply a color correction matrix in the latent dimensions during deep supervision, and pass through the final layer.

2.3. Dynamic Patch Training

Dynamic patch training enhances the 2D scanning model by optimizing its scanning technique for various image resolutions. In ExpoMamba, 2D scanning involves sequentially processing image patches to encode feature representations. We create batches of different resolution images where in a given batch the resolution is fixed and we dynamically randomize the different batch resolutions of input patches during training. This way the model learns to adapt its scanning and encoding process to different scales and levels of detail (Fig 5). This variability helps the model become more efficient at capturing and processing fine-grained details across different image resolutions, ensuring consistent performance. Consequently, the model's ability to handle mixed-exposure conditions is improved, as it can effectively manage diverse resolutions and adapt its feature extraction process dynamically, enhancing its robustness and accuracy in real-world applications.

3. Experiments and Implementation details

In this section, we evaluate our method through a series of experiments. We begin by outlining the datasets used, experimental setup, followed by a comparison of our method against state-of-the-art techniques using four quantitative metrics. We also perform a detailed ablation study (Appx E, Tab. 5) to analyze the components of our proposed method.

3.1. Datasets

To test the efficacy of our model, we evaluated ExpoMamba on four datasets: **(1) LOL** (Wei et al., 2018a), which has v1 and v2 versions. LOLv2 (Yang et al., 2020a) is divided into real and synthetic subsets. The training and testing sets are split into 485/15, 689/100, and 900/100 on LOLv1, LOLv2-real, and LOLv2-synthetic with $3 \times 400 \times 600$ resolution images. **(2) LOL4K** is an ultra-high definition dataset with $3 \times 3,840 \times 2,160$ resolution images, containing 8,099 pairs of low-light/normal-light images, split into 5,999 pairs for training and 2,100 pairs for testing. **(3) SICE** (Cai et al., 2018) includes 4,800 images, real and synthetic, at various exposure levels and resolutions, divided into training, validation, and testing sets in a 7:1:2 ratio.

Table 2. Evaluation on the UHD-LOL4K dataset. Symbols †, ‡, §, △, and ⋆ denote traditional, supervised CNN, unsupervised CNN, zero-shot, and transformer-based models, respectively.

Methods	UHD-LOL4K			
	PSNR ↑	SSIM ↑	LPIPS ↓	MAE ↓
BIMEF† (Ying et al., 2017)	18.1001	0.8876	0.1323	0.1240
LIME† (Guo et al., 2016)	16.1709	0.8141	0.2064	0.1285
NPE† (Wang et al., 2013)	17.6399	0.8665	0.1753	0.1125
SRIE† (Fu et al., 2016b)	16.7730	0.8365	0.1495	0.1416
MSRCR† (Jobson et al., 1997)	12.5238	0.8106	0.2136	0.2039
RetinexNet‡ (Wei et al., 2018b)	21.6702	0.9086	0.1478	0.0690
DSLR‡ (Lim & Kim, 2020)	27.3361	0.9231	0.1217	0.0341
KinD‡ (Zhang et al., 2019b)	18.4638	0.8863	0.1297	0.1060
Z_DCE§ (Guo et al., 2020a)	17.1873	0.8498	0.1925	0.1465
Z_DCE++§ (Li et al., 2021)	15.5793	0.8346	0.2223	0.1701
RUAS△ (Liu et al., 2021c)	14.6806	0.7575	0.2736	0.1690
ELGAN△ (Jiang et al., 2021)	18.3693	0.8642	0.1967	0.1011
Uformer⋆ (Wang et al., 2022b)	29.9870	0.9804	0.0342	0.0376
Restormer⋆ (Zamir et al., 2022)	36.9094	0.9881	0.0226	0.0117
LLFormer⋆ (Wang et al., 2023b)	37.3340	0.9862	0.0200	0.0116
UHD-Four (Li et al., 2023)	35.1010	0.9901	0.0210	-
ExpoMamba$_s$	28.3300	0.9730	0.0820	0.0315
ExpoMamba$_l$	35.2300	0.9890	0.0630	0.0451

We use dynamic adjustment for both 's' and 'l' ExpoMamba models during inference.

Table 3. Results for our Exposure Mamba approach over SICE-v2 (Cai et al., 2018) datasets.

