

Self-Supervised Video Desmoking for Laparoscopic Surgery

Renlong Wu¹, Zhilu Zhang¹(✉), Shuohao Zhang¹, Longfei Gou²,
Haobin Chen², Lei Zhang³, Hao Chen²(✉), and Wangmeng Zuo¹

¹ Harbin Institute of Technology, China

² Southern Medical University, China

³ Hong Kong Polytechnic University, China

hirenlongwu@gmail.com, cszlzhang@outlook.com,
yhyzshrby@163.com, Calvin_smu@163.com, HaoBin_Chen@outlook.com,
cszlzhang@comp.polyu.edu.hk, chenhao.05@163.com, wzmzuo@hit.edu.cn

Abstract. Due to the difficulty of collecting real paired data, most existing desmoking methods train the models by synthesizing smoke, generalizing poorly to real surgical scenarios. Although a few works have explored single-image real-world desmoking in unpaired learning manners, they still encounter challenges in handling dense smoke. In this work, we address these issues together by introducing the self-supervised surgery video desmoking (SelfSVD). On the one hand, we observe that the frame captured before the activation of high-energy devices is generally clear (named pre-smoke frame, PS frame), thus it can serve as supervision for other smoky frames, making real-world self-supervised video desmoking practically feasible. On the other hand, in order to enhance the desmoking performance, we further feed the valuable information from PS frame into models, where a masking strategy and a regularization term are presented to avoid trivial solutions. In addition, we construct a real surgery video dataset for desmoking, which covers a variety of smoky scenes. Extensive experiments on the dataset show that our SelfSVD can remove smoke more effectively and efficiently while recovering more photo-realistic details than the state-of-the-art methods. The dataset, codes, and pre-trained models are available at <https://github.com/ZcsrenlongZ/SelfSVD>.

Keywords: Laparoscopic Surgery Desmoking · Video Desmoking · Self-Supervised Learning

1 Introduction

Laparoscopy is employed to capture videos of the surgical sites to aid surgeons' decision-making, and it has found extensive application in the medical field [39]. However, during the surgery, the activation of high-energy devices (*e.g.*, electrocautery and ultrasonic scalpel) leads to the destruction and vaporization of proteins and fats, as well as the evaporation of liquid water, inevitably causing

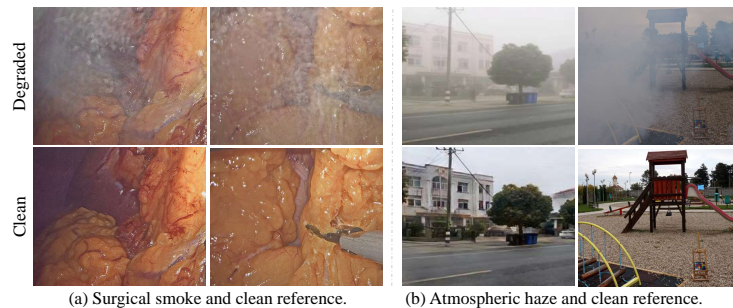


Fig. 1: The comparison of surgical smoke in (a) and atmospheric haze in (b). Degraded images are in the top row and their clean reference images are in the bottom row.

smoke [73]. The smoke may obscure specific tissue details, consequently diminishing the quality of laparoscopic imaging and impeding surgeons in making informed judgments [1]. Since it cannot be easily and quickly removed in vivo, post-processing laparoscopic images for desmoking [18] has become an effective and convenient manner to assist surgeons in observing the surgical sites clearly.

It is worth noting that although surgical smoke and atmospheric haze are somewhat similar in conformation, they are not exactly the same. As shown in Fig. 1, the latter is usually locally homogeneous and follows the atmospheric scattering model, while the former is difficult to physically simulate and contains more diverse situations, such as mist, droplets, and streaks. Consequently, the pre-trained dehazing models [5, 9, 16, 19, 29] are generally ineffective for desmoking. It is significant to give deliberate attention to both data and methods tailored for surgical smoke removal.

Recently, several attempts [36, 40, 59, 60, 71] have been made towards desmoking, while still facing some challenges. In this task, collecting paired data is difficult and even infeasible. To circumvent this problem, most existing methods employ a 3D graphics rendering engine [8, 23] to generate simulated smoky images [8, 49, 60] and videos [51]. However, the models trained on the synthetic data do not generalize well in real surgical scenarios, due to the domain gap between the synthetic smoke and the real-world one. In addition, some works [24, 43, 50, 53, 57] adopt unpaired learning manners [17, 74] for real-world single-image smoke removal. But they are unsatisfactory in handling dense smoke due to the inherent ill-posed nature. Additionally, few real-world video desmoking methods are explored.

