FF-Former: Swin Fourier Transformer for Nighttime Flare Removal

Dafeng Zhang[1*] Jia Ouyang[1*] Guanqun Liu[1] Xiaobing Wang[1] Xiangyu Kong[1] Zhezhu Jin[1]

[1] Samsung Research China - Beijing (SRC-B)

{dfeng.zhang, jia.ouyang, guanqun1.liu, x0106.wang, xiangyu.kong, zz777.jin}@samsung.com

Abstract

In the process of removing nighttime flare, it is crucial to have a large receptive field due to the fact that flare can occupy a substantial portion of an image, even potentially the entire image. However, the conventional window-based Transformer approaches restrict the receptive field within the window, limiting its ability to capture global features. And the flare can cause the dark regions to become brighter and result in a loss of contrast and alteration of the frequency characteristics of the image. To address these challenges, we introduce FF-Former, which is based on Fast Fourier Convolution (FFC) and is designed to extract global frequency features for enhancing nighttime flare removal. To achieve this, we incorporate a Spatial Frequency Block (SFB) after the Swin Transformer, which forms the Swin Fourier Transformer Block (SFTB). This configuration enables the establishment of long dependencies and the extraction of global features. Unlike the traditional Transformer, which relies on global self-attention, the SFB module only performs convolution computation, making it both effective and efficient. Additionally, during the training phase, we optimize the loss function to preserve the light source points after nighttime flare removal. Experimental results on both real-world and synthetic benchmarks demonstrate that the proposed FF-Former significantly improves the performance of nighttime flare removal.

1. Introduction

In theory, a perfect camera should be able to converge all rays from a single point source to a single focal point. However, in reality, lenses scatter and reflect light along unintended paths, leading to the appearance of halos that produce brightness in radial areas of the image. This phenomenon, referred to as flare, can negatively impact downstream visual tasks such as semantic segmentation and depth estimation. Therefore, a reliable flare removal algorithm is essential and has garnered significant attention both in industry and academia.

To address the negative impact of flare on image quality, some high-end cameras adopt advanced optical designs and materials that reduce the flare. Some lenses also add glass elements to minimize reflections from specular surfaces. An anti-reflective (AR) coating is a common solution, but it can be costly and only optimized for specific wavelengths and angles of light. In response to these limitations, many cost-effective software-based solutions have been developed to address the issue. These techniques involve detecting the flare based on its unique shape, location, or intensity, then using image patching to restore the affected areas. However, these methods are only effective for certain types of flare, such as bright spots.

AlexNet [12] revolutionized the application of AI, but learning-based flare removal algorithms have been rarely explored. This is primarily due to the challenge of collecting a large amount of perfectly aligned images with and without lens flare. However, recent advancements in semi-synthetic data based on physical principles, such as the flare7K dataset [8] introduced last year, provide a valuable benchmark for studying the complex task of nighttime flare removal. The recent launch of the Nighttime Flare Removal competition at the Mobile Intelligent Photography and Imaging Workshop 2023 has further fueled interest and development in deep learning-based algorithms for nighttime flare removal.

As we are well aware, Convolutional Neural Networks (CNN) have long been the backbone of computer vision algorithms [5, 10]. However, with the advent of the Transformer [23] framework in 2017 proposed by Google, its application has gradually spread to the field of computer vision. The recent advancements in the Transformer structure, such as the VIT [9] and Swin Transformer [14], have further solidified its position as the new dominant force in visual modeling. The key to the success of the Transformer lies in its attention mechanism and large receptive field, which has been proven to be critical for visual tasks through various studies. In particular, the flare removal task requires a large

*Equal contribution.

receptive field due to the extent of the flare, which often covers a large area or even the entire image. Thus, global information is crucial in accurately identifying the flare.

Based on this inspiration, we introduce FF-Former, a U-shape network based Fast Fourier Convolution (FFC) [7] for nighttime flare removal. To address the issue of insufficient receptive field in window-based Transformer, we present the Spatial Frequency Block (SFB) after Swin Transformer to analyze and perceive flare from a global perspective while retaining detailed information during image restoration. The SFB comprises of two branches, a spatial module and a frequency module, with FFC utilized in the frequency branch to extract global information and a residual module based on CNN in the spatial branch to enhance local detailed feature representation. Additionally, we optimize the loss function during the training phase to preserve light source points in the deflared image. The results of extensive experiments on both real-world and synthetic benchmarks show that our FF-Former outperforms current state-of-the-art (SOTA) methods in terms of nighttime flare removal performance.

