Progressive Frequency-Aware Network for Laparoscopic Image Desmoking

Jiale Zhang and Wenfeng Huang, Xiangyun Liao(✉), and Qiong Wang(✉)

Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{xy.liao,wangqiong}@siat.ac.cn

Abstract. Laparoscopic surgery offers minimally invasive procedures with better patient outcomes, but smoke presence challenges visibility and safety. Existing learning-based methods demand large datasets and high computational resources. We propose the Progressive Frequency-Aware Network (PFAN), a lightweight GAN framework for laparoscopic image desmoking, combining the strengths of CNN and Transformer for progressive information extraction in the frequency domain. PFAN features CNN-based Multi-scale Bottleneck-Inverting (MBI) Blocks for capturing local high-frequency information and Locally-Enhanced Axial Attention Transformers (LAT) for efficiently handling global low-frequency information. PFAN efficiently desmokes laparoscopic images even with limited training data. Our method outperforms state-of-the-art approaches in PSNR, SSIM, CIEDE2000, and visual quality on the Cholec80 dataset and retains only 629K parameters. Our code and models are made publicly available at: https://github.com/jlzcode/PFAN.

Keywords: Medical Image Analysis · Vision Transformer · CNN.

1 Introduction

Laparoscopic surgery provides benefits such as smaller incisions, reduced post-operative pain, and lower infection rates [14]. The laparoscope, equipped with a miniature camera and light source, allows visualization of surgical activities on a monitor. However, visibility can be hindered by smoke from laser ablation and cauterization. Reduced visibility negatively impacts diagnoses, decision-making, and patient health during intraoperative imaging and image-guided

Jiale Zhang and Wenfeng Huang contribute equally to this work.

This research was supported by multiple grants, including: The National Key Research and Development Program of China (2020YFB1313900), National Natural Science Foundation of China (62072452), Shenzhen Science and Technology Program (JCYJ20200109115627045, JCYJ20220818101408019, JCYJ20200109115201707) and Regional Joint Fund of Guangdong (2021B1515120011).

surgery, and hampers computer vision algorithms in laparoscopic tasks such as depth estimation, surgical reconstruction and lesion identification. Although smoke evacuation equipment is commonly used, its high cost and impracticality make image processing-based approaches a more attractive alternative [27]. However, traditional image processing algorithms have limitations in efficacy and can cause visual distortions. Approaches based on atmospheric scattering models inaccurately treat smoke as a homogeneous scattering medium, potentially leading to tissue misidentification and surgical accidents. End-to-end deep learning approaches show promise, but acquiring large training datasets is difficult and time-consuming, especially for medical applications. Moreover, most deep learning-based models have large parameter counts, making them unsuitable for resource-constrained medical devices. Laparoscopic models must be adaptable to various smoke concentrations and brightness levels, applicable across different surgical environments, lightweight, and effective with limited datasets.

In this study, we propose the Progressive Frequency-Aware Net (PFAN), an efficient, lightweight model built within the generative adversarial networks (GANs) framework for laparoscopic smoke removal. We address smoke removal by focusing on the image frequency domain, integrating high-frequency and low-frequency features to translate smoke-filled images into clear and no-artifacts smoke-free images. With only 629K parameters, PFAN demonstrates remarkable laparoscopic image desmoking results. In summary, the contributions of this work include:

(1) Our proposed PFAN model effectively combines CNN and ViT to take into account frequency domain features of laparoscopic images. PFAN employs the MBI (CNN-based) and LAT (ViT-based) components to sequentially extract high and low-frequency features from the images. This approach establishes a robust feature extraction framework by leveraging the CNNs' local high-frequency feature extraction capabilities and the Transformers' global low-frequency feature extraction strengths.

(2) Our work introduces two innovations to the PFAN model: the Multi-scale Bottleneck-Inverting (MBI) Block, which extracts local high-frequency features using a multi-scale inverted bottleneck structure, and the Locally-Enhanced Axial Attention Transformer (LAT), which efficiently processes global low-frequency information with Squeeze-Enhanced Axial Attention and Locally-Enhanced Feed Forward Layer.

(3) The lightweight PFAN model effectively removes smoke from laparoscopic images with a favorable performance-to-complexity balance. It is suitable for resource-constrained devices. Evaluation results indicate its superiority over state-of-the-art methods, highlighting its effectiveness in removing surgical smoke from laparoscopic images.

