International Journal of Computer Assisted Radiology and Surgery (2023) 18:1417–1427
https://doi.org/10.1007/s11548-023-02835-z

ORIGINAL ARTICLE

Surgical smoke removal via residual Swin transformer network

Feng Wang[1,2] · Xinan Sun[1,2] · Jinhua Li[1,2]

Received: 25 August 2022 / Accepted: 9 January 2023 / Published online: 23 January 2023

Abstract

Purpose In robot-assisted minimally invasive surgery (RMIS), smoke produced by laser ablation and cauterization causes degradation in the visual quality of the operating field, increasing the difficulty and risk of surgery. Therefore, it is important and meaningful to remove fog or smoke from the endoscopic video to maintain a clear visual field.

Methods In this paper, we propose a novel method for surgical smoke removal based on the Swin transformer. Our method firstly uses convolutional neural network to extract shallow features, then uses the Swin transformer block to further extract deep features and finally generates smoke-free images.

Results We conduct quantitative and qualitative experiments on the proposed method, and we also validate the desmoking results in the surgical instrument segmentation task. Extensive experiments on synthetic and real dataset show that the proposed approach has good performance and outperforms the state-of-the-art surgical smoke removal methods.

Conclusion Our method effectively removes surgical smoke, improves image quality and reduces the risk of RMIS. It provides a clearer visual field for the surgeon, as well as for subsequent visual tasks, such as instrument segmentation, 3D scene reconstruction and surgery automation.

Keywords Surgical image · Endoscope · Surgical smoke · Desmoking · Transformer

Introduction

In robot-assisted minimally invasive surgery (RMIS), surgeons perform surgical operations by endoscope to observe the tissues and organs. As a result, the clarity of the surgical image plays an important role in the surgery. However, the smoke and fog produced by surgical temperature difference or laser ablation during the surgery seriously obstructs the field and increases the surgical operating risk as shown in Fig. 1a. Meanwhile, smoke also affects the processing of subsequent visual tasks, such as instrument segmentation, 3D scene reconstruction and surgery automation. Therefore, surgical smoke removal is essential to maintain a clear surgical field for surgeons to perform endoscopic surgery safely, accurately and efficiently.

✉ Jinhua Li
lijinhua@tju.edu.cn

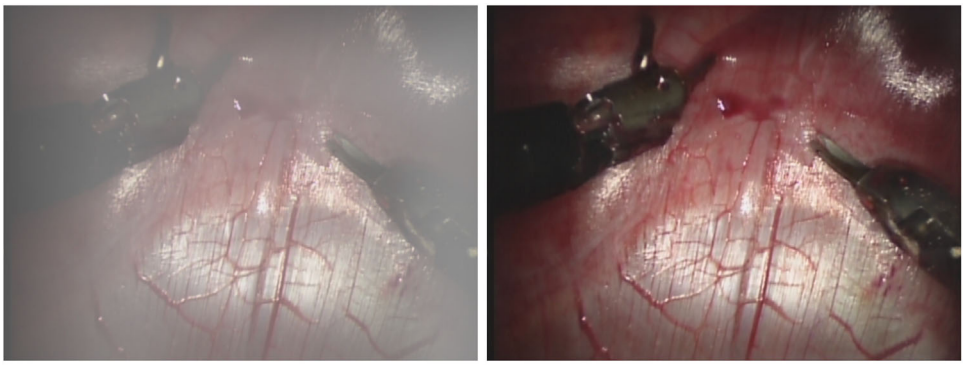
1 School of Mechanical Engineering, Tianjin University, 135 Yaguan Road, Jinnan District, Tianjin 300350, China

2 Institute of Medical Robotics and Intelligent Systems, Tianjin University, No. 2 Huake Fifth Road, Binhai Hi-Tech Industrial Development Area, Tianjin 300392, China

At present, there are two main types of surgical image desmoking or dehazing methods. One is based on mechanical methods, including endoscope lens warming strategies, anti-fogging materials and equipment modifications [1]. These methods are costly and time-consuming with repeated disruptions, and unsuitable for image-guided surgery. The other is vision-based methods, including traditional algorithms and deep learning-based algorithms.

Traditional methods are mostly based on the atmospheric scattering model (ASM) [2] which indicates that regions farther from the camera in a single hazy image have larger haze concentrations. These methods attempt to estimate transmission map by physical priors, such as dark channel prior [3] and color line prior [4], and then restore image by ASM. These approaches based on handcrafted features are often sensitive to image variations such as changes in illumination and viewpoints. More importantly, these physical priors are not inappropriate for endoscopic image, causing inaccurate transmission estimates and undesirable desmoking results.

With the development of deep learning techniques, many methods based on convolutional neural network (CNN) have been proposed for haze and smoke removal. These methods do not rely heavily on statistical priors or haze-relevant



(a) smoky image (b) our desmoking result

Fig. 1 Examples of smoky image in endoscopic surgery and our desmoking result

attributes of the images to extract features. To guide the training process, most supervised dehazing models require several types of supervision information, such as transmission map, atmospheric light and haze-free image label. However, most of them still rely on ASM and should be trained by synthesizing images from depth information. Instead of depending on ASM, more works have explored a range of effective deep learning approaches for the direct learning of clear images, including attention mechanisms [5], knowledge distillation [6] and contrastive learning [7]. The CNN-based approaches usually have two major problems. First, CNNs traverse the entire image with the same convolutional kernel when conducting image feature extraction, so this is not the optimal solution to restore different image regions. Second, convolution is ineffective for long-range dependency modeling due to the notion of local processing. As a replacement for CNN, transformer [8] designs a self-attention mechanism that records global interactions across contexts and shows potential in a variety of visual challenges. We explore merging the above two ideas, employing the transformer as a feature extractor and incorporating the advantages of CNN to solve the image desmoking problem.