Our contributions can be summarized as follows:

- We present a novel solution for nighttime flare removal, the FF-Former network, which addresses the issue of limited receptive field in traditional window-based Transformer approaches.
- We also enhance the performance by implementing the Light Source Mask Loss Function, which guarantees the preservation of the light source point even after the removal of flare.
- Comprehensive experiments conducted on both real-world and synthetic nighttime flare removal datasets demonstrate that our approach outperforms state-of-the-art (SOTA) methods in a significant manner.

2. Related Works

Now that the flare task has attracted much attention, we briefly review these methods. These methods are mainly divided into three categories: (a) Hardware Solutions (b) Software Solutions (c) Data-driven Solutions.

2.1. Hardware Solutions

Most hardware solutions focus on improving the camera's optics to eliminate flare, such as optimized lens barrel designs, lens hoods or reflective coatings. A widely used technique is to apply anti-reflective (AR) coatings on lens elements to reduce internal reflections by destroying interference, for example, Boynto *et al*. [2] built a fluid-filled camera, Raskar *et al*. [18] Inserting a transparent mask on top of the imaging sensor, Macleod *et al*. [15] replaced the circular polarizer with a neutral density filter, etc. however, AR coatings are expensive to add to all optical surfaces, and the thickness of such coatings can only be optimized for specific wavelengths and incident angles. In addition, they can only reduce flare during the capture process, but cannot deal with flare existing image, and these hardware solutions can hardly eliminate the entire flare artifacts [16, 18].

2.2. Software Solutions

In order to solve the above problems many software-based solutions were subsequently derived, but these methods are basically two-stage methods: first identify the flare, and then repair the scene of the halo area. For example, Chabert *et al*. [3] used a series of thresholds to binarize the image, calculated the contour features of the binarized image to obtain a series of potential flare candidate regions, then reconstructed these candidate regions. Vitoria *et al*. [21] detect flare points by overexposing features near the flare point and create a flare point mask to remove flare. Asha *et al*. [1] considered the bright spot problem that often appears in the background under the sun light source or /flashing light source, so as to detect the light source point, and then fill the bright spot area to repair the image. Due to so many variable, flare is often a difficult factor in image quality to measure. the above methods which based on hand-crafted features are only used for limited types of flares, it is easy to treat local bright areas as flares and it is difficult to distinguish different types of flares. Therefore, it is still unrealistic to simply remove them through physical algorithms.

2.3. Data-driven Solutions

Recently deep learning-based methods have achieved great success on various low-level vision tasks, but due to the difficulty of collecting large numbers of perfectly aligned images with and without lens flares, the development of learning-based flare removal algorithms has been slow, and only some related work has emerged in the last few years.

Wu *et al*. [24] proposed a synthesis method based on generating paired training data via light source-guided single-image flare removal (SIFR), but did not generalize well to real-world data. Qiao *et al*. [17] proposed a new learning framework to learn how to remove flare artifacts using unpaired data. Wu *et al*. [25] simulate the optical causes of flares, generate synthetic pairs of flare-damaged and clean images, making it possible to train neural networks to remove lens flare, and demonstrate that data synthesis methods are crucial for accurate flare removal Important. Last year, Dai *et al*. [8] produced the first nighttime flare removal benchmark dataset Flare7K dataset which provided a valuable benchmark for studying this challenging nighttime flare removal task.