This work aims to bring surgery video desmoking into the real world, which is more practical significance. Actually, sufficiently leveraging surgery video has the potential to address both the lack of real-world paired data and limited performance problems. On the one hand, the frames captured before the activation of high-energy devices usually have less smoke and similar contents as subsequent smoky ones. We designate the almost clear frame as a pre-smoke (PS) frame. PS frame can provide effective supervision information for smoky frames. Thus,

its reasonable utilization will make self-supervised real-world video desmoking feasible. On the other hand, the surgical smoke level fluctuates over time, either intensifying or diminishing. The neighboring frames may be complementary to the current one in smoke removal. Thus, video desmoking is expected to outperform single-image desmoking significantly in alleviating ill-posed problems.

Specifically, we propose a novel self-supervised surgery video desmoking (SelfSVD) method in this work. On the one hand, we utilize PS frame as misaligned supervision for smoky ones. In order to calculate the loss function between desmoking output and supervision accurately, a pre-trained optical flow estimation network (*e.g.*, PWC-Net [54]) is deployed to align the output with PS frame. On the other hand, PS frame is also available during the test stage, and is particularly important for handling complex and dense smoke in the real world. Thus, we further feed PS frame into models to help recover more details. Unfortunately, such a manner can easily lead to trivial solutions. To address this issue, we introduce a masking strategy and a regularization term. Besides, we suggest enhancing PS frame as better supervision by the above pre-trained SelfSVD model, and then taking it to fine-tune the model for improving visual effects.

We note that there is a dearth of real-world surgery video datasets for desmoking. In order to fill this gap, we collect multiple laparoscopic videos from professional hospitals and construct a laparoscopic surgery video desmoking (LSVD) dataset, which holds promise to benefit future studies. Extensive experiments are conducted on the dataset. The results show that our SelfSVD achieves better results than state-of-the-art methods in smoke removal and fine-scale detail recovery. Furthermore, we also design a lightweight model that can be inferred in real-time.

The contributions can be summarized as follows:

- We suggest utilizing the internal characteristics of real-world surgery videos for effective self-supervised video desmoking, and propose a SelfSVD solution, in which the pre-smoke (PS) frame serves as an unaligned supervision.
- We propose to take PS frame as an additional input to further improve desmoking performance, where a masking strategy and a regularization term are introduced to prevent trivial solutions.
- We construct a real-world laparoscopic surgery video desmoking (LSVD) dataset. Extensive experiments on the dataset demonstrate that our SelfSVD outperforms the state-of-the-art methods.

2 Related Work

2.1 Supervised Desmoking

Several studies [40, 55, 58–60, 71, 75] have endeavored to address the surgical smoke removal problem. Traditional approaches [55, 58, 75] estimate the transmission components based on the atmospheric scattering model, but easily lead to color and structure artifacts. Later works [36, 40, 59, 60, 71] adopt learn-based manners and design various networks based on convolutional neural networks

(CNN) and vision transformers [13]. For example, Lin *et al.* [36] and Ma *et al.* [40] modify U-net [48] architecture for surgical desmoking. Wang *et al.* [59] propose an encoder-decoder architecture with a Laplacian image pyramid decomposition strategy. Wang *et al.* [60] employ Swin Transformer blocks [35] to enhance feature extraction. Zheng *et al.* [71] further develop a system jointly detecting and removing surgical smoke. However, these methods generally simulate surgery smoke for training. Due to the domain gap between the synthetic smoke and the real-world one, they can't generalize well to real surgery scenes.