In this paper, we propose a surgical image desmoking model based on Swin transformer [9]. Our method first utilizes convolutional layers to capture shallow features, then employs Swin transformer block (STB) to extract deep features and ultimately outputs smoke-free images. The contributions of this work are summarized as follows:

- We propose a novel transformer-based method for surgical image desmoking which provides promising results.
- We introduce additional perceptual and structural losses on the basis of the pixel-to-pixel reconstruction loss of the Swin transformer framework to make the desmoking results more realistic.
- We adopt a rich set of objective and subjective evaluation criteria. To further validate the experimental results, we also employed a task-based evaluation method to demonstrate the desmoking performance by improving the computer vision task.

Related work

In recent years, many approaches have been explored to investigate general image dehazing and desmoking tasks to restore outdoor scenes affected by weather conditions. Typically, methods for smoke removal are either prior-based or learning-based.

Prior-based methods

Single image desmoking is an extremely ill-posed problem, and a variety of handcrafted and prior-based algorithms have been proposed to solve it. According to ASM, the hazing process is usually formulated as:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

where $I(x)$ is the observed hazy or smoky image, $J(x)$ is the scene radiance to be recovered, $t(x)$ is the transmission map and A is the global atmospheric light.

Prior-based dehazing and desmoking approaches leverage statistical features of clean images to estimate transmission maps, which are requisite by Eq. (1). He et al. [3] adopted dark channel prior (DCP), which posits that there exists at least one channel for every pixel whose value is close to zero. Berman et al. [10] assumed that a large number of distinct colors can accurately approximate the colors of a clear image, and introduce a non-local method for single image dehazing based on this prior. These methods have been demonstrated to be effective for image dehazing and desmoking, but their performance is intrinsically limited since the assumed priors are not appropriate for surgical images.

Learning-based methods

The most commonly used learning-based methods rely on training the synthetic images to obtain dehazing results. The supervised dehazing methods are classified into two categories, depending on whether ASM is used or not. The first is to estimate the parameters for dehazing process by combining ASM with CNN. These dehazing methods based on physical prior typically utilize depth information and atmospheric light values. The second is to directly learn the mapping from a hazy image to a haze-free image by data pairs.

Ren et al. [11] utilized an end-to-end approach to design a dehazing network via multi-scale convolutional neural networks with holistic edges by learning the mapping between hazy images and their transmission maps. Li et al. [12] proposed an all-in-one dehazing network (AOD-Net) to integrate the process of parameter estimation into a simple framework through the reformulated ASM. Instead of estimating the transmission map and atmospheric light, several end-to-end approaches have been presented to recover the clean image directly. Qin et al. [5] designed a feature attention (FA) module that included the channel attention and pixel attention when using CNN for feature extraction. Hong et al. [6] set the teacher network (T) as the image reconstruction task and made the student network (S) imitate this process. Reiter [13] proposed an image classifier based on CNNs with a recurrent architecture to provide temporal context in smoke recognition task.

CNN completes the image feature extraction from local to global information by continuously stacking convolutional layers, which gradually expand the perceptual field until the entire image is covered. Many studies have shown that its actual perceptual field is much smaller than the theoretical one that is not conducive to fully exploiting contextual information for feature extraction. CNNs perform better with low data volumes because of its inductive bias. Conversely, when a large amount of data is available, the inductive bias of CNNs limits the overall capability of the model. As a replacement for CNN, transformer [8] adopts the self-attention mechanism that allows to embed information globally across the overall image. Transformer performs self-attention across pixel patches to entirely provide the convolutional inductive bias. Transformer has the advantage of capturing global contextual information using attention mechanism, allowing it to create a long-range reliance on the image and extract features. Furthermore, Shikhar et al. [14] discovered that the transformer network not only outperforms CNN in image classification tasks, but also has a stronger shape bias and is more consistent with human perception after comparing with CNN. Recently, Swin transformer [9], which combines the advantages of CNN and transformer, has showed significant potential. Liang et al. [15] proposed a strong baseline model for image restoration which showed impressive performance on low-level vision tasks. Therefore, by combining the benefits of CNN and transformer, our proposed method has the advantage of CNN in processing images of huge size due to the local attention mechanism and the advantage of transformer in modeling long-range dependencies leveraging the shifted window.

Method

The architecture of our proposed method is shown in Fig. 2a. The network is comprised of three modules which is similar to [15]: feature extraction, further feature extraction and smoke-free image restoration modules.

Feature extraction

To extract the shallow feature, the first module $H_s(\cdot)$ consists of three convolutional layers. Given the input image $x \in R^{3\times H\times W}$ (H and W are the image height and width, respectively, and 3 means the R, G and B channels), the module generates a feature map $F_s \in R^{C\times H\times W}$ with C channels as

$$F_s = H_s(x) \tag{2}$$

Further feature extraction

Then, the further feature extraction module $H_f(\cdot)$ consists of N Swin transformer blocks (as shown in Fig. 2b) [9] and a 3 $\times$ 3 convolutional layer. We extract deep feature F_d from F_s as

$$F_d = H_f(F_s) \tag{3}$$

The Swin transformer block consists of a shifted window-based multi-head self-attention (MSA) module, followed by a two-layer multilayer perceptron (MLP) with GELU nonlinearity in between. Before each MSA and MLP module, a LayerNorm (LN) layer is applied, followed by a residual connection. Using a convolutional layer at the end of feature extraction can establish a better foundation for the later aggregation of shallow and deep features.