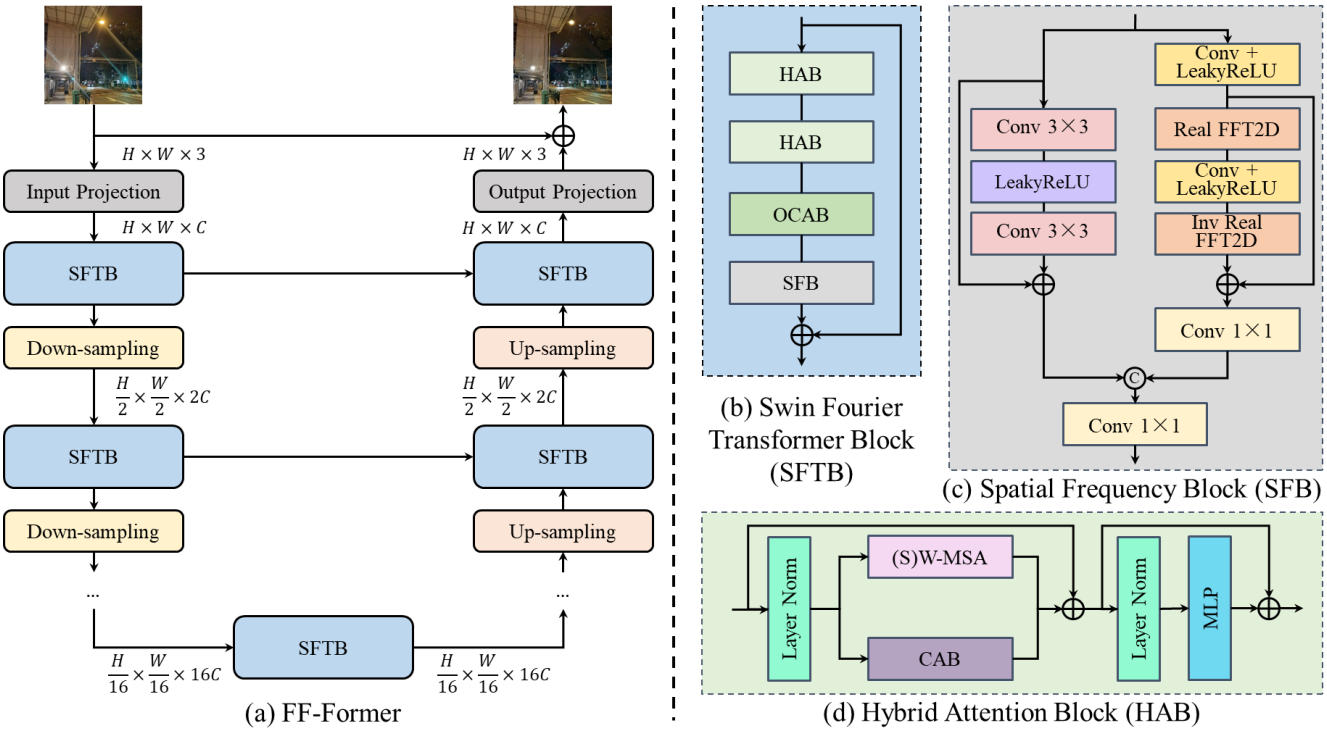


Figure 1. The network architecture of FF-Former is composed of multiple Swin Fourier Transformer Block (SFTB). SFTB is composed of Hybrid Attention Block (HAB), Overlapping Cross-Attention Block (OCAB) and a Spatial Frequency Block (SFB), in which the HAB and OCAB extract local features in the windows, while SFB extract global features via Fast Fourier Convolution.

3. Methodology

The Swin Transformer-based image restoration method achieves better performance than CNN. However, due to the high resolution of the image in the low-level task, only the window-based Transformer can be used to balance the computing resources, which limits the ability of the Transformer to extract global features. Therefore, in the image restoration task, the U-shaped network structure is favored by researchers. It extracts the global features by increasing the receptive field step by step through multiple down-sampling. Unfortunately, it is inevitable to lose the rich detail information of the input image after multiple down-sampling. In order to improve the ability of the model to extract global features and protect the details of the input image, we propose an U-shape network based Fast Fourier Convolution, named FF-Former, for nighttime flare removal, as shown in Figure 1. Specifically, we use the Swin Fourier Transformer Block (SFTB) to extract the global information of input features in the Encoder, Bottleneck and Decoder of FF-Former, while not losing the local detail information. SFTB is composed of Hybrid Attention Block (HAB), Overlapping Cross-Attention Block (OCAB) and a Spatial Frequency Block (SFB), in which the HAB extracts local features in the windows, while SFB uses Fast Fourier Convolution to extract global features.

In this paper, we input the nighttime flare-corrupted image $I_{FC} \in \mathbb{R}^{H\times W\times 3}$ into FF-Former and output one nighttime deflared image $I_{DF} \in \mathbb{R}^{H\times W\times 3}$.

Firstly, we project the input nighttime flare-corrupted images into high dimensional space to extract shallow feature $I_S \in \mathbb{R}^{H\times W\times C}$ by using a simple 3×3 convolutional layer with $LeakyReLU$. The projection can be formulated as,

$$F_S = \sigma(Conv(I_{FC})) \tag{1}$$

where C denotes the channel number of the shallow feature. $H \times W$ denotes the spatial dimension. $Conv(\cdot)$ represents 3×3 convolutional layer, $\sigma(\cdot)$ is the nonlinear activation function and $\sigma = LeakyReLU()$ in this paper. Then, we take the shallow feature as input of Encoder to extract the multi-scale features $F_{E_i} \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^i C}$ and bottleneck feature $F_{BF} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 16C}$,

$$F_{BF} = Encoder(F_S) \tag{2}$$

$$F_{E_{i+1}} = Down^{\downarrow 2^i}(SFTB(F_{E_i})), i \in 1, 2, 3, 4 \tag{3}$$

where $Encoder(\cdot)$ represents Encoder module of FF-Former, which is composed of 4-level $SFTB(\cdot)$ and down-sampling module $Down^{\downarrow 2^i}(\cdot)$. And the spatial dimension decreases gradually as the level increases, while number of