Method	SICE-v2						#params
	Underexposure		Overexposure		Average		
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	
HE (Pitas, 2000)	14.69	0.5651	12.87	0.4991	13.78	0.5376	-
CLAHE (Reza, 2004)	12.69	0.5037	10.21	0.4847	11.45	0.4942	-
RetinexNet (Wei et al., 2018a)	12.94	0.5171	12.87	0.5252	12.90	0.5212	0.84M
URetinexNet (Wu et al., 2022)	12.39	0.5444	7.40	0.4543	12.40	0.5496	1.32M
Zero-DCE (Guo et al., 2020b)	16.92	0.6330	7.11	0.4292	12.02	0.5211	0.079M
Zero-DCE++ (Li et al., 2021)	11.93	0.4755	6.88	0.4088	9.41	0.4422	0.010M
DPED (Ignatov et al., 2017)	16.83	0.6133	7.99	0.4300	12.41	0.5217	0.39M
KIND (Zhang et al., 2019a)	15.03	0.6700	12.67	0.6700	13.85	0.6700	0.59M
DeepUPE (Wang et al., 2019)	16.21	0.6807	11.98	0.5967	14.10	0.6387	7.79M
SID (Chen et al., 2018a)	19.51	0.6635	16.79	0.6444	18.15	0.6540	-
SID-ENC (Huang et al., 2022)	21.36	0.6652	19.38	0.6843	20.37	0.6748	-
SID-L (Huang et al., 2022)	19.43	0.6644	17.00	0.6495	18.22	0.6570	11.56M
RUAS (Liu et al., 2021a)	16.63	0.5589	4.54	0.3196	10.59	0.4394	0.0014M
SCI (Ma et al., 2022)	17.86	0.6401	4.45	0.3629	12.49	0.5051	0.0003M
MSEC (Afifi et al., 2021)	19.62	0.6512	17.59	0.6560	18.58	0.6536	7.04M
CMEC (Nsampi et al., 2021)	17.68	0.6592	18.17	0.6811	17.93	0.6702	5.40M
LCDPNet (Wang et al., 2022a)	17.45	0.5622	17.04	0.6463	17.25	0.6043	0.96M
DRBN (Yang et al., 2020b)	17.96	0.6767	17.33	0.6828	17.65	0.6798	0.53M
DRBN+ERL (Huang et al., 2023)	18.09	0.6735	17.93	0.6866	18.01	0.6796	0.53M
DRBN-ERL+ENC (Huang et al., 2023)	22.06	0.7053	19.50	0.7205	20.78	0.7129	0.58M
ELCNet (Huang & Belongie, 2017)	22.05	0.6893	19.25	0.6872	20.65	0.6861	0.018M
ELCNet+ERL (Huang et al., 2023)	22.14	0.6908	19.47	0.6982	20.81	0.6945	0.018M
FECNet (Huang et al., 2019)	22.01	0.6737	19.91	0.6961	20.96	0.6849	0.15M
FECNet+ERL (Huang et al., 2023)	22.35	0.6671	20.10	0.6891	21.22	0.6781	0.15M
IAT (Cui et al., 2022)	21.41	0.6601	22.29	0.6813	21.85	0.6707	0.090M
ExpoMamba$_s$	22.59	0.7161	20.62	0.7392	21.61	0.7277	41M

Our 's': smallest model outperforms all the baselines.

3.2. Experimental setting

The proposed network is a single-stage end-to-end training model. The patch sizes are set to 128×128, 256×256, and 324×324 with checkpoint restarts and batch sizes of 8, 6, and 4, respectively, in consecutive runs. For dynamic patch training, we use different patch sizes simultaneously. The optimizer is `RMSProp` with a learning rate of 1×10^{-4}, a weight decay of 1×10^{-7}, and momentum of 0.9. A linear warm-up cosine annealing (Loshchilov & Hutter, 2016) scheduler with 15 warm-up epochs is used, starting with a learning rate of 1×10^{-4}. All experiments were carried out using the PyTorch library (Paszke et al., 2019) on an NVIDIA A10G GPU.

Loss functions. To optimize our ExpoMamba model we use a set of loss functions:

$$\mathbf{L} = \mathbf{L}_{l1} + \mathbf{L}_{vgg} + \mathbf{L}_{ssim} + \mathbf{L}_{lpips} + \lambda \cdot \mathbf{L}_{overexposed} \tag{7}$$

4. Results

The best performance for Tab. 1, Tab. 2, and Tab. 3 are marked with Red, Green, and Blue, respectively.

Tab. 1 compares our performance to 31 state-of-the-art baselines, including lightweight and heavy models. We evaluate ExpoMamba's performance using `SSIM`, `PSNR`, `LPIPS`, and `FID`. ExpoMamba achieves an inference time of **36 ms**, faster than most baselines (Fig. 1) and the fastest among comparable models. Models like DiffLL (Jiang et al., 2023), CIDNet (Feng et al., 2024), and LLformer (Wang et al., 2023b) have comparable results but much longer inference times. Traditional algorithms (e.g., MSRCR (Jobson et al., 1997), MF (Fu et al., 2016a), BIMEF (Ying et al., 2017), SRIE (Fu et al., 2016b), FEA (Dong et al., 2011), NPE (Wang et al., 2013), LIME (Guo et al., 2016)) generally perform poorly on LOL4K (Tab. 2). Fig. 1.b shows that increasing image resolution to 4K significantly increases inference time for transformer models due to their quadratic complexity. Despite being a 41 million parameter model, ExpoMamba demonstrates remarkable storage efficiency, consuming $\sim 1/4^{th}$ memory (2923 Mb) compared to CIDNet, which, despite its smaller size of 1.9 million parameters, consumes 8249 Mb. This is because ExpoMamba's state expansion fits inside the GPU's high-bandwidth memory and removes the quadratic bottleneck which significantly reduces memory footprint. Current SOTA models CIDNet (Feng et al., 2024) and LLformer (Wang et al., 2023b) are slower and less memory-efficient.

5. Conclusion

We introduced *ExpoMamba*, a model designed for efficient and effective low-light image enhancement. By integrating frequency state-space components within a U-Net variant, *ExpoMamba* leverages spatial and frequency domain processing to address computational inefficiencies and high-resolution challenges. Our approach combines robust feature extraction of state-space models, enhancing low-light images with high fidelity and achieving impressive inference speeds. Our novel dynamic patch training strategy significantly improves robustness and adaptability to real-world hardware constraints, making it suitable for real-time applications on edge devices. Experimental results show that *ExpoMamba* is much faster and comparably better than numerous existing transformer and diffusion models, setting a new benchmark in low light image enhancement.