Image restoration

Similar to the first module, the image restoration module restores the smoke-free image using convolutional layers. The calculation is formulated as

$$I_r = H_r(F_s + F_d) \tag{4}$$

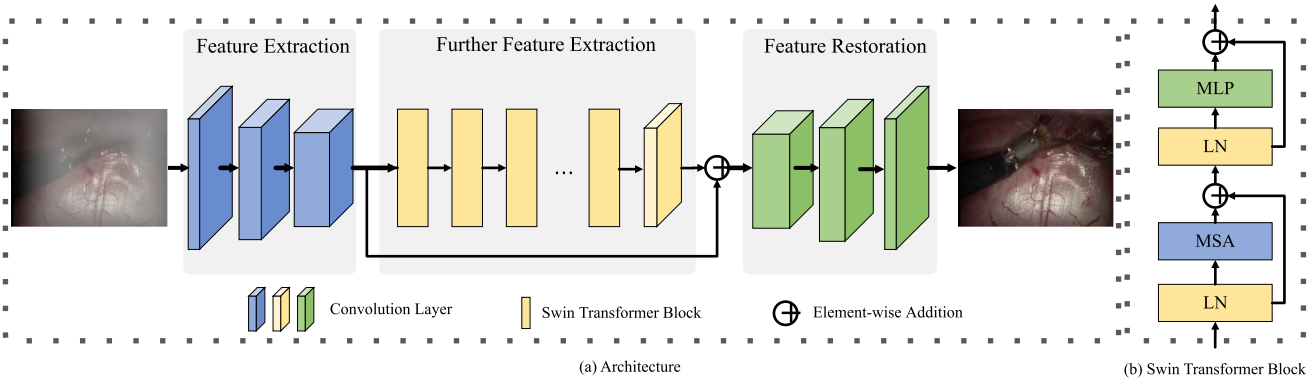


Fig. 2 Architecture of our proposed network for image desmoking

where $H_{\mathrm{r}}(\cdot)$ is the function of the reconstruction module. We provide skip connections which connect the shallow and deep feature maps to enhance feature fusion.

Loss function

We propose a supervised surgical image desmoking method to achieve the smoky-to-clear translation by introducing three loss functions, namely reconstruction loss, perceptual loss and SSIM loss. We employ the reconstruction loss which measures the pixel-wise difference between the desmoked and the ground truth images. The perceptual loss allows to assess the semantic difference between two images and measures visual similarity by comparing feature-level content. SSIM loss provides a measure of similarity by comparing the luminance, contrast and structural similarity information of two images. We optimize the parameters by a combination of the three losses, and the total loss function is defined as

$$L_{\text{total}} = \lambda_c L_{\text{Char}} + \lambda_p L_{\text{Prec}} + \lambda_s L_{\text{SSIM}} \tag{5}$$

where $\mathrm{L_{Char}}$, $\mathrm{L_{Prec}}$ and $\mathrm{L_{SSIM}}$ indicate reconstruction loss, perceptual loss and SSIM loss, respectively, and λ_c, λ_p and λ_s are trade-off weights.

Reconstruction loss

Instead of using the L2 loss function, we propose to train the network with the robust Charbonnier loss [16] to better handle outliers and improve the performance. As previously mentioned, I_{r} denotes the smoke-free image of input image x, and I_{gt} denotes the clean image. The Charbonnier loss is defined as

$$L_{\text{Char}} = \sqrt{||I_r - I_{gt}||^2 + \varepsilon^2} \tag{6}$$

where ε is a constant that is empirically set to 10^{-3}.

Perceptual loss

To evaluate the visual difference between the estimated image and the ground truth, the perceptual loss uses multi-scale features obtained from a pre-trained deep neural network. The definition of the perceptual loss is

$$L_{\text{Prec}} = \sum_{i=1}^{3} \frac{1}{C_i H_i W_i} ||\Phi_i(I_{gt}) - \Phi_i(I_{\mathrm{r}})||_2^2 \tag{7}$$

where $\Phi(\cdot)$ is the 16-layer VGG network pre-trained on the ImageNet dataset, $\Phi_i(\cdot)$ denote the ith VGG16 feature maps, and C_i, H_i and W_i indicate the dimension of $\Phi_i(\cdot)$.

SSIM loss

SSIM loss preserves the contrast in high-frequency regions (edge and details) better than the other loss functions. Including the SSIM factor in the loss function can retain the rich structural information of the restored image. The loss function is formulated as

$$\mathrm{SSIM} = \frac{(2\mu_r\mu_{gt}+C_1)(2\sigma_{r,gt}+C_2)}{(\mu_r^2+\mu_{gt}^2+C_1)(\sigma_r^2+\sigma_{gt}^2+C_2)} \tag{8}$$

$$L_{\text{SSIM}} = 1 - \mathrm{SSIM} \tag{9}$$

where μ_r and μ_{gt} represent the mean values of the I_r and I_{gt}, σ_r and σ_{gt} represent the standard deviations of the I_r and I_{gt}, $\sigma_{r,gt}$ represents the covariance of the I_r and I_{gt}, and C_1 and C_2 are constants set to avoid the denominator being 0.

Experiment

We use a vast scope of objective and subjective evaluation metrics. We conduct comprehensive tests on both synthetic and real-world smoke data sets, and our method outperform

the most advanced algorithms in both subjective visual and quantitative comparisons.

Implementation details

We implement our algorithm on the Pytorch framework. All the experiments are conducted on two NVIDIA 3080 GPUs. The entire training process is optimized by the ADAM solver with default parameters that β_1 and β_2 take the default values of 0.9 and 0.999, respectively. The learning rate is set to 10^{-4} with a decay rate of 0.5 for every 10 epochs with a batch size of 8. The patch size is 64×64, and the window size is set to 8 by default. The channel number and attention head number are generally set to 180 and 6, respectively. The trade-off weights in loss function are set to $\lambda_c = 1$, $\lambda_p = 0.01$, and $\lambda_s = 0.5$.