2.2 Supervised Dehazing

Dehazing [5, 11, 19, 20, 29, 30, 37, 45, 47, 63, 64, 67, 72] is of great relevance to surgical desmoking, which aims to recover clean components from hazy ones affected by adverse weather conditions. Li *et al.* [29] reformulate the atmospheric scattering model and propose AODNet for image dehazing. UHD [63] introduces the infinite approximation of Taylor's theorem with Laplacian pyramid pattern. DeHamer [19] further combine CNN [48] and vision Transformer [13] together for performance improvement. Compared to the above single-image methods, video ones [30, 37, 47, 64, 67, 67] can leverage temporal clues between consecutive frames for more effective restoration. Li *et al.* [30] first build an united video detection and dehazing framework that focuses on temporal information fusion. Ren *et al.* [47] incorporate global semantic priors for smooth transmission map estimation. Recently, MAPNet [64] explores physical haze priors to guide spatial information extraction and proposes a spatial-temporal alignment strategy to guide temporal features aggregation, achieving state-of-the-art performance. However, employing their pre-trained models directly does not produce satisfactory desmoking results due to the discrepancy between atmospheric haze and surgical smoke. Moreover, the lack of real-world paired data severely restricts the possibility of starting training from scratch. In this work, we propose a self-supervised framework for video desmoking (SelfSVD) to address these issues.

2.3 Unsupervised Desmoking and Dehazing

In contrast to supervised methods, unsupervised ones [8, 9, 15, 16, 24, 31–33, 43, 50, 53, 57, 65, 66, 70, 73] can be trained without paired supervision. For surgical smoke removal, Cyclic-DesmokeGAN [57] proposes a real-world image desmoking model based on CycleGAN [74]. Desmoke-LAP [43] and DCP-Pixel2Pixel [50] introduce dark channel prior [21] into loss and network design, respectively. MS-CycleGAN [53] adapts a model pre-trained on synthetic data to real-world ones.

Apart from the above unsupervised desmoking methods, unsupervised dehazing ones [15, 31–33, 65, 66, 70] are also widely explored. YOLY [31] and ZID [32] perform dehazing in a zero-shot manner by disentangling the hazy image into the clean component and other ones. RefinetNet [70], DistentGAN [65], and USID-Net [33] build their models under GAN [17] framework. D^4 [66] converts the transmission map estimation into density and depth image prediction, achieving better results. CycleDehaze [15] introduces a Laplacian pyramid network to

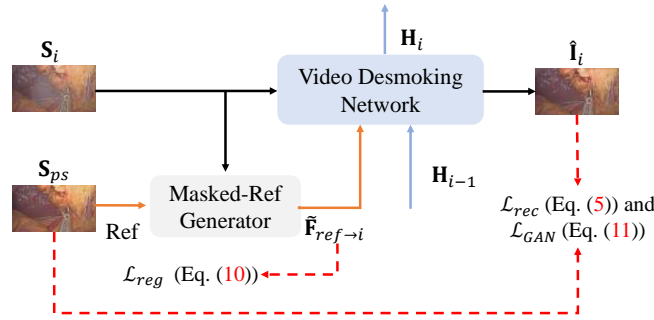


Fig. 2: The illustration of processing the i -th smoky frame S_i . PS frame (S_{ps}) is taken as both supervision and reference (Ref) input. A masking strategy with the masked-ref generator as shown in Fig. 4 and a regularization term as Eq. (10) are introduced to prevent trivial solutions. H_{i-1} is the temporal features from previous frames and H_i is the temporal features for subsequent ones.

handle high-resolution images efficiently. However, these methods suffer from unstable training and fail to process dense smoke. In contrast, our SelfSVD enables more stable and finer-scale restoration by sufficiently utilizing pre-smoke frames.

3 Proposed Method

We first describe the approach towards self-supervised learning, *i.e.*, taking pre-smoke (PS) frame as supervision in Sec. 3.1. Then we detail how to take PS frame as reference input and introduce the solutions to prevent trivial solutions in Sec. 3.2. Next, we introduce the way to enhance PS frame as better supervision for improving visual effects in Sec. 3.3. Finally, the details about the network architecture and learning objective are provided in Sec. 3.4.

3.1 Taking PS Frame as Supervision

Given a smoky video consisting of N frames $\{S_i\}_{i=1}^N$, video desmoking aims to restore the corresponding clean components $\{\hat{I}_i\}_{i=1}^N$ with temporal clues, *i.e.*,

$$\{\hat{I}_i\}_{i=1}^N = \mathcal{D}(\{S_i\}_{i=1}^N; \Theta_{\mathcal{D}}), \quad (1)$$

where \mathcal{D} denotes the video desmoking model with the parameter $\Theta_{\mathcal{D}}$. However, the paired smoky-clean videos are difficult to acquire in the real world. Several studies [40, 55, 58–60, 71, 75] utilize synthetic data for training, but the domain gap between synthetic smoke and real one hinders their effective applications in the real world. A few single-image unpaired methods [24, 43, 50, 53, 57] are explored to overcome this issue, but their training is less stable and their ill-posed nature makes them less effective in processing dense smoke. In contrast to