References

Adhikarla, E., Zhang, K., Yu, J., Sun, L., Nicholson, J., and Davison, B. D. Robust computer vision in an ever-changing world: A survey of techniques for tackling distribution shifts, 2023. URL https://arxiv.org/abs/2312.01540.

Adhikarla, E., Zhang, K., VidalMata, R. G., Aithal, M., Madhusudhana, N. A., Nicholson, J., Sun, L., and Davison, B. D. Unified-egformer: Exposure guided lightweight transformer for mixed-exposure image enhancement, 2024. URL https://arxiv.org/abs/2407.13170.

Afifi, M., Derpanis, K. G., Ommer, B., and Brown, M. S. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9157–9167, 2021.

Brateanu, A., Balmez, R., Avram, A., and Orhei, C. Lyt-net: Lightweight yuv transformer-based network for low-light image enhancement. *arXiv preprint arXiv:2401.15204*, 2024.

Cai, J., Gu, S., and Zhang, L. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

Chen, C., Chen, Q., Xu, J., and Koltun, V. Learning to see in the dark. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3291–3300. Computer Vision Foundation / IEEE Computer Society, 2018a. doi: 10.1109/CVPR.2018.00347. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Learning_to_See_CVPR_2018_paper.html.

Chen, C., Chen, Q., Xu, J., and Koltun, V. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3291–3300, 2018b.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., and Gao, W. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12299–12310, 2021a.

Chen, Y., Zeng, Q., Ji, H., and Yang, Y. Skyformer: Remodel self-attention with gaussian kernel and nystr\"om method. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2122–2135. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/10a7cdd970fe135cf4f7bb55c0e3b59f-Paper.pdf.

Chen, Y., Zhang, S., Liu, F., Chang, Z., Ye, M., and Qi, Z. Transhash: Transformer-based hamming hashing for efficient image retrieval. *CoRR*, abs/2105.01823, 2021c. URL https://arxiv.org/abs/2105.01823.

Chiu, C.-C. and Ting, C.-C. Contrast enhancement algorithm based on gap adjustment for histogram equalization. *Sensors*, 16(6):936, 2016.

Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., Qiao, Y., and Harada, T. You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0238.pdf.

Dale-Jones, R. and Tjahjadi, T. A study and modification of the local histogram equalization algorithm. *Pattern Recognition*, 26(9):1373–1381, 1993.

Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., and Lu, Y. Fast efficient algorithm for enhancement of low lighting video. In *ICME*, pp. 1–6, 2011.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Duman Keles, F., Wijewardena, P. M., and Hegde, C. On the computational complexity of self-attention. In Agrawal, S. and Orabona, F. (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 597–619. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/duman-keles23a.html.

Feng, Y., Zhang, C., Wang, P., Wu, P., Yan, Q., and Zhang, Y. You only need one color space: An efficient network for low-light image enhancement, 2024.

Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., and Ré, C. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Fu, X., Zeng, D., Huang, Y., Liao, Y., Ding, X., and Paisley, J. A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, 129:82–96, 2016a.

Fu, X., Zeng, D., Huang, Y., Zhang, X.-P., and Ding, X. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, pp. 2782–2790, 2016b.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Guo, C., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., and Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, pp. 1780–1789, 2020a.

Guo, C. G., Li, C., Guo, J., Loy, C. C., Hou, J., Kwong, S., and Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1780–1789, June 2020b.

Guo, L., Wang, C., Yang, W., Huang, S., Wang, Y., Pfister, H., and Wen, B. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14049–14058, 2023.

Guo, X., Li, Y., and Ling, H. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016.

Guo, Z., Perminov, S., Konenkov, M., and Tsetserukou, D. Hawkdrive: A transformer-driven visual perception system for autonomous driving in night scene. *arXiv preprint arXiv:2404.04653*, 2024.

Hu, L., Chen, H., and Allebach, J. P. Joint multi-scale tone mapping and denoising for hdr image enhancement. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 729–738, 2022. doi: 10.1109/WACVW54805.2022.00080.

Huang, J., Xiong, Z., Fu, X., Liu, D., and Zha, Z.-J. Hybrid image enhancement with progressive laplacian enhancing unit. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pp. 1614–1622, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350855. URL https://doi.org/10.1145/3343031.3350855.

Huang, J., Liu, Y., Fu, X., Zhou, M., Wang, Y., Zhao, F., and Xiong, Z. Exposure normalization and compensation for multiple-exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6043–6052, 2022.

Huang, J., Zhao, F., Zhou, M., Xiao, J., Zheng, N., Zheng, K., and Xiong, Z. Learning sample relationship for exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9904–9913, 2023.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization, 2017.

Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., and Van Gool, L. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Jiang, H., Luo, A., Fan, H., Han, S., and Liu, S. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., and Wang, Z. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 30:2340–2349, 2021.

Jobson, D. J., Rahman, Z.-u., and Woodell, G. A. A multi-scale retinex for bridging the gap between color images and the human observation of scenes. *IEEE TIP*, 6(7): 965–976, 1997.

Kansal, S., Purwar, S., and Tripathi, R. K. Image contrast enhancement using unsharp masking and histogram equalization. *Multimedia Tools and Applications*, 77: 26919–26938, 2018.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/2006.16236.

Khan, M. F., Khan, E., and Abbasi, Z. A. Segment dependent dynamic multi-histogram equalization for image contrast enhancement. *Digital Signal Processing*, 25: 198–223, 2014.