Datasets

Because of suffering from limited data in the surgical scenes, we use a large-scale synthetic dataset named RESIDE: V0 [17] for pre-training. We adopt all of the synthesized hazy/clean image pairs of indoor training set (ITS) and outdoor training set (OTS) in RESIDE: V0 dataset. In addition, we employ a 3D graphics rendering engine [18] to render smoke onto surgical images from 2017 EndoVis challenge dataset [19] as additional training data. We repurpose the weights of the pre-trained models to these images via fine-tuning to further improve performance. The physically based haze formation model to generate smoke as in [20] is not used since various spatial domains of the image might be influenced by different levels of smoke in surgical scenes. We generate different smoky images by using random parameters of position, density and intensity. In the fine-tuning phase, we unfreeze the base model and train the entire model end to end with a low learning rate and obtain incremental improvements for surgical scene by adjusting the weights appropriately.

Finally, we sample 120 non-smoky images and 120 real smoky images from the Hamlyn center laparoscopic/endoscopic video datasets [21] for testing. The non-smoky images are rendered using the same method to obtain the synthetic test dataset, and the real smoky images are used as the real test dataset.

Comparison methods

We employ two widely used metrics to evaluate each method for synthetic images with ground truth, i.e., peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

PSNR is a full-reference image quality evaluation metric which computes the differences between processed images and ground truths and is often employed in low-level image restoration. Given a desmoked image I_r and a clean image I_{gt}, the PSNR is defined as:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right) \tag{10}$$

where

$$\mathrm{MSE} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(I_{gt}(i,\ j) - I_r(i,\ j)\right)^2 \tag{11}$$

where W and H are the width and height of the desmoked and ground truth images, $I_{gt}(i,\ j)$ is the value of the clean image at location $(i,\ j)$ and $I_r(i,\ j)$ corresponds to the value of the generated smoke-free image. A better quality desmoked image equates to a larger value of PSNR.

The SSIM metric [22] is a well-known quality evaluation method to measure the similarity between two images and defined as Eq. (8). A good desmoking method has a high value of SSIM.

Since there is no ground truth in real-world cases, we adopt the Fog Aware Density Evaluator (FADE) [23] to evaluate the haze density of images. FADE is a non-reference method that does not require the original foggy or smoky image. A lower value of FADE implies better desmoking performance. Additionally, we use three well-known non-reference image quality evaluation metrics: natural image quality evaluator (NIQE), blind/referenceless image spatial quality evaluator (BRISQUE) [24] and spatial spectral entropy-based quality (SSEQ). For all metrics NIQE, BRISQUE and SSEQ, the lower value indicates better result. Finally, we assess the visual quality on both synthetic and real smoky images.

Table 1 Comparison with the different methods using the PSNR and SSIM. The best and suboptimal performances are indicated by bold and italics, respectively

Metrics	PSNR ↑	SSIM ↑
DCP	16.42	0.6709
NLD	18.25	0.6793
DehazeNet	19.71	0.6918
AOD-net	18.11	0.6667
FFA-net	*19.77*	0.6515
4KDehazing	18.43	*0.7913*
PSD	10.59	0.5733
D4	17.08	0.5600
Ours (pre-trained)	21.15	0.7875
Ours	**23.55**	**0.8567**