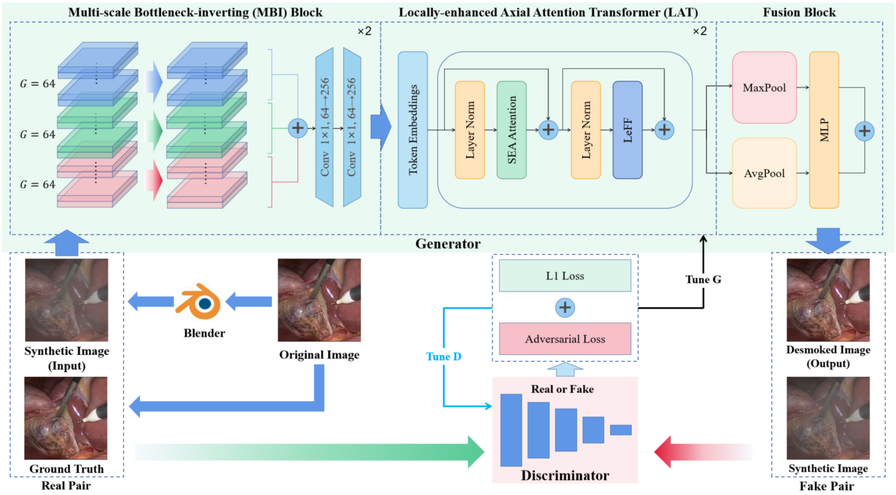

Fig. 1. The flowchart of PFAN illustrates a framework consisting of a generator network (G) and a discriminator network (D). Within this proposed approach, the generator G incorporates Multi-scale Bottleneck-Inverting (MBI) Blocks and Locally-Enhanced Axial Attention Transformer (LAT) Blocks.

2 Related Work

2.1 Traditional Theory-Based Desmoking Methods

Traditional desmoking techniques include image restoration and enhancement. Restoration methods, like the dark channel prior (DCP) by He et al. [7], use atmospheric degradation and depth information, but face limitations in laparoscopic imaging. Enhancement techniques, such as Retinex algorithm [22], wavelet-based algorithms [24], improving local contrast, increasing visibility and interpretability [18]. Wang et al. [29] created a variational desmoking approach, but it relies on assumptions regarding smoke's heterogeneous nature and varying depths.

2.2 Deep Learning-Based Desmoking Methods

Deep learning advances have fostered diverse frameworks for laparoscopic image smoke removal. Sabri et al. [2] employed synthetic surgical smoke images with different smoke densities and utilized CNNs to remove smoke in a supervised setting, while DehazeNet [23] , AOD-Net [16] and DM^{2}F-Net [4] relied on atmospheric scattering models, inappropriate for surgical environments. GANs [6], using game-theoretic approaches, generate realistic images. Techniques like Pix2Pix [13] employ conditional GANs for domain mapping. In medical imaging, GANs have been effective in PET-CT translation and PET image denoising [1]. However, methods based on convolutional neural networks struggle with low-frequency information extraction, such as contour and structure. Vision Transformers (ViT) [5] excel in low-frequency extraction, but their complexity restricts use in resource-limited medical devices.

3 Methodology

Fig. 1 depicts our proposed PFAN, a lightweight CNN-ViT-based approach within a GAN architecture for desmoking in laparoscopic images, it extracts information progressively in the frequency domain by leveraging the strengths of CNNs and ViTs. In order to obtain the necessary corresponding smoky and non-smoky images, we integrate a graphics rendering engine into our learning framework to generate paired training data without manual labeling.

3.1 Synthetic Smoke Generation

We employ the Blender engine to generate smoke image pairs for model training, offering two advantages over physically-based haze formation models [23] and Perlin noise functions [3]. First, laparoscopic surgical smoke is localized and depth-independent, making traditional haze models unsuitable. Second, modern rendering engines provide realistic and diverse smoke shapes and densities using well-established, physically-based built-in models. With Blender's render engine, denoted by ϕ, we generate the smoke evolution image sequence, S_{Smoke}, by adjusting parameters such as smoke source density, intensity, temperature, location (S_d, S_i, S_t, S_l), and light location and intensity (L_l, L_i):

$$S_{Smoke} = \phi(S_d, S_i, S_t, S_l, L_l, L_i) \tag{1}$$

Let I_{Smoke} represent one frame of the smoke image sequence. To create a synthetic smoke evolution image sequence (I_{Syn}) within the surgical scene, we overlay a randomly generated frame of smoke evolution image sequence (I_{Smoke}) onto each smoke-free laparoscopic image ($I_{Smoke-free}$). The following formula represents this process:

$$I_{Syn} = I_{Smoke-free} + I_{Smoke} \tag{2}$$

The synthesized laparoscopic image sequence shows the evolution process of smoke. In the first frame of the synthesized image sequence, smoke is only present at a specific location within the image, simulating the situation of burning lesion areas in laparoscopic surgery. As time progresses, it disperses from the burning point outwards according to random density, temperature, and intensity parameters. The synthesis of an extensive range of realistic images depicting simulated surgical smoke is made possible through the utilization of a robust rendering engine. By incorporating variations in smoke, such as location, intensity, density, and luminosity, over-fitting is prevented in the network's training.