channels increases. $F_{BF} = F_{E_3}$ denotes the input feature with flare of Bottleneck module. We use the Bottleneck module to remove nighttime flare in the latent space, and get the nighttime deflared feature $F_{BD} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 16C}$,

$$F_{BD} = Bottleneck(F_{BF}) \tag{4}$$

where $Bottleneck(\cdot)$ represents Bottleneck module of FF-Former, which has same components as Encoder module in each level. Next, we use the Decoder to restore the deflared feature to the original scale and obtain the reconstructed feature $F_R \in \mathbb{R}^{H \times W \times C}$,

$$F_R = Decoder(F_{BD}) \tag{5}$$

$$F_{D_{i+1}} = Up^{\uparrow 2^i}(SFTB(F_{D_i})), i \in 1, 2, 3, 4 \tag{6}$$

where $Decoder(\cdot)$ represents Encoder module of FF-Former, which is also composed of 4-level SFTB and upsampling module $Up^{\downarrow 2^i}(\cdot)$. And we fuse the encoder features in each level to protect the details information of the reconstructed features. F_{D_i} represents the decoding features from each level of Decoder. Finally, we input the reconstructed feature into the output projection module to get the deflared image, and use the sigmoid function to ensure that the deflared value is between 0 and 1.

$$I_{DF} = sigmoid(\sigma(Conv(I_{F_R})) + I_{FC}) \tag{7}$$

where $\sigma(Conv(\cdot))$ represents the output projection, which is also composed of a simple 3×3 convolutional layer with $LeakyReLU()$. Summarily, I_{DF} can also be represented as follows,

$$I_{DF} = FF - Former(I_{FC}) \tag{8}$$

where $FF-Former(\cdot)$ denotes the function of FF-Former.

3.1. Swin Fourier Transformer Block (SFTB)

SFTB is our core module, which is composed of Hybrid Attention Block (HAB), Overlapping Cross-Attention Block (OCAB) and a Spatial Frequency Block (SFB). The purpose of SFTB is to solve the problem that the window-based Transformer cannot establish long-term dependence. Therefore, we introduce Fast Fourier Convolution (FFC) into SFTB to solve the problem of insufficient receptive field of Swin Transformer without increasing too much computation. And SFTB uses the powerful modeling ability of Swin Transformer to extract local features and the global perception ability of SFB to extract global features. It further improves the glare identification and image restoration ability by fusing local and global information. SFB consists of two branches: spatial module of left and frequency module on right. We use the frequency module to analyze flare from a global perspective, and use the spatial module to retain more details information when restoring the nighttime flare-corrupted image.

3.1.1 HAB and OCAB

HAT [6] demonstrates incorporating a channel attention block (CAB) into the Swin Transformer block can activate more pixels for restoring image and further enhance the representation ability of the network. Following HAT, we place the CAB into the Swin Transformer in parallel with the (shifted) window-based multi-head self-attention ((S)W-MSA) module, which improves the feature representation from channel and spatial dimensions. The HAB is formulated as,

$$X = LN(X) \tag{9}$$

$$X = (S)WMSA(LN(X)) + \alpha CAB(X) + X \tag{10}$$

$$X = MLP(LN(X)) + X \tag{11}$$

where $(S)WMSA(\cdot)$ is the standard Swin Transformer with the (shifted) window-based multi-head self-attention, $LN(\cdot)$ is the LayerNorm (LN) layer and $MLP(\cdot)$ is the multi-layer perceptron module. $CAB(\cdot)$ denotes the channel attention block, which consists of two 3×3 convolution layers with a $GELU$ activation function between them and a channel attention module. Specifically, $X \in \mathbb{R}^{W^2 \times C}$ is the local window feature. Then X is sent into the Overlapping Cross-Attention Block (OCAB) to further use more information.

3.1.2 Spatial Frequency Block (SFB)

Although HAB and OCAB increase the receptive field by increasing the window size of Swin Transformer and establishing overlapping cross attention, their receptive field is still limited to the window size, which means that they can only extract local features. To break this limitation, we use Fast Fourier Convolution (FFC) in Spatial Frequency Block (SFB) after OCAB to establish a long dependency to extract global features. SFB has the same global receptive field as Transformer with global self-attention, but only the convolution computation, which is very effective and efficient.