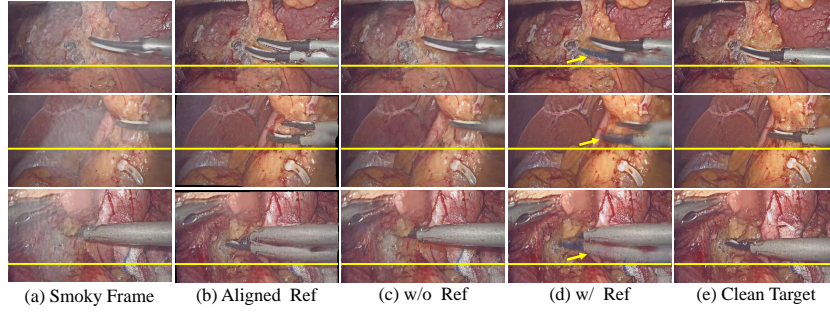


Fig. 3: Examples of trivial solutions. When inputting PS frame as Ref naively, the imperfect optical flow between Ref and the smoky frame leads to trivial solutions, as indicated by yellow arrows. The same positions are marked with yellow lines.

these approaches, we propose to utilize the internal characteristics of real-surgery video for self-supervised video desmoking to address these issues together, as shown in Fig. 2. Specifically, we note that the pre-smoke frame (PS frame, \mathbf{S}_{ps}) is clearer than subsequent smoky ones and can be regarded as their supervision.

Motivated by this, we introduce SelfSVD to achieve self-supervised learning, where we can train the video desmoking network \mathcal{D} as,

$$\Theta_{\mathcal{D}}^* = \arg \min_{\Theta_{\mathcal{D}}} \mathcal{L}(\mathcal{D}(\{\mathbf{S}_i\}_{i=1}^N; \Theta_{\mathcal{D}}), \mathbf{S}_{ps}), \quad (2)$$

where \mathcal{L} denotes the learning objective.

It is worth noting that $\hat{\mathbf{I}}_i$ is not spatially aligned with \mathbf{S}_{ps} . Taking \mathbf{S}_{ps} as supervision directly may lead to blurry results [34, 67, 69]. Instead, we adopt a deformation-based learning objective to tolerate the misalignment. In particular, a pre-trained optical flow network \mathcal{O} is first used to estimate the optical flow $\Psi_{ps \rightarrow i}$ from \mathbf{S}_{ps} to $\hat{\mathbf{I}}_i$, *i.e.*,

$$\Psi_{ps \rightarrow i} = \mathcal{O}(\mathbf{S}_{ps}, \hat{\mathbf{I}}_i). \quad (3)$$

Then $\hat{\mathbf{I}}_i$ is back warped towards \mathbf{S}_{ps} with a warping operation \mathcal{W} according to the estimated optical flow $\Psi_{ps \rightarrow i}$, *i.e.*,

$$\hat{\mathbf{I}}_{i \rightarrow ps} = \mathcal{W}(\hat{\mathbf{I}}_i, \Psi_{ps \rightarrow i}). \quad (4)$$

Finally, $\hat{\mathbf{I}}_{i \rightarrow ps}$ is spatially aligned with \mathbf{S}_{ps} . The reconstruction loss \mathcal{L}_{rec} of the desmoking model can be written as,

$$\mathcal{L}_{rec} = \sum_{i=1}^N \|\mathbf{V}_i \odot (\hat{\mathbf{I}}_{i \rightarrow ps} - \mathbf{S}_{ps})\|_1. \quad (5)$$

\odot is a pixel-wise multiply operation. \mathbf{V}_i is a mask that indicates the valid positions of optical flow. The j -th element of \mathbf{V}_i can be calculated as,

$$\mathbf{V}_i^j = \text{sgn}(\max(0, [\mathcal{W}(\mathbf{1}, \Psi_{ps \rightarrow i})]_j - \tau)). \quad (6)$$

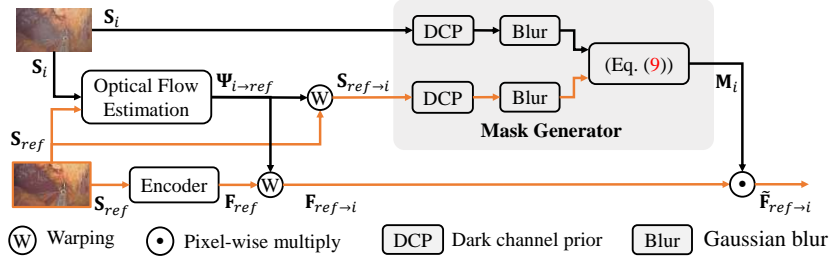


Fig. 4: The structure of masked-ref generator. The mask generator is used to generate a mask M_i , which is employed to produce the masked reference features $\tilde{F}_{ref \rightarrow i}$.