3.2 Multi-scale Bottleneck-Inverting (MBI) Block

The MBI Block is designed to efficiently extract high-frequency features, drawing inspiration from various well-established neural networks [26, 20, 12, 11]. Here, we denote input smoke images as $\mathcal{X}_{Smoke} \in \mathbb{R}^{H\times W\times 3}$, and the set of high-frequency information extracted by each MBI Block can be defined as $\{\mathcal{X}_{HF} = \mathcal{X}_{HF_1}, ..., \mathcal{X}_{HF_k}\}$. Within each MBI Block, group convolution is represented as GConv, and the multi-scale feature can be obtained as:

$$\mathcal{X}_{MS} = GConv_{i,g}(\mathcal{X}_{Smoke}) + GConv_{j,g}(\mathcal{X}_{Smoke}) + GConv_{k,g}(\mathcal{X}_{Smoke}) \quad (3)$$

Here, i, j, and k represent the size of the receptive field, which were set to 3, 7, and 11, respectively. We choose GELU [9] instead of RELU as the activation function following each convolution layer, given its smoother properties and proven higher performance. The parameter g indicates that, during group convolution, input features will be divided into g groups. In this paper, this value is set to 64, which matches the feature channels, resulting in a significant reduction of parameters by 1/64 in comparison to standard convolution. Next, we merge the multi-scale feature $\mathcal{X}_{MS}$ and expand it to a high-dimensional representation using point-wise convolution. Following this, features are projected back to a low-dimensional representation through point-wise convolution, represented as

$$\mathcal{X}_{HF} = PwConv_{high \rightarrow low}(PwConv_{low \rightarrow high}(\mathcal{X}_{MS})) \quad (4)$$

3.3 Locally-Enhanced Axial Attention Transformer (LAT) Block

Applying ViT models to desmoke laparoscopic images faces challenges. ViT's multi-head self-attention layer applies global attention, neglecting differing frequencies and local high-frequency information. Additionally, ViT's computational cost increases quadratically with token count, limiting its use with high-resolution feature maps. To overcome these issues, we introduce the Locally-Enhanced Axial Attention Transformer (LAT) Block. It combines streamlined squeeze Axial attention for global low-frequency semantics and a convolution-based enhancement branch for local high-frequency information. The LAT Block captures long-range dependencies and global low-frequency information with low parameter counts.

Given the features at MBI Block outputs, $\mathcal{X}_{MBI}$, LAT first reshapes the input into patch sequences using $H \times H$ non-overlapping windows. And then Squeeze-Enhanced Axial Attention computes attention maps ($\mathcal{X}_{Sea}$) for each local window. To further process the information, LAT replaces the multi-layer perceptron (MLP) layers in a typical ViT with a Locally-Enhanced Feed Forward Layer. Additional skip connections enable residual learning and produce $\mathcal{X}_{LAT}$.

$$\mathcal{X}_{Sea} = SEA(LN(\mathcal{X}_{MBI})) + \mathcal{X}_{MBI}, \quad \mathcal{X}_{LAT} = LEFF(LN(\mathcal{X}_{Sea})) + \mathcal{X}_{Sea} \quad (5)$$

Here, $\mathcal{X}_{Sea}$ and $\mathcal{X}_{LAT}$ correspond to the outputs of the Squeeze-Enhanced Axial Attention and LEFF modules, respectively. LN denotes layer normalization [15]. We discuss Squeeze-Enhanced Axial Attention and LEFF in detail in subsequent sections.

Squeeze-Enhanced Axial Attention (SEA) The Squeeze-Enhanced Axial Attention utilized in the Locally-Enhanced Axial Attention Transformer (LAT)

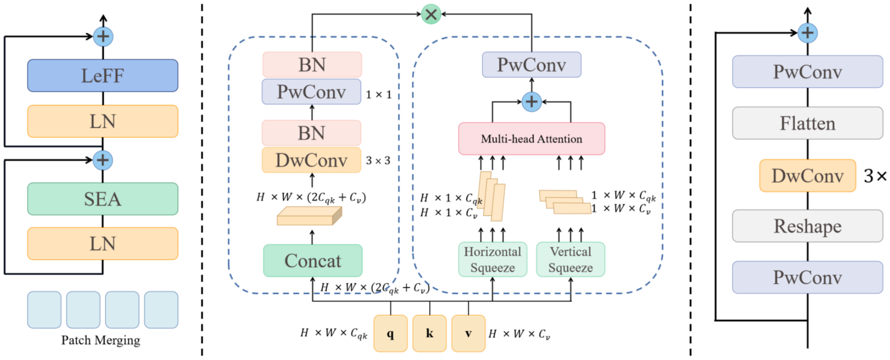

Fig. 2. Left: the schematic illustration of the proposed Locally-Enhanced Axial Attention Transformer Block. Middle: Squeeze-Enhanced Axial Attention Layer. Right: Locally-Enhanced feed-forward network.