Kitaev, N., Kaiser, L., and Levskaya, A. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Land, E. H. and McCann, J. J. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

Lazzarini, V. *Frequency-Domain Techniques*, pp. 223–271. Springer International Publishing, Cham, 2017. ISBN 978-3-319-63504-0. doi: 10.1007/978-3-319-63504-0_7. URL https://doi.org/10.1007/978-3-319-63504-0_7.

Li, C., Guo, C. G., and Loy, C. C. Learning to enhance low-light image via zero-reference deep curve estimation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3063604.

Li, C., Guo, C.-L., Zhou, M., Liang, Z., Zhou, S., Feng, R., and Loy, C. C. Embedding fourier for ultra-high-definition low-light image enhancement. *arXiv preprint arXiv:2302.11831*, 2023.

Liba, O., Murthy, K., Tsai, Y.-T., Brooks, T., Xue, T., Karnad, N., He, Q., Barron, J. T., Sharlet, D., Geiss, R., et al. Handheld mobile photography in very low light. *ACM Trans. Graph.*, 38(6):164–1, 2019.

Lim, S. and Kim, W. Dslr: Deep stacked laplacian restorer for low-light image enhancement. *IEEE TMM*, 23:4272–4284, 2020.

Lin, Z., He, Z., Wang, P., Tan, B., Lu, J., and Bai, Y. Snr-net: A deep learning-based network for banknote serial number recognition. *Neural Processing Letters*, 52:1415–1426, 2020.

Liu, J., Dejia, X., Yang, W., Fan, M., and Huang, H. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 129:1153–1184, 2021a. doi: 10.1007/s11263-020-01418-8.

Liu, L., Qu, Z., Chen, Z., Ding, Y., and Xie, Y. Transformer acceleration with dynamic sparse attention. *CoRR*, abs/2110.11299, 2021b. URL https://arxiv.org/abs/2110.11299.

Liu, R., Ma, L., Zhang, J., Fan, X., and Luo, Z. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, pp. 10561–10570, 2021c.

Liu, X., Wu, Z., Li, A., Vasluianu, F.-A., Zhang, Y., Gu, S., Zhang, L., Zhu, C., Timofte, R., Jin, Z., et al. Ntire 2024 challenge on low light image enhancement: Methods and results. *arXiv preprint arXiv:2404.14248*, 2024.

Liu, Y. Design of a two-branch network enhancement algorithm for deep features in visually communicated images. *Signal, Image and Video Processing*, pp. 1–12, 2024.

Lore, K. G., Akintayo, A., and Sarkar, S. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Lu, J., Yao, J., Zhang, J., Zhu, X., Xu, H., Gao, W., XU, C., Xiang, T., and Zhang, L. Soft: Softmax-free transformer with linear complexity. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21297–21309. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b1d10e7bafa4421218a51b1e1f1b0ba2-Paper.pdf.

Ma, L., Ma, T., Liu, R., Fan, X., and Luo, Z. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5637–5646, June 2022.

Mehta, R. and Sivaswamy, J. M-net: A convolutional neural network for deep brain structure segmentation. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pp. 437–440. IEEE, 2017.

Morikawa, C., Kobayashi, M., Satoh, M., Kuroda, Y., Inomata, T., Matsuo, H., Miura, T., and Hilaga, M. Image and video processing on mobile devices: a survey. *the visual Computer*, 37(12):2931–2949, 2021.

Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., and Ré, C. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022.

Nsampi, N. E., Hu, Z., and Wang, Q. Learning exposure correction via consistency modeling. In *Proc. Brit. Mach. Vision Conf.*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pitas, I. *Digital Image Processing Algorithms and Applications*. John Wiley & Sons, Inc., USA, 1st edition, 2000. ISBN 0471377392.

Pokle, A., Geng, Z., and Kolter, J. Z. Deep equilibrium approaches to diffusion models. *Advances in Neural Information Processing Systems*, 35:37975–37990, 2022.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., and Jagersand, M. U2-net: Going deeper with nested u-structure for salient object detection. In *Pattern Recognition 2020*, volume 106, pp. 107404, 2020.

Ren, X., Yang, W., Cheng, W.-H., and Liu, J. Lr3m: Robust low-light enhancement via low-rank regularized retinex model. *IEEE Transactions on Image Processing*, 29: 5862–5876, 2020.

Reza, A. M. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44, 2004. URL https://api.semanticscholar.org/CorpusID:41505656.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. Efficient attention: Attention with linear complexities. *CoRR*, abs/1812.01243, 2018. URL http://arxiv.org/abs/1812.01243.

Shrivastav, P. Advancements and challenges in low-light object detection. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pp. 1351–1356. IEEE, 2024.

Singh, K., Kapoor, R., and Sinha, S. K. Enhancement of low exposure images via recursive histogram equalization algorithms. *Optik*, 126(20):2619–2625, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. URL https://arxiv.org/abs/2010.02502.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6), dec 2022. ISSN 0360-0300. doi: 10.1145/3530811. URL https://doi.org/10.1145/3530811.

Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J. F., Mehri, S., Rostamzadeh, N., Bengio, Y., and Pal, C. J. Deep complex networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1T2hmZAb.

Valin, J. The daala directional deringing filter. *CoRR*, abs/1602.05975, 2016. URL http://arxiv.org/abs/1602.05975.

Wang, H., Xu, K., and Lau, R. W. Local color distributions prior for image enhancement. In *European Conference on Computer Vision*, pp. 343–359. Springer, 2022a.