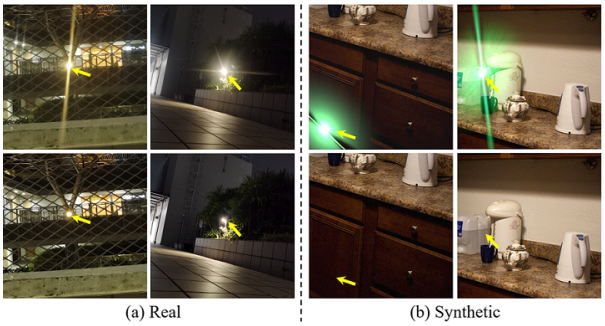
The architecture of SFB is shown in Figure 1(c) and is composed of two components: a spatial module for local information on the left and a frequency module for global information on the right. We concatenate the left and right outputs, and perform a convolution operation to obtain the final result. The formula of SFB is as follows,

$$X_{spatial} = Spatial(X) \tag{12}$$

$$X_{frequency} = Frequency(X) \tag{13}$$

$$X = Conv(([X_{spatial}||X_{frequency}])) \tag{14}$$

where $Spatial(\cdot)$ is the spatial convolution module and $Frequency(\cdot)$ represents the frequency FFC module. $X_{spatial}$ and $X_{frequency}$ denote the spatial (local) feature



(a) Real (b) Synthetic

Figure 2. Light source in real and synthetic dataset.

and frequency (global) feature. $Conv(\cdot)$ is a 3×3 convolution layer and $||$ denotes the concatenation operator.

In the frequency module, we transform the features from spatial into the frequency domain to extract the global information by using the 2-D Fast Fourier Transform (FFT). Then we perform a 1×1 convolution to extract the global feature. According to the convolution theorem [11], convolution in frequency domain equals point-wise multiplication in the spatial domain. This implies that convolution in frequency model of SFB incurs a global update. Finally, we perform inverse 2-D FFT operation to obtain spatial domain features. The $X_{frequency}$ is formulated as,

$$X = \sigma(Conv(X)) \tag{15}$$

$$X = IFFT(\sigma(Conv(FFT(X)))) + X \tag{16}$$

$$X_{frequency} = Conv(X) \tag{17}$$

where $FFT(\cdot)$ and $IFFT(\cdot)$ denote 2-D Fast Fourier Transform (FFT) and inverse 2-D FFT, separately. $Conv(\cdot)$ is 1×1 convolution layer and $\sigma = LeakyReLU()$. The spatial module is a simple residual block and is used for protecting the local information. The $X_{spatial}$ is formulated as,

$$X_{spatial} = Conv(\sigma(Conv(X))) + X \tag{18}$$

where $Conv(\cdot)$ is the 3×3 convolution layer.

3.2. Light Source Mask Loss Function

The nighttime flare is the diffraction phenomenon when shooting light source with defective or stained lens. Dai *et al*. [8] does not consider the light source in the background image when synthesizing the training data set, but the real image contains the light source, as shown in the Figure 2. So, the neural network is also removing light sources while removing nighttime flare, which will increase the burden of network and affect the deflared performance. In order to reduce the impact of light source on performance, we propose a light source mask loss function, as shown in the Figure 3. We generate the light source mask image based on the brightness of the flare, and incorporate it into the background image. To maintain the light source in the deflared

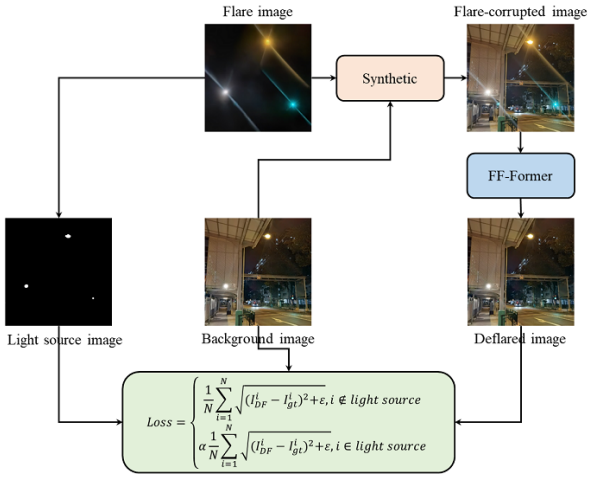


Figure 3. Light source mask loss function.

image, the loss within the light source is multiplied by a small factor. This simple approach effectively improves the performance of the model while preserving the light source in the deflared image to make it look more realistic and natural. The loss function is,

$$Loss = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\sqrt{(I_{DF}^{i}-I_{gt}^{i})^2+\varepsilon}, i \notin Light \\ \alpha\frac{1}{N}\sum_{i=1}^{N}\sqrt{(I_{DF}^{i}-I_{gt}^{i})^2+\varepsilon}, i \in Light \end{cases} \tag{19}$$

where $Light$ represents the pixels in the light source area.