\mathcal{W} is warping operation. \max and sgn are the maximum and sign function, respectively. $\mathbf{1}$ is an all-1 matrix, $[\cdot]_j$ denotes the j -th element. τ is set to 0.999.

3.2 Taking PS Frame as Reference Input

We note that PS frame is available during both training and testing. Thus, beyond taking PS frame as supervision, we can further take the PS frame as a reference (Ref) input to guide smoke removal. Denote by S_{ref} the Ref (*i.e.*, S_{ps}), Eq. (2) can be modified as,

$$\Theta_{\mathcal{D}}^* = \arg \min_{\Theta_{\mathcal{D}}} \mathcal{L}(\mathcal{D}(\{S_i\}_{i=1}^N, S_{ref}; \Theta_{\mathcal{D}}), S_{ps}). \quad (7)$$

Denote the features of Ref and the i -th smoky frame by F_{ref} and F_i respectively (see Sec. 3.4 for details). The spatial misalignment between them should be addressed first. Thus, we estimate the optical flow $\Psi_{i \rightarrow ref}$ from the i -th smoky frame S_i to S_{ref} . Then, we back warp F_{ref} to F_i according to $\Psi_{i \rightarrow ref}$, generating the warped reference features $F_{ref \rightarrow i}$.

When $\Psi_{i \rightarrow ref}$ is perfectly estimated, $F_{ref \rightarrow i}$ is naturally perfectly aligned with F_i . In this case, the desmoking model can produce better results. However, it is not realistic to achieve perfect optical flow estimation due to the interference of surgical smoke and content occlusion. Some contents (*e.g.*, areas where high-energy devices move significantly) in F_{ref} may not be correctly mapped to the corresponding positions in F_i , being kept in $F_{ref \rightarrow i}$. Moreover, the desmoking model tends to utilize features from $F_{ref \rightarrow i}$ rather than F_i , as the former contains more clean information relevant to supervision. Thus, the model easily over-fits the Ref, as shown in Fig. 3. To circumvent the issue, we introduce a hard masking strategy and supplement a soft regularization term, as described below.

Masking Strategy. In regions with significantly inaccurate optical flow, we prefer to forget the reference information to avoid output contents being inconsistent with smoky inputs. As shown in Fig. 4, we suggest generating a mask M_i to indicate these regions in $F_{ref \rightarrow i}$, and exclude corresponding features, *i.e.*,

$$\tilde{F}_{ref \rightarrow i} = M_i \odot F_{ref \rightarrow i}. \quad (8)$$

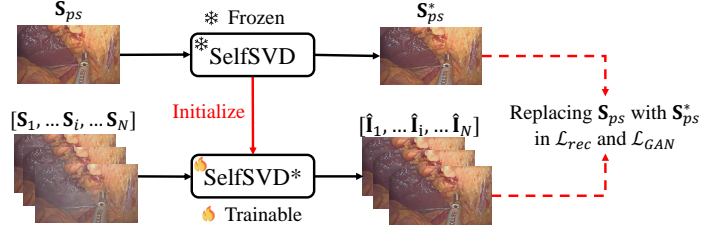


Fig. 5: The illustration of enhancing PS frame (\mathbf{S}_{ps}) as supervision. We regard \mathbf{S}_{ps} as a frame with less smoke and feed it into a pre-trained SelfSVD model, generating a cleaner result \mathbf{S}_{ps}^* . Then, \mathbf{S}_{ps}^* is taken as improved supervision to fine-tune the SelfSVD model by \mathcal{L}_{rec} and \mathcal{L}_{GAN} (*i.e.*, replacing \mathbf{S}_{ps} with \mathbf{S}_{ps}^* in Eq. (5) and Eq. (11)), getting an improved model named SelfSVD*.