Wang, R., Zhang, Q., Fu, C.-W., Shen, X., Zheng, W.-S., and Jia, J. Underexposed photo enhancement using deep illumination estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Wang, S., Zheng, J., Hu, H.-M., and Li, B. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE TIP*, 22(9):3538–3548, 2013.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. Linformer: Self-attention with linear complexity, 2020.

Wang, T., Zhang, K., Shao, Z., Luo, W., Stenger, B., Kim, T., Liu, W., and Li, H. Lldiffusion: Learning degradation representations in diffusion models for low-light image enhancement. *CoRR*, abs/2307.14659, 2023a. doi: 10.48550/ARXIV.2307.14659. URL https://doi.org/10.48550/arXiv.2307.14659.

Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., and Lu, T. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2654–2662, 2023b.

Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.-P., Kot, A. C., and Wen, B. Exposurediffusion: Learning to expose for low-light image enhancement. *arXiv preprint arXiv:2307.07710*, 2023c.

Wang, Z., Cun, X., Bao, J., and Liu, J. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pp. 17683–17693, 2022b.

Wang, Z., Zheng, J.-Q., Zhang, Y., Cui, G., and Li, L. Mamba-unet: Unet-like pure visual mamba for medical image segmentation, 2024.

Wei, C., Wang, W., Yang, W., and Liu, J. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, pp. 155. BMVA Press, 2018a. URL http://bmvc2018.org/contents/papers/0451.pdf.

Wei, C., Wang, W., Yang, W., and Liu, J. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018b.

Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., and Jiang, J. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5891–5900, 2022. doi: 10.1109/CVPR52688.2022.00581.

Xian, X., Zhou, Q., Qin, J., Yang, X., Tian, Y., Shi, Y., and Tian, D. Crose: Low-light enhancement by cross-sensor interaction for nighttime driving scenes. *Expert Systems with Applications*, pp. 123470, 2024.

Xiao, Z. and Hou, Z. Phase based feature detector consistent with human visual system characteristics. *Pattern Recognition Letters*, 25(10):1115–1121, 2004.

Yang, W., Wang, S., Fang, Y., Wang, Y., and Liu, J. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Yang, W., Wang, S., Fang, Y., Wang, Y., and Liu, J. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3063–3072, 2020b.

Yang, Y., Xing, Z., and Zhu, L. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*, 2024.

Ying, Z., Li, G., and Gao, W. A bio-inspired multi-exposure fusion framework for low-light image enhancement. *arXiv preprint arXiv:1711.00591*, 2017.

Yuan, S., Li, J., Ren, L., and Chen, Z. Multi-frequency field perception and sparse progressive network for low-light image enhancement. *Journal of Visual Communication and Image Representation*, 100:104133, 2024.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. Big bird: Transformers for longer sequences. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pp. 5728–5739, 2022.

Zhang, Q., Tao, M., and Chen, Y. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022.

Zhang, R., Guo, L., Huang, S., and Wen, B. Rellie: Deep reinforcement learning for customized low-light image enhancement. *CoRR*, abs/2107.05830, 2021a. URL https://arxiv.org/abs/2107.05830.

Zhang, Y., Zhang, J., and Guo, X. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pp. 1632–1640, New York, NY, USA, 2019a. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350926. URL https://doi.org/10.1145/3343031.3350926.

Zhang, Y., Zhang, J., and Guo, X. Kindling the darkness: A practical low-light image enhancer. In *ACMMM*, pp. 1632–1640, 2019b.

Zhang, Y., Guo, X., Ma, J., Liu, W., and Zhang, J. Beyond brightening low-light images. *Int. J. Comput. Vision*, 129(4):1013–1037, apr 2021b. ISSN 0920-5691. doi: 10.1007/s11263-020-01407-x. URL https://doi.org/10.1007/s11263-020-01407-x.

Zheng, Z. and Zhang, J. Fd-vision mamba for endoscopic exposure correction, 2024.

Zhou, D., Yang, Z., and Yang, Y. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028*, 2023a.

Zhou, M., Huang, J., Li, C., Yu, H., Yan, K., Zheng, N., and Zhao, F. Adaptively learning low-high frequency information integration for pan-sharpening. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3375–3384, 2022.

Zhou, M., Huang, J., Guo, C.-L., and Li, C. Fourmer: An efficient global modeling paradigm for image restoration. In *International Conference on Machine Learning*, pp. 42589–42601. PMLR, 2023b.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., and Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024a.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., and Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024b.

Appendix

A. Related Work

A.1. Conventional and Deep CNN-Based Methods

Traditional methods for low-light image enhancement often rely on histogram equalization (HE) (Dale-Jones & Tjahjadi, 1993; Singh et al., 2015; Khan et al., 2014) and Retinex theory (Land & McCann, 1971; Ren et al., 2020). HE based methods aim to adjust the contrast of the image by uniformly distributing the pixel intensities, which can sometimes lead to overenhancement and noise amplification, which were later investigated more carefully by CegaHE (Chiu & Ting, 2016), UMHE (Kansal et al., 2018), etc. Retinex theory, which decomposes an image into illumination and reflectance, provides a more principled approach to enhancement but still faces limitations in complex lighting conditions.