4. Experiments

4.1. Datasets

Flare7K is a larger and more realistic nighttime flare removal dataset than previous works, which contains 5,000 scattering flares and 2,000 reflective flares. We add flare image to the background image to generate nighttime flare-corrupted image for training our FF-Former. For fair comparison, we use the same data augmentation strategy as Dai *et al*. [8]. We recover the linear luminance of flare image and flare-free image by using an inverse gamma correction strategy with $\gamma \sim U(1.8, 2.2)$. We also randomly multiply the RGB values with $U(0.5, 1.2)$ and add a Gaussian noise with variance sampled from a scaled chi-square distribution $\sigma^2 \sim 0.01\chi^2$ to improve the robustness of model. Then we carry out a series of affine transformations on flare images to enhance the diversity of flare. We also random blur the flare image with the kernel size in $U(0.1, 3)$ and add the offset in $U(-0.02, 0.02)$ to control the brightness of the entire image. Finally, we add the flare image to background image to generate the flare-corrupted image. We test FF-Former on real and synthetic nighttime flare images.

Data\Method		Input	Previous work			Network trained on Flare7k dataset				
			Zhang [27]	Sharma [20]	Wu [25]	U-Net [19]	HINet [4]	Restormer* [26]	Uformer [22]	FF-Former(Our)
Real-world	PSNR↑	22.56	21.02	20.49	24.61	26.11	26.74	26.28	26.98	**27.35**
	SSIM↑	0.857	0.784	0.826	0.871	0.879	0.882	0.883	0.890	**0.901**
	LPIPS↓	0.078	0.174	0.112	0.060	0.055	0.048	0.054	0.047	**0.044**
Synthetic	PSNR↑	22.77	21.04	20.01	27.88	29.07	29.97	29.45	30.47	**30.88**
	SSIM↑	0.921	0.841	0.865	0.952	0.958	0.959	0.950	0.965	**0.969**
	LPIPS↓	0.060	0.136	0.111	0.031	0.022	0.021	0.025	**0.017**	0.019

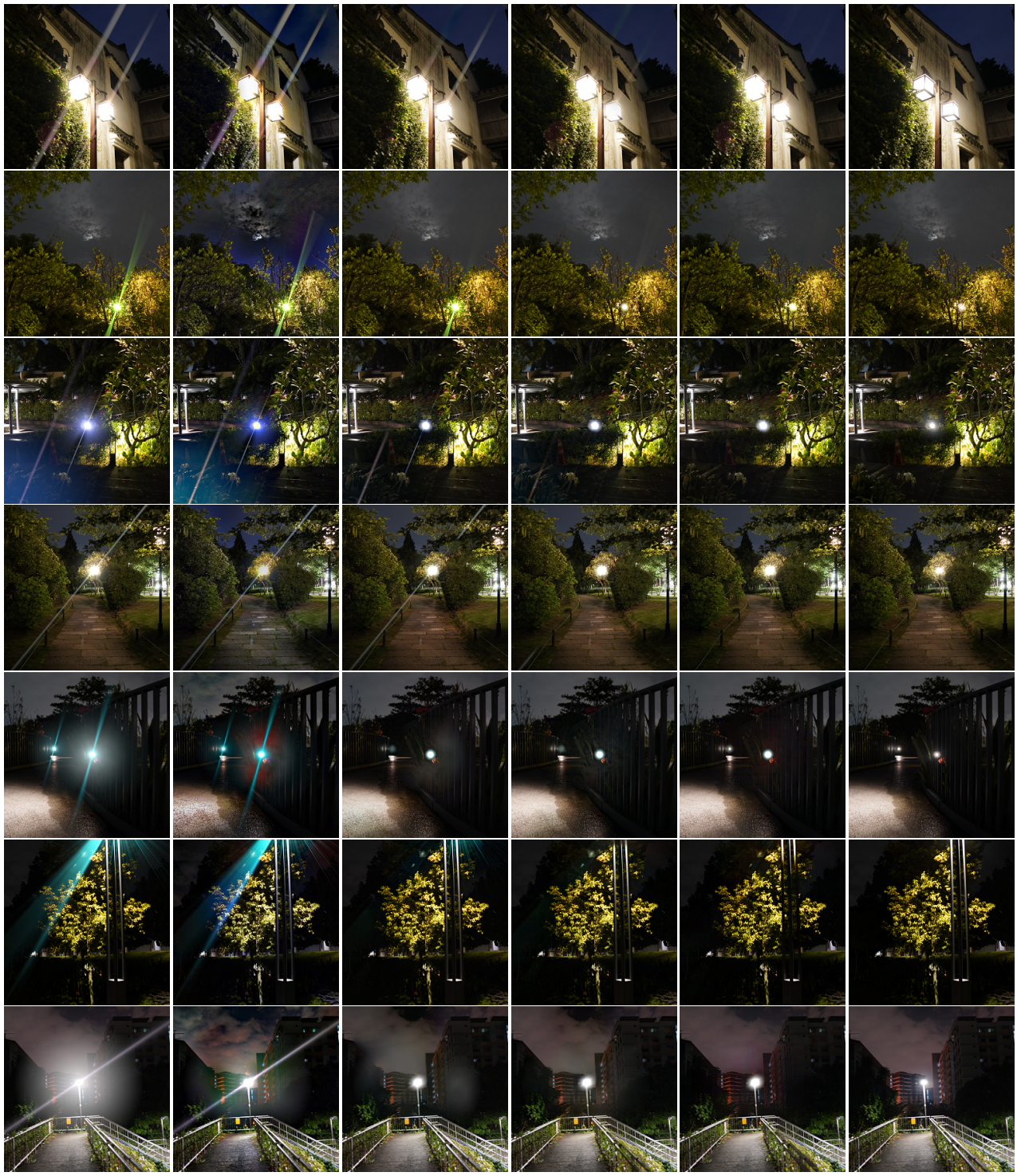
Table 1. Quantitative comparison of synthetic and real nighttime flare-corrupted data. The benchmark of the image restoration methods for nighttime flare removal is listed on the right part of the table. "*" denotes models with reduced parameters due to the limited GPU memory. The best results are in bold faces.