As shown in Fig. 4, to obtain \mathbf{M}_i , we first align \mathbf{S}_{ref} to \mathbf{S}_i , obtaining the warped reference image $\mathbf{S}_{ref \rightarrow i}$. Then, we process $\mathbf{S}_{ref \rightarrow i}$ and \mathbf{S}_i with DCP [21] and large-kernel Gaussian blur operations to alleviate the disturbance of surgery smoke. Next, we divide $\mathbf{S}_{ref \rightarrow i}$ and \mathbf{S}_i into P non-overlapping patches respectively, *i.e.*, $\{\mathbf{S}_{ref \rightarrow i}^p\}_{p=1}^P$ and $\{\mathbf{S}_i^p\}_{p=1}^P$. For each pair $\mathbf{S}_{ref \rightarrow i}^p$ and \mathbf{S}_i^p , their structure should be significantly distinct when optical flow is estimated inaccurately. So we adopt structural similarity (SSIM) metric [61] to detect such patch, *i.e.*,

$$m_i^p = \text{sgn} \left(\max \left(0, \text{SSIM}(\mathbf{S}_{ref \rightarrow i}^p, \mathbf{S}_i^p) - \epsilon \right) \right). \quad (9)$$

m_i^p is the value of \mathbf{M}_i in the p -th patch. ϵ is a threshold and is set to 0.92. The patch size is set to 8. Please see more details about mask in the Suppl.

Regularization Strategy. The areas where optical flow is slightly wrong may lead to potential trivial solutions, as they are not easily detected explicitly. We use ℓ_1 regularization [3, 12] to constraint $\mathbf{F}_{ref \rightarrow i}$ in these areas (indicated by mask \mathbf{M}_i) to be sparse. The regularization loss \mathcal{L}_{reg} can be written as,

$$\mathcal{L}_{reg} = \|\mathbf{M}_i \odot \mathbf{F}_{ref \rightarrow i}\|_1. \quad (10)$$

3.3 Enhancing PS Frame as Supervision

For model training, there are two possible ways to introduce PS frame. One is to always adopt the starting frame in surgery, and another one is to select it dynamically as the surgery proceeds. We adopt the latter way, as the video contents may change significantly, causing starting frames to provide insufficient information for long-distance ones. However, in this manner, some smoke generated by previous activation of high-energy devices may remain in \mathbf{S}_{ps} . Thus, naively taking \mathbf{S}_{ps} as supervision may not be the optimal solution.

Taking this into account, we improve the training of SelfSVD by enhancing \mathbf{S}_{ps} as better supervision, dubbed SelfSVD*, as shown in Fig. 5. In particular, a SelfSVD model is first pre-trained under the supervision of \mathbf{S}_{ps} . Then, we regard

\mathbf{S}_{ps} as a frame with less smoke and feed it into the pre-trained SelfSVD, getting a cleaner result \mathbf{S}_{ps}^* . Finally, a SelfSVD* model is obtained by fine-tuning the pre-trained SelfSVD with \mathbf{S}_{ps}^* as supervision, and the original SelfSVD model can be detached during testing.

3.4 Network Architecture and Learning Objective

Network Architecture. Considering the practical application, the desmoking model needs to process surgery videos online, rather than offline. Therefore, we design a video desmoking network based on the unidirectional recurrent neural network [62]. According to the function of each component, it can be divided into five modules, *i.e.*, feature encoder, masked-ref generator, alignment, fusion, and reconstruction module. When processing the i -th smoky frame \mathbf{S}_i , we first pass \mathbf{S}_i to the encoder to obtain its features \mathbf{F}_i . \mathbf{S}_{ref} and \mathbf{S}_i are fed into the masked-ref generator to get the masked reference features $\tilde{\mathbf{F}}_{ref \rightarrow i}$, which has been introduced in Sec. 3.2. Then, we deploy the alignment module based on the optical flow to align previous temporal features \mathbf{H}_{i-1} to \mathbf{F}_i , getting the warped ones $\mathbf{H}_{i-1 \rightarrow i}$. Finally, the fusion module takes \mathbf{F}_i , $\tilde{\mathbf{F}}_{ref \rightarrow i}$ and $\mathbf{H}_{i-1 \rightarrow i}$ as inputs to get the fused feature representations, which are passed to the reconstruction module for generating the restored clean component $\hat{\mathbf{I}}_i$. Please see more details about the network architecture in the Suppl.