is designed to extract global information in a succinct way. Initially, we compute $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$ by $\mathbf{q} = W_q * \mathcal{X}, \mathbf{k} = W_k * \mathcal{X}, \mathbf{v} = W_v * \mathcal{X}$, where $\mathcal{X} \in \mathbb{R}^{H \times W \times C}$. $W_q, W_k \in \mathbb{R}^{C_{qk} \times C}$ and $W_v \in \mathbb{R}^{C_v \times C}$ are learnable weights. Then, a horizontal squeeze $\mathbf{q}_h$ is executed by averaging the query feature map along the horizontal direction and a vertical squeeze $\mathbf{q}_v$ is applied in the vertical direction.

$$\mathbf{q}_h = \frac{1}{W}\left(\mathbf{q}^{(C_{qk},H,W)} \mathbb{1}_W\right)^{\rightarrow(H,C_{qk})}, \quad \mathbf{q}_h = \frac{1}{H}\left(\mathbf{q}^{(C_{qk},W,H)} \mathbb{1}_H\right)^{\rightarrow(W,C_{qk})} \tag{6}$$

The notation $\mathbf{z}^{\rightarrow(\cdot)}$ represents the permutation of tensor $\mathbf{z}$'s dimensions, and $\mathbb{1}_m \in \mathbb{R}_m$ is a vector with all elements equal to 1. The squeeze operation on $\mathbf{q}$ is also applied to $\mathbf{k}$ and $\mathbf{v}$, resulting in $\mathbf{q}_h, \mathbf{k}_h, \mathbf{v}_h \in \mathbb{R}^{H \times C_{qk}}, \mathbf{q}_v, \mathbf{k}_v, \mathbf{v}_v \in \mathbb{R}^{W \times C_{qk}}$. The squeeze operation consolidates global information along a single axis, thereby significantly improving the subsequent global semantic extraction process, as demonstrated by the following equation.

$$\mathbf{y}_{(i,j)} = \sum_{p=1}^{H} softmax_p \left({\mathbf{q}_h^i}^{\top} \mathbf{k}_h^p\right) \mathbf{v}_h^p + \sum_{p=1}^{W} softmax_p \left({\mathbf{q}_v^j}^{\top} \mathbf{k}_v^p\right) \mathbf{v}_v^p \tag{7}$$

As can be seen, in Squeeze-Enhanced Axial Attention, each position of the feature map only propagates information in two squeezed axial features, while in traditional self-attention (as in the following equation), each position of the feature map calculates self-attention with all positions.

$$\mathbf{y}_{(i,j)} = \sum_{p \in \mathcal{G}_{(i,j)}} softmax_p \left(\mathbf{q}_{(i,j)}^{\top} \mathbf{k}_p\right) \mathbf{v}_p \tag{8}$$

The traditional global self-attention is as above, where $\mathcal{G}(i,j)$ means all positions on the feature map at location (i,j). When a conventional attention module is applied to a feature map with dimensions $H \times W \times C$ the time complexity becomes $O(H^2W^2(C_{qk}+C_v))$, resulting in low efficiency. However, with SEA, the time complexity for squeezing q, k, v is $O((H+W)(2C_{qk}+C_v))$ and the attention operation takes $O((H^2+W^2)(C_{qk}+C_v))$ time. Consequently, our squeeze Axial attention successfully lowers the time complexity to $O(HW)$, ensuring a more efficient and faster process.

Locally-Enhanced Feed-Forward Network (LEFF) Adjacent pixels play a crucial role in image desmoking, as demonstrated in [29], which highlights their essential contribution to image dehazing and denoising. However, previous research [27] has highlighted the limited ability of the Feed-Forward Network (FFN) within the standard Transformer to effectively utilize local context. To address this limitation, we introduce a depth-wise convolutional block to LAT, inspired by recent studies [17]. As depicted in Fig. 2 (Right), we begin by applying a linear projection layer to each token to augment its feature dimension. Subsequently, we reshape the tokens into 2D feature maps and implement a 3×3 depth-wise convolution to capture local information. Afterward, we flatten the features back into tokens and reduce the channels using another linear layer to align with the input channel dimensions. *LeakyReLU* serves as the activation function following each linear or convolution layer.