Convolutional Neural Networks (CNNs) have significantly advanced this field. Early works like LLNet (Lore et al., 2017) used autoencoders to enhance low-light image visibility. The SID (See-in-the-Dark) network (Chen et al., 2018b) leveraged raw image data for better enhancement by training on paired low-light and normal-light images. Other works in paired training include DSLR (Lim & Kim, 2020), DRBN (Yang et al., 2020b), KinD (Zhang et al., 2019a), KinD++ (Zhang et al., 2021b), MIRNet (Zamir et al., 2020), ReLLIE (Zhang et al., 2021a), DDIM (Song et al., 2020), SCI (Ma et al., 2022), RAUS (Liu et al., 2021a), Restormer (Zamir et al., 2022), CIDNet (Feng et al., 2024), LLFormer (Wang et al., 2023b), SNRNet (Lin et al., 2020), Uformer (Wang et al., 2022b), and CDEF (Valin, 2016). Methods like RetinexNet (Wei et al., 2018b), which decompose images into illumination and reflectance components, also show considerable promise but often struggle with varying lighting conditions.

A.2. Foundation Models in LLIE

Transformer Models. Such approaches have gained popularity for modeling long-range dependencies in images. LLFormer (Wang et al., 2023b) leverages transformers for low-light enhancement by focusing on global context, significantly improving image quality. Fourmer (Zhou et al., 2023b) introduces a Fourier transform-based approach within the transformer architecture, while IAT (Cui et al., 2022) adapts ISP-related parameters to address low-level and high-level vision tasks. IPT (Chen et al., 2021a) uses a multi-head, multi-tail shared pre-trained transformer module for image restoration. LYT-Net (Brateanu et al., 2024) addresses image enhancement with minimal computing resources by using YUV colorspace for transformer models. Despite their effectiveness, these transformer models often require substantial computational resources, limiting their practicality on edge devices.

Diffusion Models. Diffusion models have shown great potential in generating realistic and detailed images. The ExposureDiffusion model (Wang et al., 2023c) integrates a diffusion process with a physics-based exposure model, enabling accurate noise modeling and enhanced performance in low-light conditions. Pyramid Diffusion (Zhou et al., 2023a) addresses computational inefficiencies by introducing a pyramid resolution approach, speeding up enhancement without sacrificing quality. (Saharia et al., 2022) handles image-to-image tasks using conditional diffusion processes. Models like (Zhang et al., 2022) and deep non-equilibrium approaches (Pokle et al., 2022) aim to reduce sampling steps for faster inference. However, starting from pure noise in conditional image restoration tasks remains a challenge for maintaining image quality while cutting down inference time (Guo et al., 2023).

Hybrid Modelling. Hybrid models includes learning features in both spatial and frequency domain has been another popular area in image enhancement/restoration tasks. Mostly it has been explored in three sub-categories: **(1)** Fourier Transform (Yuan et al., 2024), Fourmer (Zhou et al., 2023b), FD-VisionMamba (Zheng & Zhang, 2024); **(2)** Wavelet Transform; **(3)** Homomorphic Filtering (). Such methods demonstrate that leveraging both spatial and frequency information can significantly improve enhancement performance.

State-Space Models. Recent advancements reveal the efficacy of state space models (SSM) as a robust architecture in foundation model era for sequence modeling, offering a fresh perspective beyond conventional RNNs, CNNs, and Transformers. Pioneering this shift, the S4 (Gu et al., 2021) model demonstrated superior performance in managing long-range dependencies by employing the HiPPO matrix (Fu et al., 2022) to define state dynamics systematically. Initially introduced for audio processing, SSMs have emerged as a alternative, later expanded into language and vision domains for handling long-range model dependencies and temporal dynamics becoming a strong competitor for current transformer based methods. The V-Mamba architecture (Zhu et al., 2024b; Yang et al., 2024) combines state-space models with U-Net frameworks to capture detailed image aspects at multiple scales, proving effective in biomedical image segmentation.

Furthermore, the S4 architecture (Gu et al., 2021; Nguyen et al., 2022) extends this idea by incorporating linear state-space models for fast and efficient sequence modeling, making it suitable for real-time applications.

B. The Importance of Inference Time over FLOPs in Real-World Applications

In our paper, we use inference time as a measure because inference time, unlike the abstract measure of FLOPs (Floating Point Operations Per Second), reflects actual performance in real-world applications, being influenced not only by hardware speed but also by model design and optimization.

In practical scenarios, wherein systems requiring real-time processing like autonomous vehicles and interactive AI applications, the agility of model inference directly impacts usability and user experience. Moreover, as inference constitutes the primary computational expense post-deployment, optimizing inference time enhances both the cost-effectiveness and the energy efficiency of AI systems. Thus, we focused on minimizing inference time, rather than merely reducing FLOPs, ensuring that AI models are not only theoretically efficient but are also pragmatically viable in dynamic real-world environments. We believe that this approach not only accelerates the adoption of AI technologies but also drives advancements in developing models that are both performant and sustainable.