Learning Objective. To further improve the visual quality, we adopt adversarial loss [41] to train our desmoking networks, which can be written as,

$$\mathcal{L}_{GAN} = \frac{1}{2} \mathbb{E}_{\mathbf{S} \sim \mathcal{P}_{\mathbf{S}}} [\mathcal{DISC}(\mathcal{D}(\mathbf{S}, \mathbf{S}_{ref})) - 1]^2, \quad (11)$$

where \mathbf{S} denotes the smoky video $\{\mathbf{S}_i\}_{i=1}^N$, and \mathcal{DISC} is the discriminator [74] (see the Suppl. for detailed structure). The discriminator is trained by,

$$\begin{aligned} \mathcal{L}_{DISC} = & \frac{1}{2} \mathbb{E}_{\mathbf{S}_{ps} \sim \mathcal{P}_{\mathbf{S}_{ps}}} [\mathcal{DISC}(\mathbf{S}_{ps}) - 1]^2 \\ & + \frac{1}{2} \mathbb{E}_{\mathbf{S} \sim \mathcal{P}_{\mathbf{S}}} [\mathcal{DISC}(\mathcal{D}(\mathbf{S}, \mathbf{S}_{ref}))]^2. \end{aligned} \quad (12)$$

Overall, combined Eq. (5), Eq. (10) and Eq. (11), the loss terms of SelfSVD can be written as,

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{GAN} \mathcal{L}_{GAN}, \quad (13)$$

where λ_{reg} and λ_{GAN} are set to 0.05 and 1.0, respectively. When training SelfSVD* model, \mathbf{S}_{ps} in Eq. (13) is replaced with \mathbf{S}_{ps}^* , as stated in Sec. 3.3.

4 LSVD Dataset

Cyclic-DesmokeGAN [57] and Desmoke-LAP [43] collect 1,200 smoky images and 3,000 ones from cholecystectomy and hysterectomy surgery recordings, respectively. However, they are only used for single-image desmoking, being unsuitable

for our video desmoking. Sengar *et al.* [51] propose a video desmoking dataset but the smoke is simulated by the 3D graphics rendering engine [23]. As far as we know, real-world laparoscopic surgery video desmoking datasets are currently scarce, which implies a high demand for this dataset.

In this work, we construct the LSVD dataset by collecting laparoscopic surgery videos of 40 patients from professional hospitals. We first select the frames manually where surgeons start activating the high-energy devices in each video. Then, we take its preceding frame as PS frame and several subsequent ones (until the one in which smoke is nearly dismissed) as smoky frames. Finally, 486 video clips are collected, where each clip contains 20~50 frames with a resolution of 1080×1920 . 416 clips are used for the training set, and the remaining 70 ones are used for the testing set. The dataset covers diverse and complex real-world surgical smoke. We provide some examples in the Suppl.

5 Experiments

5.1 Implementation Details

For optical flow estimation, we adopt a pre-trained PWC-Net [54]. During training, we randomly crop patches and augment them with random flips. The batch size is set to 4 and the patch size is set to 256×256 . SelfSVD is trained with ADAM optimizer [28] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ for 100k iterations. Cosine annealing strategy [38] is employed to decrease the learning rate from 1×10^{-4} to 1×10^{-7} steadily. For the training of SelfSVD*, we fine-tune the pre-trained SelfSVD model for additional 40k iterations and set the initial learning rate to 5×10^{-5} . All experiments are conducted with PyTorch [44] on a single Nvidia GeForce RTX A6000 GPU.

SelfSVD and SelfSVD* keep the same number of residual blocks [22] as BasicVSR++ [7], and the computation costs are generally similar to the video processing methods. Moreover, to make the model inference cost consistent with some single-image processing methods, we present the lightweight versions by reducing the numbers and channels of residual blocks, dubbed SelfSVD-S and SelfSVD*-S respectively. Please see more details in the Suppl.

5.2 Evaluation Configurations

On the one hand, as smoky frames have no paired clean frames, we have to utilize PS frame to evaluate desmoking methods. To remove the possible smoke in PS frame, we first feed it into a pre-trained SelfSVD model and take the pre-processed PS frame as the clean target. Then we adopt the aligned PSNR [26] and SSIM [61] as the reference evaluation metrics, following similar works [2, 14, 62]. Specifically, we align desmoking results to the target by a pre-trained optical flow network (*i.e.*, PWC-Net [54]) and calculate PSNR and SSIM between aligned results and the target. Simultaneously, we also report the metrics when taking the original PS frame as the target in the Suppl. On the other hand, we employ no-reference metrics (*i.e.*, FADE [10], NIQE [42], and PI [4]) to assess the desmoking

Table 1: Quantitative comparison on LSVD dataset. \uparrow denotes the higher metric the better, and \downarrow denotes the lower one the better. The best results in each category are marked in **bold**.