Bold and italics to indicate the best and suboptimal performance, respectively

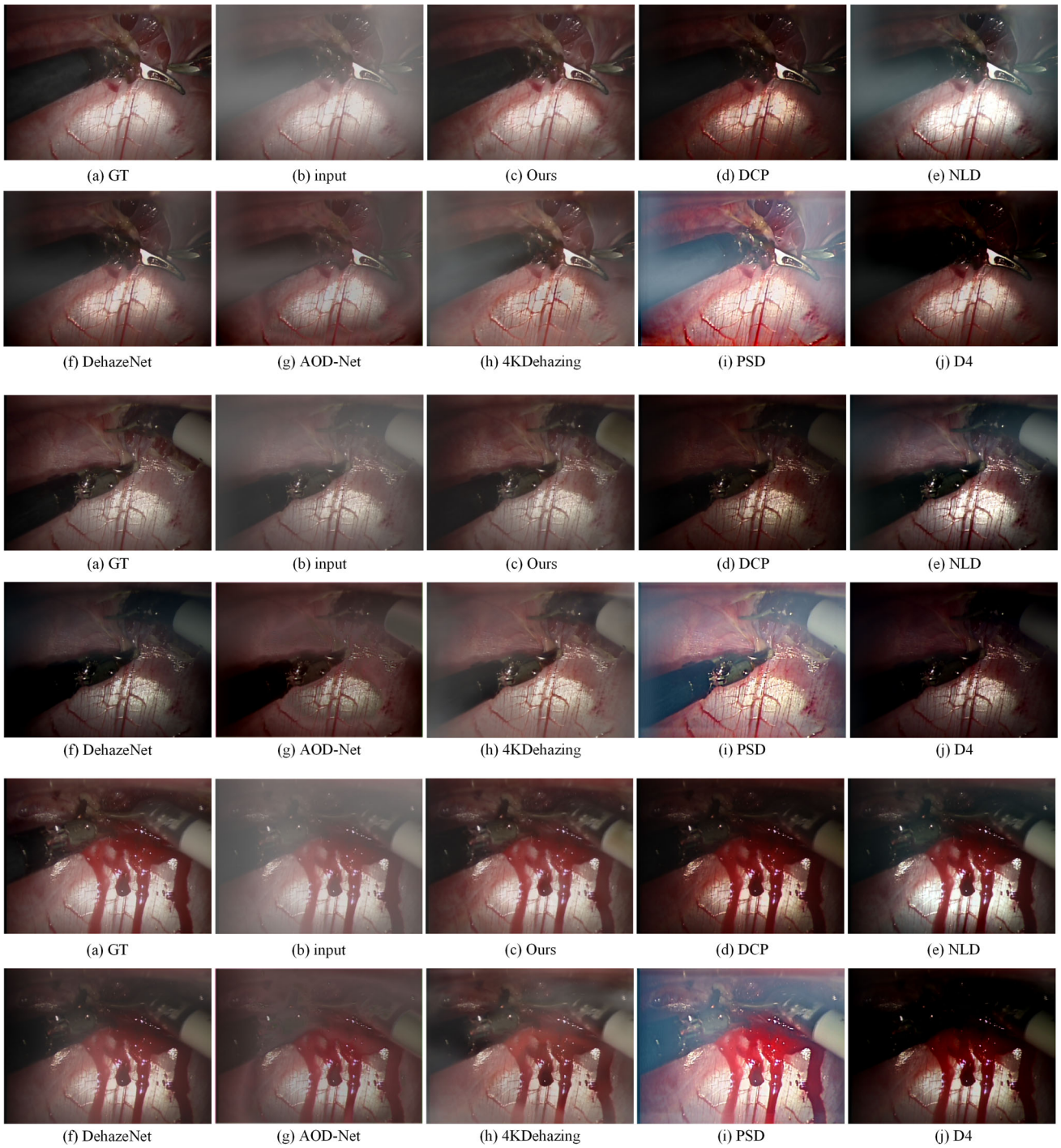


Fig. 3 Visual comparisons of desmoking results on synthetic images

We compare the proposed method with eight state-of-the-art desmoking and dehazing methods including both traditional image processing approaches (DCP [3], NLD [10]) and the most recent deep learning-based methods (DehazeNet [20], AOD-Net [12], FFA-Net [5], 4KDehazing [25], PSD [26] and D4 [27]).

Experiments on synthetic dataset

The effectiveness of our approach for smoke removal is assessed using both performance metrics and visual quality with our generated synthetic test dataset. Table 1 shows the quantitative results of our synthetic datasets with other approaches. First, our pre-trained model as well as the