Fusion Block We employ Channel Attention [31] as the Fusion Block in our approach to enhance the cross-channel feature fusion capabilities. The Channel Attention mechanism models inter-dependencies between channels of features, enabling adaptive adjustment of feature responses across different channels, and assigning corresponding weights. Embedding channel attention can facilitate adaptive enhancement and fusion of convolution and corresponding Transformer features in the LAT module. The attention map, $\mathcal{X}_{CA}$, can be calculated using the function, where σ represents the *Sigmoid* function.

$$\mathcal{X}_{CA} = \sigma \left(LEFF(AvgPool(\mathcal{X}_{LAT})) + LEFF(MaxPool(\mathcal{X}_{LAT})) \right) \tag{9}$$

Afterward, the low-frequency information $\mathcal{X}_{LF}$ is acquired as described in (10).

$$\mathcal{X}_{LF} = \mathcal{X}_{LAT} \cdot \mathcal{X}_{CA} \tag{10}$$

To achieve the smoke-free result, $\mathcal{X}_{Smoke-free}$, the low-frequency information of the original input smoke image $\mathcal{X}_{LF}$ is combined with the high-frequency information $\mathcal{X}_{HF}$, which is the output of MBI blocks.

$$\mathcal{X}_{Smoke-free} = \mathcal{X}_{HF} + \mathcal{X}_{LF} \tag{11}$$

4 Experiment

4.1 Data Collections

We used images from the Cholec80 dataset [28], consisting of 80 cholecystectomy surgery videos by 13 surgeons. We sampled 1,500 images at 20-second intervals from these videos, selecting 660 representative smoke-free images. As detailed in Section 3.1, we added synthetic random smoke, yielding 660 image pairs, divided in an 8:1:2 ratio for training, validation, and testing. Synthetic smoky images were generated according to Section 3.1. Importantly, each dataset contained distinct videos, ensuring no overlap.

4.2 Implementation Details

Our experiments utilized six NVIDIA RTX 2080Ti GPUs. Initially, we trained the Discriminator (PatchGAN) for one epoch to provide a rough smoke mask, followed by iterative training of the Discriminator and Generator while freezing the PatchGAN's parameters during the Generator's training. We employed an Adam solver with a learning rate of 0.0002, momentum parameters $\beta 1 = 0.5$ and $\beta 2 = 0.999$, and a batch size of 6. Consistent with prior research, random cropping was used for generating training and validation patches.

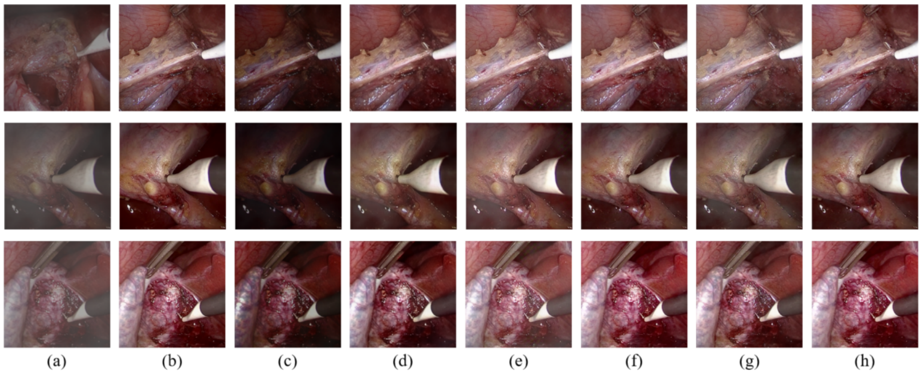

Fig. 3. Comparison experiments between SOTAs. (a) Input (b) Ground Truth, (c) Dark Channel Prior(DCP) [7] (d) CycleGAN + ResNet, (e) CycleGAN + U-Net, (f) Pix2Pix + ResNet, (g) Pix2Pix + U-Net, and (h) Ours.

5 Result

Table 1. Quantitative results. The best and second-best results are highlighted and underlined, respective

Model		Parameters↓	PSNR↑	SSIM↑	CIEDE2000↓
DCP		/	27.6250	0.5528	35.9952
CycleGAN	U-Net	54414K	28.7449	0.7621	10.3298
CycleGAN	ResNet6	7841K	29.0250	0.7826	9.5821
CycleGAN	ResNet9	11383K	29.0926	0.7802	9.2868
Pix2Pix	U-Net	54414K	29.2967	0.7073	8.8060
Pix2Pix	ResNet6	7841K	29.8249	0.8358	6.9364
Pix2Pix	ResNet9	11383K	29.8721	0.8417	6.7046
Pix2Pix	Uformer	85605K	29.7030	0.8026	8.0602
Ablation Models					
w/o Multi-scale		613K	29.9970	0.8692	6.9362
w/o Fusion Block		629K	29.4425	0.7814	8.1200
w/o MBI		540K	29.7599	0.9029	6.9149
w/o LAT		90K	28.8936	0.7857	10.1284
Ours		629K	**30.4873**	**0.9061**	**5.4988**

In our quantitative evaluations, we assess desmoking performance by comparing smoke-free images to their desmoked counterparts using the following metrics: number of Parameters, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [10], and CIEDE2000 [21] (which represents color reconstruction accuracy for the human visual system). We compare the proposed method with eight state-of-the-art desmoking and dehazing methods including both a traditional image processing approach (DCP [7]) and the most recent deep learning-based methods (original CycleGAN [19] (with U-Net [25]), CycleGAN with ResNet [8] (6Blocks), CycleGAN with ResNet (9Blocks), original Pix2Pix [13] (with U-Net), Pix2Pix with ResNet (6Blocks), Pix2Pix with ResNet (9Blocks), Pix2Pix with Uformer [30]).