(a) Real input (b) Zhang [27] (c) Wu [25] (d) Dai [8] (e) FF-Former(Our) (f) GT

Figure 4. Visual comparison of flare removal on real-world nighttime flare images.

4.2. Implementation Details

SFTB is the basic module for our FF-Former, and the channel numbers C is 32 in the first SFTB of Encoder. We set 4-level in the Encoder and Decoder module for extracting multi-scale features. Following Dai *et al*. [8], we crop the input flare free and flare-corrupted images into 512×512 with batch size of 2 to train our FF-Former. We use the



(a) Input (b) Zhang [27] (c) Wu [25] (d) Dai [8] (e) FF-Former(Our) (f) GT

Figure 5. Visual comparison on synthetic nighttime flare images.

Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to optimize the Light Source Mask Loss Function. We only use Charbonnier L1 loss function [13] and set α to 0.05. The initial learning rate is 1e-4 and we use CosineAnnealingLR with 600,000 maximum iterations and 1e-7 minimum learning rate to adjust learning rate. we also use horizontal and vertical flip for

Spatial Frequency Block	Real-world	Synthetic
✗	27.23	30.75
✓	27.35	30.88

Table 2. Performance in the real-world and synthetic nighttime flare datasets with/without Spatial Frequency Block (SFB).

Light Source Mask Loss	Real-world	Synthetic
✗	27.17	30.62
✓	27.35	30.88

Table 3. Performance in the real-world and synthetic nighttime flare datasets with/without using Light Source Mask Loss Function.

data enhancement.

4.3. Comparison to state-of-the-arts methods

The quantitative results of our FF-Former on real-world and synthetic nighttime flare dataset achieves the best performance compared to other models, which is present in Table 1. HINet [4], Restormer [26] and Uformer [22] are the state-of-the-art methods for image restoration, and Dai *et al*. [8] trains them in the Flare7k dataset to remove the nighttime flare, in which Uformer performs best than others. However, Uformer can only extract local features by using window-based Transformer, and does not have the ability to identify nighttime flare images and restore deflared images from a global perspective. Our FF-Former solves the above problems. Especially, FF-Former improves the PSNR of Uformer from 26.98 dB and 30.47 dB to 27.35 dB and 30.88 dB in real-world and synthetic nighttime flare datasets respectively, 0.37 dB and 0.41 dB higher than its. It demonstrates the effectiveness of our proposed method and represents a major improvement over the nighttime flare removal task. The SSIM results of our FF-Former has the same conclusion as PSNR.

We also conduct a series qualitative comparison with Dai in the real-world and synthetic nighttime flare datasets, as shown on Figure 4 and Figure 5. From the visual results in Figure 4, our method can better restore the round flare with a larger radius, and has better results for the flare near the light source and retain more details. These all prove the effectiveness of our RFB and light source mask loss function. From the visual results in the first and second rows of Figure 5, our method can also remove longer streak flare, demonstrating our RFB's ability to perceive global flare. The fifth row of Figure 5 also shows that our method can better retain the light source information. Although the LPIPS of our FF-Former is smaller than Uformer, we can see from the visual results in the seventh row of Figure 5 that our results are more realistic and natural, and the flare removal is cleaner.