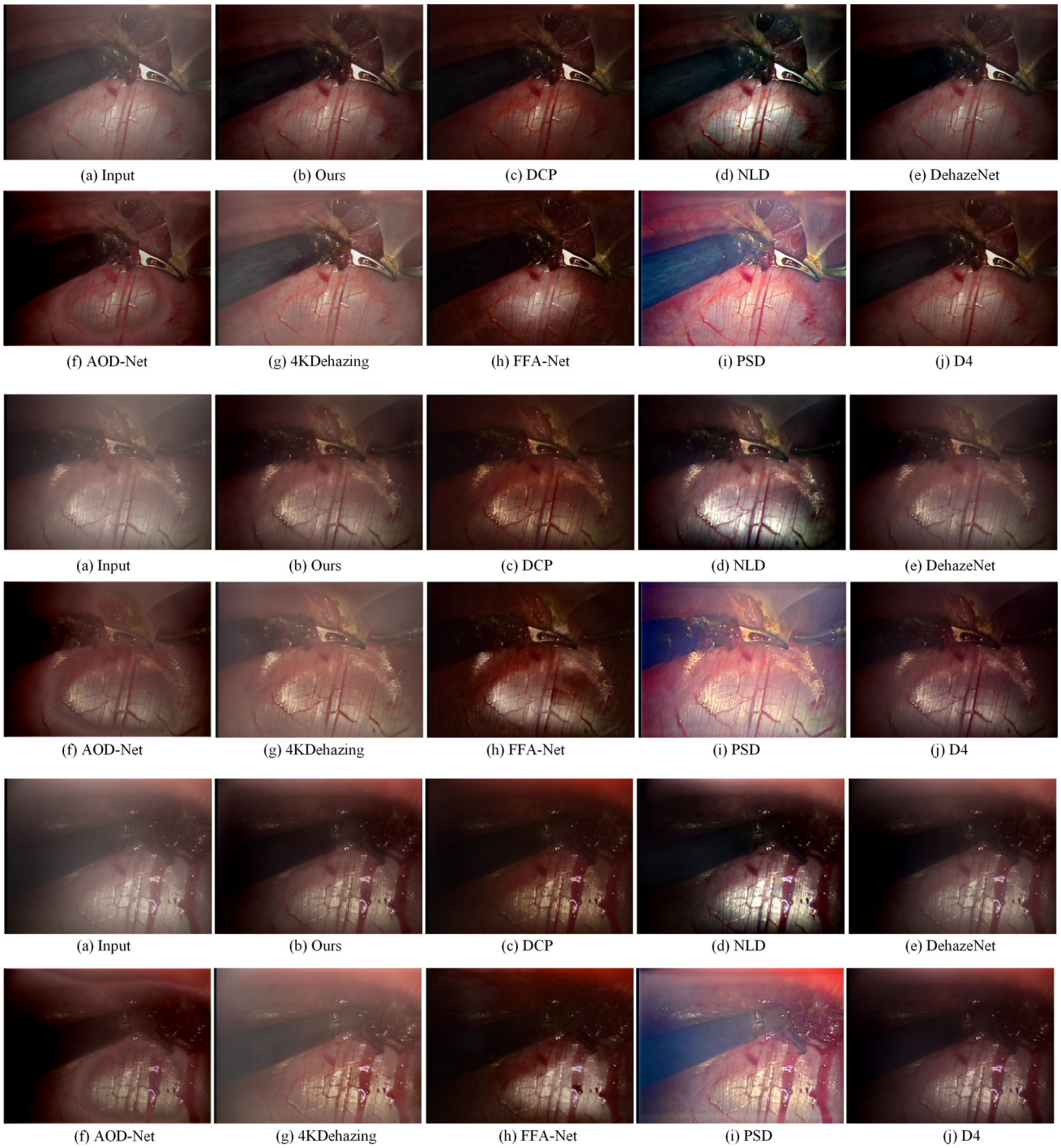


Fig. 4 Visual comparisons of desmoking results on real smoky images

final model obtains the highest PSNR and SSIM values and outperforms all of the other dehazing and desmoking methods, which demonstrates the robustness in terms of smoke removal task. Secondly, the handcrafted feature-based approaches [3, 10] do not perform effectively as expected. The performance of PSD is also unsatisfactory because it is based on a large number of physical priors. As a result, it is obvious that the physical priors of ASM are not applicable to surgical smoke removal. Moreover, compared to other indirect approaches, those methods [25, 27] that directly predict clear images based on end-to-end trainable networks achieve better results. FFA-Net also achieved better results, probably due to the use of the attention mechanism.

We also give the subjective comparison of desmoking results in Fig. 3. We observe that DCP and D4 often produce images with lower brightness than the original scene. The results of NLD and PSD suffer from some color distortion and seem unrealistic. DehazeNet, AOD-Net and FFA-Net do not remove the smoke completely and tend to output low-brightness images. Compared with these methods, our results are cleaner and visually pleasing with brighter details and sharper edges. Although optimizing SSIM loss directly makes our results ideal in SSIM metric, our proposed method also have better results on PSNR metric as well as subjective evaluation.

Experiments on real smoky dataset

We further evaluate our algorithm on real smoky images. Figure 4 illustrates the real smoky images and the desmoking results from state-of-the-art methods. Overall, the desmoking results are less effective than the synthetic dataset as shown. NLD and D4 suffers from serious color distortions. Images recovered using AOD-Net and 4KDehazing are still a little hazy, particularly in area at the edge. Although the desmoking results of DehazeNet are generally satisfactory, color distortion does occasionally occur.

For quantitative comparison, we observe that our method does not have the lowest FADE score as shown in Table 2. It is mainly because FADE is a patch-based evaluator to assess the fog density for an entire hazy image. However, it is based on statistical regularities observed on natural foggy and fog-free images, which always consider sharpness, contrast and saturation of the image. The fog aware features are derived from a reliable space domain natural scene statistics (NSS) model, so the FADE score is based on a natural scenario rather than a surgical scenario. Therefore, it is likewise acceptable that our experiment result is higher than NLD by only a small margin in terms of FADE. However, Fig. 4 shows that NLD has significant color distortion, while our method has a better visual result.

Our method acquires relatively good results in terms of NIQE, BRISQUE and SSEQ which are general non-reference image quality evaluation indicators based on natural scene statistics. DCP has the best and suboptimal performance in terms of NIQE and BRISQUE, respectively, while PSD has the best results in terms of SSEQ. However, there is no clear conclusion that which one of these general non-reference metrics can better evaluate the image quality. Figure 4 shows that PSD has serious color shifting and DCP has lower brightness than the original scene. Our intuition is that these metrics do not reflect the desmoking quality from the human vision. We find that Guo et al. [28] also had the same idea that the existing non-reference metrics are not fully suitable to assess the desmoking performance. As a result,

Table 2 Comparison with the different methods using the FADE, NIQE, BRISQUE and SSEQ. The best and suboptimal performances are indicated by bold and italics, respectively

Metrics	FADE ↓	NIQE ↓	BRISQUE ↓	SSEQ ↓
DCP	0.5700	**6.5710**	*25.04*	19.73
NLD	**0.4851**	7.0743	32.79	16.56
DehazeNet	0.6647	7.3431	29.16	30.48
AOD-net	0.5663	7.6234	**15.15**	30.95
4KDehazing	0.9475	*7.0084*	31.32	*15.91*
PSD	*0.5206*	7.6055	51.64	**11.67**
D4	0.6741	7.1164	27.72	22.53
Ours	0.6309	7.0539	26.15	19.56