Table 1 demonstrates our model's superior performance compared to alternative methods based on synthetic datasets. The highest PSNR and SSIM values, and the lowest CIEDE2000 value, emphasize our approach's effectiveness in smoke removal tasks. Fig. 3 presents a subjective evaluation of desmoking results, emphasizing previous approaches' limitations in adequately removing smoke. Non-deep learning methods often produce low-brightness, color-shifted images due to DCP's unsuitability for surgical applications with complex lighting and varied smoke. Although deep learning techniques better restore brightness, CycleGAN and Pix2Pix cannot fully eliminate smoke, as evidenced by residual smoke in some image portions (Fig. 3). These methods also result in unclear tissue contours due to CNN-based models' restricted global low-frequency feature extraction. In contrast, our methodology yields cleaner images with enhanced brightness, sharp details, and distinct edges.

5.1 Evaluation under Different Smoke Densities

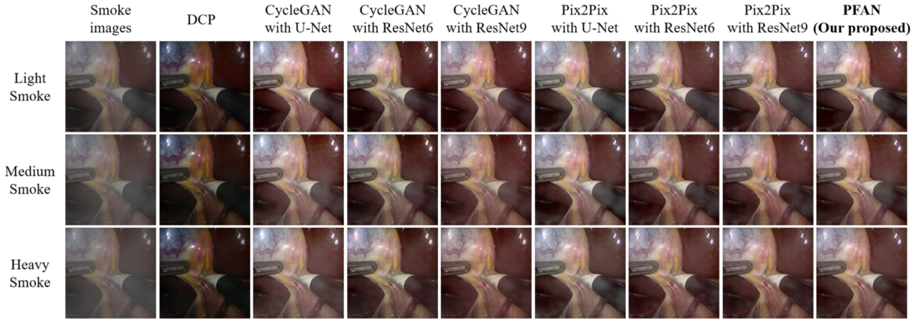

Fig. 4. Qualitative comparison between SOTAs under different smoke densities

Smoke impairs image information, often irreversibly, depending on thickness. To evaluate networks' desmoking performance at varying densities, we analyzed light, medium, and heavy smoke levels. We generated test sets for each density level with fixed starting positions and temperatures. Fig.4 displays rendered smoke images (I_{Syn}) and desmoked results from five methods and our approach. DCP struggles to restore dark-red tissue colors, whereas deep learning-based techniques perform better using context. Pix2Pix produces similar results but falters for some images, introducing artificial reflections. Our method achieves clean results with minor saturation deviations, even under dense smoke conditions. Table 2 compares our approach to five alternatives, consistently yielding the highest SSIM and PSNR while reducing CIEDE2000, outperforming other established methods.

Table 2. Quantitative comparison between SOTAs under different smoke densities

Smoke Density		Light Smoke			Medium Smoke			Heavy Smoke		
Model		PSNR↑	SSIM↑	CIEDE2000↓	PSNR↑	SSIM↑	CIEDE2000↓	PSNR↑	SSIM↑	CIEDE2000↓
DCP		27.6611	0.6215	30.1270	27.6811	0.5887	32.9143	27.6944	0.5807	33.8072
CycleGAN	U-Net	29.0426	0.7778	8.5370	28.9490	0.7607	10.7167	28.8837	0.7639	10.7521
CycleGAN	ResNet6	29.0713	0.7958	8.2868	28.7621	0.7741	11.7635	28.7647	0.7755	11.6661
CycleGAN	ResNet9	29.3232	0.8002	7.8017	28.7466	0.7650	11.9671	28.9379	0.7711	10.8202
Pix2Pix	U-Net	29.2652	0.7270	8.9004	29.4071	0.7119	9.1812	29.4474	0.7199	8.9037
Pix2Pix	ResNet6	29.9776	0.8404	6.6498	30.1833	0.8288	6.8033	30.2138	0.8344	6.2970
Pix2Pix	ResNet9	29.9492	0.8484	6.6610	30.1498	0.8372	**6.7079**	30.3287	0.8434	6.7079
Ours		**30.1209**	**0.8856**	**6.5182**	**30.2740**	**0.8704**	6.8001	**30.5223**	**0.8762**	**6.1147**

5.2 Ablation Studies

We design a series of ablation experiments to analyze the effectiveness of each of the modules we propose. The ablation results are reported in Table 1.