C. Detailed Methodology

C.1. Proof of Phase Manipulation

For an image $I(x, y)$, its Fourier transform is given by:

$$\mathbf{F}(u, v) = \iint I(x, y) e^{-i2\pi(ux+vy)}\, dx\, dy \tag{8}$$

This can be decomposed into amplitude $A(u, v)$ and phase $\phi(u, v)$:

$$\mathbf{F}(u, v) = \mathbf{A}(u, v) e^{i\phi(u,v)} \tag{9}$$

The inverse Fourier transform, which reconstructs the image from its frequency representation, is:

$$\mathbf{I}(x, y) = \iint \mathbf{A}(u, v) e^{i\phi(u,v)} e^{i2\pi(ux+vy)}\, du\, dv \tag{10}$$

Suppose that the phase component $\phi(u, v)$ is uniformly shifted by an angle $\Delta\phi$, the new phase $\phi'(u, v) = \phi(u, v) + \Delta\phi$. The modified image $I'(x, y)$ with this new phase is represented as:

$$\mathbf{I}'(x, y) = \iint \mathbf{A}(u, v) e^{i(\phi(u,v)+\Delta\phi)} e^{i2\pi(ux+vy)}\, du\, dv \tag{11}$$

Using Euler's formula $e^{i\Delta\phi} = \cos(\Delta\phi) + i\sin(\Delta\phi)$, the equation becomes:

$$\mathbf{I}' = \iint \mathbf{A}(u, v) e^{i\phi(u,v)} (\cos(\Delta\phi) + i\sin(\Delta\phi)) e^{i2\pi(ux+vy)}\, du\, dv \tag{12}$$

Given that $\cos(\Delta\phi)$ and $\sin(\Delta\phi)$ are constants for a particular $\Delta\phi$, they can be factored out of the integral:

$$\mathbf{I}'(x, y) = \cos(\Delta\phi) \cdot I(x, y) + i\sin(\Delta\phi) \cdot \tag{13}$$

$$\iint A(u, v) e^{i\phi(u,v)} e^{i2\pi(ux+vy)}\, du\, dv \tag{14}$$

This shows that the new image $\mathbf{I}'(x, y)$ is a linear combination of the original image $\mathbf{I}(x, y)$ and another image derived from the same amplitude and a phase-shifted version of the original phase components. The transformation demonstrates that even a constant shift in the phase component translates into a significant transformation in the spatial domain, affecting the structural layout and visual features of the image.

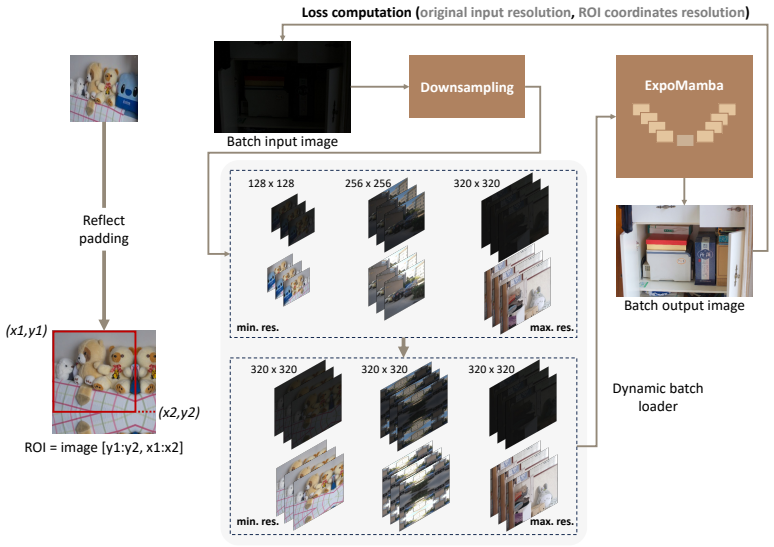

Figure 5. The downsampled images are prepared in multiple different training resolutions with padding to dynamically load the batched-images of different resolutions.

C.2. Model Robustness through Dynamic Patch Training

To address the hardware constraints in real-time scenarios such as phones or laptop webcams, which often adjust camera resolutions to optimize performance within design and battery limits, there is a critical need for models that dynamically adapt to these variations. Feeding various image resolutions to the model dynamically also helps avoid spurious correlations that are formed due to strong correlation (Adhikarla et al., 2023) in the data distribution of certain types of images. For instance, the SICEv2 dataset has relatively more mixed-exposure images, and the borders of sudden changes in exposure become more prone to spurious correlations. However, our ExpoMamba uses spatial and temporal components that are inherently designed in Vision Mamba (Zhu et al., 2024a) to handle both the spatial distribution of pixels in images and the temporal sequence of frames in videos.

C.3. Dynamic Adjustment Approximation

The Dynamic Adjustment Approximation module offers a unique way to enhance images without needing ground truth *mean* and *variance*. Instead, it dynamically adjusts brightness by using the image's own statistical properties, specifically the *median* and *mean* pixel values. Unlike previous models like KinD, LLFlow, RetinexFormer, which relied on static adjustment factors from the ground truth and often produced less accurate results otherwise, our method calculates a desired shift based on the difference between a normalized brightness value and the image's mean. Then, it adjusts the image's medians toward this shift, taking both the current median and mean into account. This leads to a more balanced and natural enhancement. Adjustment factors are carefully computed to avoid infinite or undefined values, ensuring stability. This approach simplifies the process by not requiring ground-truth data and also improves the efficiency and effectiveness of image enhancement.

$$\text{adjustment\_factors} = \frac{\text{medians} + \text{strength} \times (\text{normalized\_value} - \text{means})}{\text{medians}} \tag{15}$$

$$\text{adjusted\_img} = \text{input\_img} \times \text{adjustment\_factors} \tag{16}$$

C.4. Model Configuration

This model configuration table provides a detailed comparison between the two variants of ExpoMamba, highlighting their configurations and performance metrics. Notably, despite an increase of 125 million parameters, the memory consumption of the larger $\text{ExpoMamba}_{\text{large}}$ variant is 5690 Mb, which is a modest increase compared to transformer-based models.

Table 4. We describe two variants of our model, **s**' and **l**' represent small and large model configurations.