Bold and italics to indicate the best and suboptimal performance, respectively

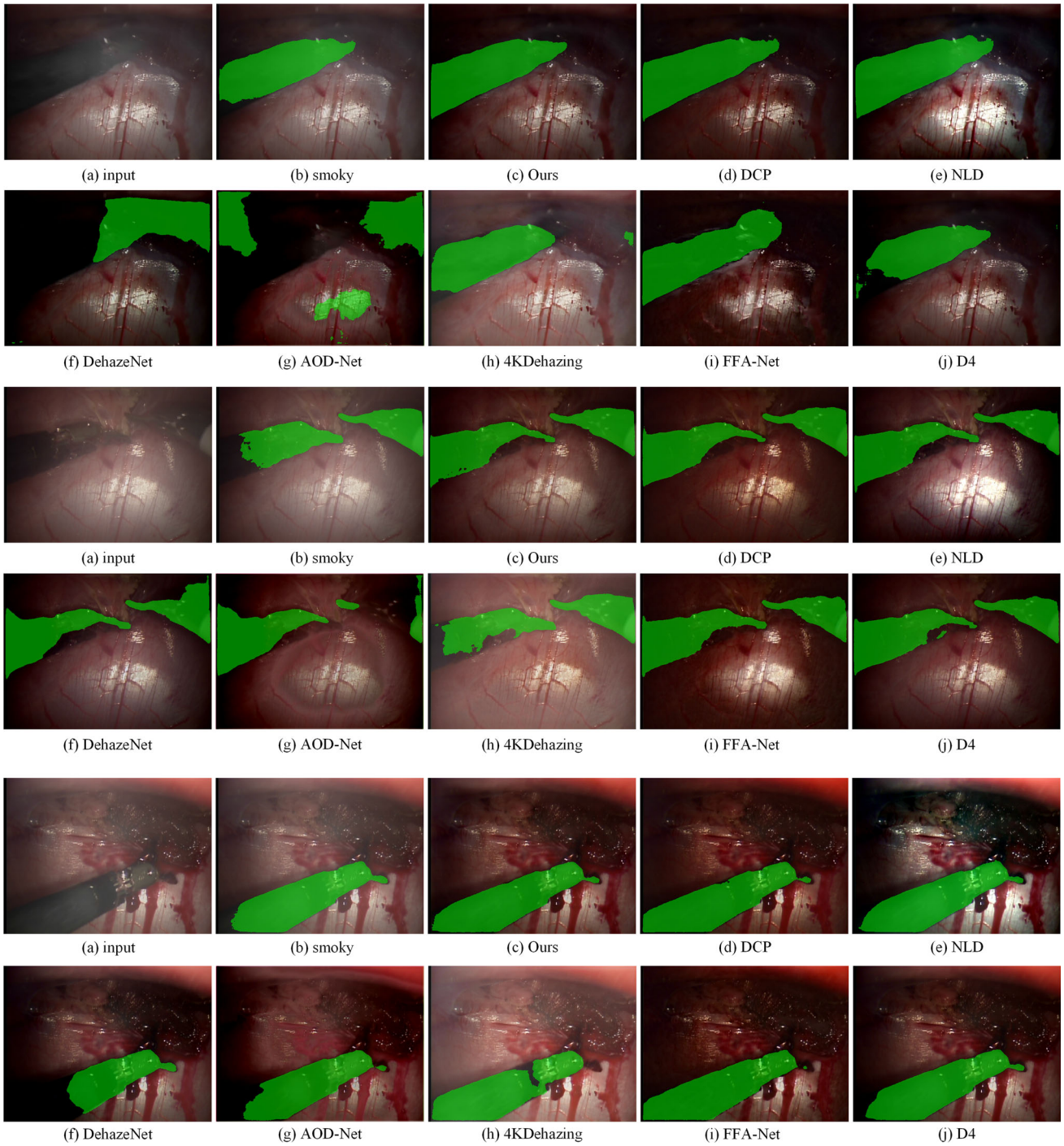


Fig. 5 Task-based comparisons of segmentation and desmoking results on real smoky images

developing more effective non-reference metrics is the most challenging and crucial work.

To better assess real smoky dataset, we evaluate task-based performance as an indirect indicator of the desmoking performance. Such a task-driven evaluation approach has still not received attention, but it has great potential for surgical applications. Therefore, we validate our desmoking results based on instrument segmentation tasks [29]. Since there is no ground truth for the segmentation results in the real smoky dataset, we only show the qualitative comparison results in Fig. 5. It is seen that most of the desmoking algorithms improve the segmentation effect. Traditional desmoking methods are generally more effective. Compared to other

deep learning-based algorithms, we obtain better segmentation results for the surgical instruments in our desmoked images. The segmentation results indirectly demonstrate the effectiveness of our proposed method and the future potential for other vision tasks.

Discussion and conclusion

In this paper, we propose a novel deep learning-based method which combines transformer with traditional CNN for surgical image desmoking. We utilize the self-attention mechanism for deep feature extraction. Compared to other desmoking algorithms which are based on various priors, our method conducts end-to-end feature extraction and generates clean images. The desmoking method augments the surgical field to clearly display surgical instruments and their environment during RMIS. Moreover, with task-based experimental validation, our proposed method could also be used as a preprocessing step for other high-level visual tasks, such as instrument segmentation, 3D scene reconstruction and surgery automation.

In order to evaluate the performance of our method, extensive experiments are conducted on synthetic and real smoky surgical images. The results of our method are satisfactory when assessed by both the subjective evaluation and the full-reference metrics. The major reason is that our method utilizes the transformer network to make sure that the desmoking output is identical to the clear image rather than focusing on designing a suitable prior.

We also observe that for the real smoky dataset, the four non-reference metrics do not reach a consensus and our method does not achieve the best results. Firstly, many factors, such as sharpness, contrast and saturation, are considered in the image quality metrics. Our method is trained and focused on restoring the realistic and natural surgical images without putting special emphasis on qualities like sharpness, contrast and saturation. Secondly, these image quality indicators are indeed inconsistent with human visual perception. Although many desmoking methods use these indicators to assess their results, there are some inaccuracies in the quantitative evaluation of desmoking. Furthermore, the desmoking results around the instrument are more important for the surgeon or other visual tasks. Therefore, the global non-reference metrics are not adequate to evaluate the desmoking results and task-based evaluation methods are more appropriate. In future work, we plan to find more accurate and reliable quantitative metrics for surgical smoke removal.

For real smoky dataset, we find that for thick smoke (as the third image in Fig. 4), all the desmoking methods do not work very well, especially around the instrument tip. The main reason is that the context information is missing for single image desmoking. We are going to extend our method to multi-frame desmoking because the temporal information between consecutive images in the videos is important.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11548-023-02835-z.

Author contributions Not applicable.

Funding This study was funded by the National Natural Science Foundation of China (Grant No. 51721003).

Data availability The public dataset used during the current study is available from (http://hamlyn.doc.ic.ac.uk/vision/ and https://endovissub2017-roboticinstrumentsegmentation.grandchallenge.org/).

Code availability Code will be publicly available with the publication of this work.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Consent to participate Not applicable.

Consent for publication Not applicable.

Informed consent This article does not contain patient data.