Effectiveness of the MBI Block: The goal of the MBI Block is to effectively capture multi-scale, high-frequency details. Fig. 5 demonstrates that removing the MBI Block results in remaining smoke and blurry edges and textures in some image portions. This limitation in high-frequency detail extraction makes it challenging to obtain satisfactory desmoking outcomes. In Table. 1, our PFAN outperforms the model without the MBI Block in terms of PSNR, SSIM, and CIEDE2000 metrics. This comparison highlights the critical role of MBI Blocks in achieving superior results.

Effectiveness of the LAT Block: The ViT-based LAT Blocks aim to extract global low-frequency information. Fig. 5 shows that the model without LAT Blocks achieves a visually similar desmoking effect to the Ground Truth (GT); however, the color appears dull and exhibits noticeable distortion compared to the original smoke-free image. The higher CIEDE2000 value indicates insufficient low-frequency feature extraction. Furthermore, the lower PSNR and SSIM values demonstrate the effectiveness of the LAT module.

Effectiveness of the Multi-scale MBI: Our approach employs group convolution with varying receptive fields in the MBI Block, facilitating multi-scale

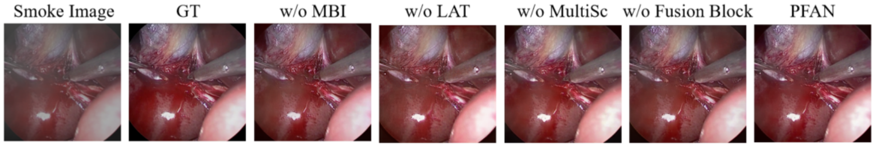

Fig. 5. Qualitative results of ablation experiments.

high-frequency information extraction. We conducted an ablation study, replacing multi-scale convolutions in the MBI Block with only 3×3 group convolutions. Fig. 5 reveals substantial improvements in smoke removal, but the tissue in the central scalpel area appears blurred. Table. 1 demonstrates the "w/o Multi-scale" model achieves comparable performance to PFAN in terms of CIEDE2000 and PSNR; however, the SSIM value is significantly inferior, highlighting the importance of Multi-scale group convolutions in the MBI Block.

Effectiveness of Fusion Block: The Fusion Block in our proposed method leverages channel attention for adaptive discriminative fusion between image Transformer features and convolutional features, enhancing the network's learning capability. Importantly, omitting channel attention leads to the most significant decline in SSIM value among the four ablation experiments. Additionally, noticeable differences in both PSNR and CIEDE2000 emerge compared to the PFAN results, underscoring channel attention's crucial role in PFAN.

6 Limitations

Our method has a few limitations. It overlooks external factors such as water vapor and pure white gauze, which can degrade image quality and then impede desmoking performance. Future iterations should incorporate these elements into training and testing to ensure clinical applicability. Moreover, our proposed single-frame desmoking method may introduce temporal discontinuity in video desmoking tasks due to smoke density fluctuations. Thus, based on our current method, further investigation into spatial-temporal convolution techniques is necessary for enhancing laparoscopic video desmoking.

7 Conclusion

In conclusion, we proposed a groundbreaking deep learning method PFAN for laparoscopic image desmoking. By incorporating the lightweight and efficient CNN-ViT-based approach with the innovative CNN-based Multi-scale Bottleneck-Inverting (MBI) Blocks and Locally-Enhanced Axial Attention Transformers (LAT), PFAN effectively captures both low and high-frequency information for desmoking analysis, even with a limited dataset. The evaluation on the synthetic Cholec80 dataset, with various smoke-dense images, showcases the superiority of PFAN compared to existing SOTAs in performance and visual effects. Additionally, PFAN maintains a lightweight design, making it a feasible and desirable choice for implementation in medical equipment. Our desmoking method enables advanced applications. It enhances surgical safety by providing real-time desmoked images, serving as a valuable reference during ablation procedures. Furthermore, beyond aiding surgeons directly, the technology can also improve the robustness of various vision-based surgical assistance systems when used as a preprocessing step.