4.4. Ablation Study

4.4.1 Impact of Spatial Frequency Block (SFB)

The nighttime flare is continuous, and it usually passes through the whole image. The existing windows-based Transformer does not have the ability to model long term dependence, that is to say, it has only local receptive field and has no ability to perceive global flare information. Therefore, we introduce Fast Fourier Convolution (FFC) into Spatial Frequency Block (SFB) to solve the problem of insufficient receptive field of Swin Transformer, and use a simple residual block to protect details. The experimental results with/without SFB on real-world and synthetic nighttime flare dataset demonstrate that our method can improve the nighttime flare removal performance, as show in Table 2. The visual results in Figure 4 and Figure 5 demonstrate that our RFB have the ability to perceive global flare.

4.4.2 Impact of Light Source Mask Loss Function

The nighttime flare must be near the light source. The existing synthetic data does not add light sources to the background image, which will cause the light sources in the image to be removed after the training. In the previous works [8], the saturated regions of nighttime flare-corrupted image are extracted and pasted back to the deflared image to recover the light source. We need to keep the light source in the deflared image, but we also need to remove the light source in the training process. Such contradictory problems will increase the burden of network training and affect the flare removal performance. We propose a light source mask loss function to alleviate the above problems. In this simple way, our method effectively improves the performance of the model, and also preserves the light source in the deflared image to make it look more realistic and natural, as show in Table 3, Figure 4 and Figure 5.

5. Conclusion

In this paper, we introduce a novel solution for nighttime flare removal, called FF-Former, which is based on an U-shape network with Fast Fourier Convolution (FFC). To overcome the limitations of receptive field in current window-based Transformer methods, we propose the Swin Fourier Transformer Block (SFTB), which offers a comprehensive analysis of nighttime flares from a global perspective while retaining important image details during the restoration process. Additionally, we incorporate a Light Source Mask Loss Function to ensure the preservation of light sources in the output images, resulting in more realistic and natural-looking results. Our extensive experiments on both real-world and synthetic benchmarks demonstrate that our FF-Former outperforms existing models, making it a promising solution for nighttime flare removal.

References

[1] CS Asha, Sooraj Kumar Bhat, Deepa Nayak, and Chaithra Bhat. Auto removal of bright spot from images captured against flashing light source. In *2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, pages 1–6. IEEE, 2019. 2

[2] Paul A Boynton and Edward F Kelley. Liquid-filled camera for the measurement of high-contrast images. In *Cockpit Displays X*, volume 5080, pages 370–378. SPIE, 2003. 2

[3] Floris Chabert. Automated lens flare removal. In *Technical report*. Department of Electrical Engineering, Stanford University, 2015. 2

[4] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 6, 8

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 4

[7] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 2

[8] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flare7k: A phenomenological nighttime flare removal dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 2, 5, 6, 7, 8

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[11] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004. 5

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[13] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018. 7

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[15] H Angus Macleod and H Angus Macleod. *Thin-film optical filters*. CRC press, 2010. 2

[16] Andreas Nussberger, Helmut Grabner, and Luc Van Gool. Robust aerial object tracking from an airborne platform. *IEEE Aerospace and Electronic Systems Magazine*, 31(7):38–46, 2016. 2

[17] Xiaotian Qiao, Gerhard P Hancke, and Rynson WH Lau. Light source guided single-image flare removal from unpaired data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4177–4185, 2021. 2

[18] Ramesh Raskar, Amit K. Agrawal, Cyrus A. Wilson, and Ashok Veeraraghavan. Glare aware photography: 4d ray sampling for reducing glare effects of camera lenses. *ACM Trans. Graph.*, 27(3):56, 2008. 2

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[20] Aashish Sharma and Robby T Tan. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2021. 6

[21] Patricia Vitoria and Coloma Ballester. Automatic flare spot artifact detection and removal in photographs. *Journal of Mathematical Imaging and Vision*, 61(4):515–533, 2019. 2

[22] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 6, 8

[23] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *NIPS*, 2017. 1

[24] Yicheng Wu, Qiurui He, Tianfan Xue, Rahul Garg, Jiawen Chen, Ashok Veeraraghavan, and J Barron. Single-image lens flare removal. *arXiv preprint arXiv:2011.12485*, 2020. 2

[25] Yicheng Wu, Qiurui He, Tianfan Xue, Rahul Garg, Jiawen Chen, Ashok Veeraraghavan, and Jonathan T Barron. How to train neural networks for flare removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2239–2247, 2021. 2, 6, 7

[26] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 6, 8

[27] Jing Zhang, Yang Cao, Zheng-Jun Zha, and Dacheng Tao. Nighttime dehazing with a synthetic benchmark. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2355–2363, 2020. 6, 7