References

1. Lawrentschuk N, Fleshner NE, Bolton DM (2010) Laparoscopic lens fogging: a review of etiology and methods to maintain a clear visual field. J Endourol 24(6):905–913. https://doi.org/10.1089/end.2009.0594
2. Narasimhan SG, Nayar SK (2003) Contrast restoration of weather degraded images. IEEE Trans Pattern Anal Mach Intell 25(6):713–724. https://doi.org/10.1109/TPAMI.2003.1201821
3. He K, Sun J, Tang X (2011) Single image haze removal using dark channel prior. IEEE Trans Pattern Anal Mach Intell 33(12):2341–2353. https://doi.org/10.1109/TPAMI.2010.168
4. Fattal R (2014) Dehazing using color-lines. ACM Trans Graph (TOG) 34(1):1–14. https://doi.org/10.1145/2651362
5. Qin X, Wang Z, Bai Y, Xie X, Jia H (2020) FFA-Net: feature fusion attention network for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence, pp. 11908–11915
6. Hong M, Xie Y, Li C, Qu Y (2020) Distilling image dehazing with heterogeneous task imitation. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 3459–3468. https://doi.org/10.1109/CVPR42600.2020.00352
7. Wu H, Qu Y, Lin S, Zhou J, Qiao R, Zhang Z, Xie Y, Ma L (2021) Contrastive learning for compact single image dehazing. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 10546–10555. https://doi.org/10.1109/CVPR46437.2021.01041
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: 31st International conference on neural information processing systems, pp. 6000–6010

9. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF international conference on computer vision (ICCV), pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
10. Berman D, Treibitz T, Avidan S (2016) Non-local image dehazing. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp. 1674–1682. https://doi.org/10.1109/CVPR.2016.185
11. Ren W, Pan J, Zhang H, Cao X, Yang M-H (2019) Single image dehazing via multi-scale convolutional neural networks with holistic edges. Int J Comput Vis 128(1):240–259. https://doi.org/10.1007/s11263-019-01235-8
12. Li B, Peng X, Wang Z, Xu J, Feng D (2017) AOD-Net: all-in-one dehazing network. In: 2017 IEEE international conference on computer vision (ICCV), pp. 4780–4788. https://doi.org/10.1109/ICCV.2017.511
13. Reiter W (2021) Co-occurrence balanced time series classification for the semi-supervised recognition of surgical smoke. Int J Comput Assist Radiol Surg 16(11):2021–2027. https://doi.org/10.1007/s11548-021-02411-3
14. Tuli S, Dasgupta I, Grant E, Griffiths T (2021) Are convolutional neural networks or transformers more like human vision? In: Proceedings of the annual meeting of the cognitive science society
15. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R (2021) SwinIR: Image restoration using swin transformer. In: 2021 IEEE/CVF international conference on computer vision workshops (ICCVW), pp. 1833–1844. https://doi.org/10.1109/ICCVW54120.2021.00210
16. Charbonnier P, Blanc-Feraud L, Aubert G, Barlaud M (1994) Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st international conference on image processing, pp. 168–172, vol. 162. https://doi.org/10.1109/ICIP.1994.413553
17. Li B, Ren W, Fu D, Tao D, Feng D, Zeng W, Wang Z (2019) Benchmarking single-image dehazing and beyond. IEEE Trans Image Process 28(1):492–505. https://doi.org/10.1109/TIP.2018.2867951
18. Chen L, Tang W, John NW, Wan TR, Zhang JJ (2020) DesmokeGCN: generative cooperative networks for joint surgical smoke detection and removal. IEEE Trans Med Imaging 39(5):1615–1625. https://doi.org/10.1109/TMI.2019.2953717
19. Allan M, Shvets A, Kurmann T, Zhang Z, Duggal R, Su Y-H, Rieke N, Laina I, Kalavakonda N, Bodenstedt S, García-Peraza LC, Li W, Iglovikov V, Luo H, Yang J, Stoyanov D, Maier-Hein L, Speidel S, Azizian M (2019) 2017 Robotic instrument segmentation challenge
20. Cai B, Xu X, Jia K, Qing C, Tao D (2016) DehazeNet: an end-to-end system for single image haze removal. IEEE Trans Image Process 25(11):5187–5198. https://doi.org/10.1109/TIP.2016.2598681
21. Ye M, Johns E, Handa A, Zhang L, Pratt P, Yang G-Z (2017) Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. Hamlyn Symp Med Robot 2:1–2
22. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861
23. Choi LK, You J, Bovik AC (2015) Referenceless prediction of perceptual fog density and perceptual image defogging. IEEE Trans Image Process 24(11):3888–3901. https://doi.org/10.1109/TIP.2015.2456502
24. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708. https://doi.org/10.1109/TIP.2012.2214050
25. Zheng Z, Ren W, Cao X, Hu X, Wang T, Song F, Jia X (2021) Ultra-high-definition image dehazing via multi-guided bilateral learning. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 16180–16189. https://doi.org/10.1109/CVPR46437.2021.01592
26. Chen Z, Wang Y, Yang Y, Liu D (2021) PSD: principled synthetic-to-real dehazing guided by physical priors. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 7176–7185. https://doi.org/10.1109/CVPR46437.2021.00710
27. Yang Y, Wang C, Liu R, Zhang L, Guo X, Tao D (2022) Self-augmented unpaired image dehazing via density and depth decomposition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2037–2046
28. Guo X, Yang Y, Wang C, Ma J (2022) Image dehazing via enhancement, restoration, and fusion: a survey. Inf Fusion 86–87:146–170. https://doi.org/10.1016/j.inffus.2022.07.005
29. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI (2018) Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA), pp. 624–628. https://doi.org/10.1109/ICMLA.2018.00100