References

1. Armanious, K., Jiang, C., Fischer, M., Thomas, K., Nikolaou, K.: MedGAN : Medical Image Translation using GANs pp. 1–16 (2016)
2. Bolkar, S., Wang, C., Cheikh, F.A., Yildirim, S.: Deep smoke removal from minimally invasive surgery videos. Proceedings - International Conference on Image Processing, ICIP pp. 3403–3407 (2018). https://doi.org/10.1109/ICIP.2018.8451815
3. Bolkar, S., Wang, C., Cheikh, F.A., Yildirim, S.: Sabri Bolkar , Congcong Wang † , Faouzi Alaya Cheikh , Sule Yildirim. 2018 25th IEEE International Conference on Image Processing (ICIP) pp. 3403–3407 (2018)
4. Deng, Z., Zhu, L., Hu, X., Fu, C.W., Xu, X., Zhang, Q., Qin, J., Heng, P.A.: Deep multi-model fusion for single-image dehazing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2453–2462 (2019)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2020), http://arxiv.org/abs/2010.11929
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020). https://doi.org/10.1145/3422622
7. He, K., Sun, J., TTang, X.: Single image haze removal using dark channel prior. 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (January 2011), 1956–1963 (2009). https://doi.org/10.1109/CVPRW.2009.5206515
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hendrycks, D., Gimpel, K.: Gaussian Error Linear Units (GELUs) pp. 1–10 (2016), http://arxiv.org/abs/1606.08415
10. Horé, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. Proceedings - International Conference on Pattern Recognition pp. 2366–2369 (2010). https://doi.org/10.1109/ICPR.2010.579
11. Huang, W., Liao, X., Qian, Y., Jia, W.: Learning hierarchical semantic information for efficient low-light image enhancement. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2023). https://doi.org/10.1109/IJCNN54540.2023.10190996
12. Huang, W., Liao, X., Zhu, L., Wei, M., Wang, Q.: Single-Image Super-Resolution Neural Network via Hybrid Multi-Scale Features. Mathematics **10**(4), 1–26 (2022). https://doi.org/10.3390/math10040653
13. Isola, P., Efros, A.A., Ai, B., Berkeley, U.C.: Image-to-Image Translation with Conditional Adversarial Networks
14. Jaschinski, T., Mosch, C.G., Eikermann, M., Neugebauer, E.A., Sauerland, S.: Laparoscopic versus open surgery for suspected appendicitis. Cochrane Database of Systematic Reviews **2018**(11) (2018). https://doi.org/10.1002/14651858.CD001546.pub4
15. Kotwal, A., Bhalodia, R., Awate, S.P.: Joint desmoking and denoising of laparoscopy images. Proceedings - International Symposium on Biomedical Imaging **2016-June**, 1050–1054 (2016). https://doi.org/10.1109/ISBI.2016.7493446
16. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: AOD-Net : All-in-One Dehazing Network pp. 4780–4788 (2017). https://doi.org/10.1109/ICCV.2017.511

17. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: LocalViT: Bringing Locality to Vision Transformers (2021), http://arxiv.org/abs/2104.05707
18. Li, Y., Miao, Q., Liu, R., Song, J., Quan, Y., Huang, Y.: Neurocomputing A multiscale fusion scheme based on haze-relevant features for single image dehazing. Neurocomputing **283**, 73–86 (2018). https://doi.org/10.1016/j.neucom.2017.12.046, https://doi.org/10.1016/j.neucom.2017.12.046
19. Liu, W., Hou, X., Duan, J., Qiu, G.: End-to-End Single Image Fog Removal Using **29**(1), 7819–7833 (2020). https://doi.org/10.1109/TIP.2020.3007844
20. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: [ConvNeXt_CVPR22] A ConvNet for the 2020s. Cvpr pp. 11976–11986 (2022), http://arxiv.org/abs/2201.03545
21. Luo, M.R., Cui, G., Rigg, B.: The development of the cie 2000 colour-difference formula: Ciede2000. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur **26**(5), 340–350 (2001)
22. Rahman, Z.u., Jobson, D.J., Woodell, G.A., Science, C.: Retinex Processing for Automatic Image Enhancement
23. Removal, I.H., Cai, B., Xu, X., Jia, K., Qing, C.: DehazeNet : An End-to-End System for Single. IEEE Transactions on Image Processing **25**(11), 5187–5198 (2016). https://doi.org/10.1109/TIP.2016.2598681
24. Rong, Z., Jun, W.L.: Improved wavelet transform algorithm for single image dehazing. Optik **125**(13), 3064–3066 (2014). https://doi.org/10.1016/j.ijleo.2013.12.077, http://dx.doi.org/10.1016/j.ijleo.2013.12.077
25. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
26. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
27. Tchaka, K., Pawar, V.M., Stoyanov, D.: Chromaticity based smoke removal in endoscopic images. Medical Imaging 2017: Image Processing **10133**(February 2017), 101331M (2017). https://doi.org/10.1117/12.2254622
28. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE Transactions on Medical Imaging **36**(1), 86–97 (2017). https://doi.org/10.1109/TMI.2016.2593957
29. Wang, C., Alaya Cheikh, F., Kaaniche, M., Beghdadi, A., Elle, O.J.: Variational based smoke removal in laparoscopic images. BioMedical Engineering Online **17**(1), 1–18 (2018). https://doi.org/10.1186/s12938-018-0590-5, https://doi.org/10.1186/s12938-018-0590-5
30. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A General U-Shaped Transformer for Image Restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
31. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)