Model Type	Configuration					
	base channel	patch size	depth	params	inference speed	Memory consumption
ExpoMamba$_{\text{small}}$	48	4	1	41 M	36 ms	2923 Mb
ExpoMamba$_{\text{large}}$	96	6	4	166 M	95.6 ms	5690 Mb

C.5. Algorithm

The following pseudocode presents the details of ExpoMamba training with FSSB blocks:

Algorithm 1 ExpoMamba Training with Frequency State Space Block (FSSB)

Input: Low-light image dataset $\mathbf{D}$, Frequency State Space Block $\mathbf{FSSB}$, Visual SSM blocks $\mathbf{VSSM_A}$ and $\mathbf{VSSM_P}$, HDR layer $\mathbf{HDR}$, ComplexConv layer $\mathbf{CC}$, Training epochs E
Output: Model parameters θ
Fourier Transform:
Initialize frequency components $\mathbf{F}(u, v)$
for each image $I \in \mathbf{D}$ **do**
  Compute $\mathbf{F}(u, v)$ via Fourier Transform: $\mathbf{F}(u, v) = \iint \mathbf{I}(x, y) e^{-i2\pi(ux+vy)} dx \cdot dy$
end for
Frequency State Space Processing:
for each frequency component $\mathbf{F}(u, v)$ **do**
  Decompose into amplitude $\mathbf{A}(u, v)$ and phase $\mathbf{P}(u, v)$
  Process $\mathbf{A}(u, v)$ and $\mathbf{P}(u, v)$ via VSSM blocks: $\mathbf{h}[t+1] = \mathbf{A}[t] \cdot \mathbf{h}[t] + \mathbf{B}[t].x[t]\ \mathbf{y}[t] = \mathbf{C}[t] \cdot \mathbf{h}[t]$
end for
Inverse Fourier Transform:
for each processed component $\hat{\mathbf{F}}(u, v)$ **do**
  Recombine modified amplitude $\mathbf{A}"(\mathbf{u}, \mathbf{v})$ and phase $\mathbf{P}"(\mathbf{u}, \mathbf{v})$
  Transform back to spatial domain using inverse Fourier Transform: $\mathbf{I}'(\mathbf{x}, \mathbf{y}) = \mathbf{F^{-1}}(\mathbf{A}"(\mathbf{u}, \mathbf{v}) + i \cdot \mathbf{P}"(\mathbf{u}, \mathbf{v}))$
end for
Training:
for each epoch $e = 1, 2, \ldots, E$ **do**
  for each batch $\mathbf{b} \in \mathbf{D}$ **do**
    Forward pass through FSSB, HDR, and ComplexConv layers
    Compute loss $\mathcal{L}$
    Backpropagation and update model parameters
  end for
end for
Return: Model parameters θ

D. Loss function

The combined loss function as shown in Eq. 7, is designed to enhance image quality by addressing different aspects of image reconstruction. The `L1` loss ensures pixel-level accuracy, crucial for maintaining sharp edges. This loss component has been widely utilized by the low light papers and has proven to be a valuable loss component for training variety of image restoration tasks. `VGG` loss, leveraging high-level features, maintains perceptual similarity. `SSIM` loss preserves structural integrity and local visual quality, which is vital for a coherent visual experience. `LPIPS` loss focuses on perceptual differences to generate natural looking image. Additionally, the overexposed regularizer detects and penalizes overexposed areas, crucial for handling `HDR` content and preserving details. It works in combination with HDR blocks to suppress artifacts in overexposed areas and control enhancement. In Eq. 7, λ is the weight for the overexposed regularization term.

E. Ablation Study

We have performed the ablation study of our model ExpoMamba over LOL-v1 dataset. We used 'DoubleConv' Block instead of regular convolutional blocks in the regular U-Net/M-Net architecture. 'Block' represents the residual block inside every upsampling blocks. We implemented two variants of HDR layer, where HDR/HDROut represent the same single layer approach with different locations for layer placement. On the other hand, HDR-CSRNet+ is a deeper network originally design for congested scene recognition is used inside FSSB instead of simple HDR layer.

- DoubleConv: Its absence results in lower PSNR and SSIM scores, confirming its importance.

Table 5. Ablation Study on various components inside our proposed model ExpoMamba.

DoubleConv	Block	FSSB	HDR	HDR-CSRNet+	HDROut	DA	**PSNR**	**SSIM**
✓	✗	✗	✗	✗	✗	✗	18.978	0.815
✗	✓	✗	✗	✗	✗	✗	19.787	0.828
✗	✗	✓	✗	✗	✗	✗	22.459	0.836
✗	✗	✗	✓	✗	✗	✗	20.576	0.823
✓	✓	✓	✗	✗	✗	✗	24.878	0.841
✓	✓	✓	✓	✗	✓	✓	25.110	0.845
✓	✓	✓	✗	✓	✓	✓	**25.640**	**0.860**

When 'DoubleConv' is not used, we default to using the standard U^2-Net/M-Net architecture's 2D convolutional block.

- Block: Inclusion of residual blocks improves performance metrics.
- FSSB: Significantly enhances model performance, indicating its crucial role.
- HDR vs. HDR-CSRNet+ vs. HDROut:
- HDR: Provides notable improvements but is outperformed by HDR-CSRNet+.
- HDR-CSRNet+: Offers the best results among the HDR variants.
- HDROut: Slightly less effective than HDR-CSRNet+.
- DA (Dynamic Adjustment during inference): Consistently boosts the model's PSNR and SSIM slightly based on input mean value.