Methods		PSNR \uparrow	SSIM \uparrow	FADE \downarrow	NIQE \downarrow	PI \downarrow	#FLOPs (G)	#Time (s)	#Params (M)
Unsupervised Image Processing	PSD [9]	14.69	0.3722	0.1715	5.44	4.37	905	0.31	31.11
	DCP [21]	17.77	0.5582	0.3217	5.53	4.25	-	0.08	-
Unpaired Image Processing	DCP-Pixel2Pixel [50]	19.65	0.5079	0.5509	6.30	4.50	195	0.02	54.40
	DistntGAN [52]	21.51	0.6037	0.6135	4.95	3.60	1572	0.22	11.38
	Desmoke-LAP [43]	22.52	0.6170	0.5386	4.75	3.93	1571	0.15	11.38
	RefineNet [70]	22.87	0.6177	0.5749	6.93	4.13	1780	0.34	0.85
	UHD [63]	20.93	0.5792	1.3573	6.26	6.18	123	0.27	34.55
Self-Supervised Image Processing	MSDesmoking [59]	21.53	0.5997	1.1290	5.20	5.19	605	0.04	8.80
	Wang <i>et al.</i> [60]	22.65	0.6078	1.1202	7.63	7.11	6584	28.20	3.19
	MSBDN [11]	22.69	0.6093	0.8995	6.07	5.40	779	0.29	31.35
	AODNet [29]	23.03	0.6116	1.0417	5.96	4.79	4	0.08	0.002
	DADFNet [20]	23.06	0.6143	0.4948	5.29	4.08	198	0.03	0.85
	Dehamer [19]	23.13	0.6184	1.1508	6.93	6.36	1564	0.41	132.45
	BasicVSR [6]	23.00	0.6168	0.5609	5.60	4.13	831	0.13	6.29
Self-Supervised Video Processing	BasicVSR++ [7]	23.35	0.6196	0.5665	5.50	3.90	1197	0.19	9.76
	MAPNet [64]	23.28	0.6152	0.9632	5.53	5.43	261	0.20	28.75
	(Ours) SelfSVD-S	23.84	0.6183	0.4787	5.04	3.94	169	0.03	1.92
	(Ours) SelfSVD*-S	24.00	0.6209	0.4404	4.87	3.92	169	0.03	1.92
	(Ours) SelfSVD	24.23	0.6216	0.4626	4.85	3.87	996	0.18	15.58
	(Ours) SelfSVD*	24.58	0.6279	0.4193	4.72	3.86	996	0.18	15.58

results. In addition, the number of model parameters and FLOPs, as well as inference time per frame when inputting 1080×1920 videos are also reported.

5.3 Comparison with State-of-the-Art Methods

Related learning-based methods are either trained on synthetic data in a supervised manner [6, 7, 11, 19, 20, 29, 59, 63, 64], or trained in an unpaired manner [43, 50, 52, 70]. For fair comparisons, we retrain their models on our LSVD dataset. For training supervised methods, there is no paired ground-truth. Instead, we take the aligned PS frame as the supervision, which is the same as the setting of training our models as stated in Sec. 3.1. Thus, they become self-supervised methods. For unpaired learning methods, we take the original PS frame as their unpaired supervision. Overall, we compare our methods (*i.e.*, SelfSVD-S and SelfSVD*-S, SelfSVD, SelfSVD*) against 16 related state-of-the-art ones, including 2 unsupervised image processing methods (*i.e.*, PSD [9], DCP [21]), 4 unpaired image ones (*i.e.*, DCP-Pixel2Pixel [50], DistntGAN [52], and Desmoke-LAP [43], RefineNet [70]), 7 self-supervised image ones (*i.e.*, UHD [63], MSDesmoking [59], Wang *et al.* [60] MSBDN [11], AODNet [29], DADFNet [20], Dehamer [19],) and 3 self-supervised video ones (*i.e.*, BasicVSR [6], BasicVSR++ [7], and MAPNet [64]). We do not compare with unsupervised and unpaired video ones, as few works explore that as far as we know.

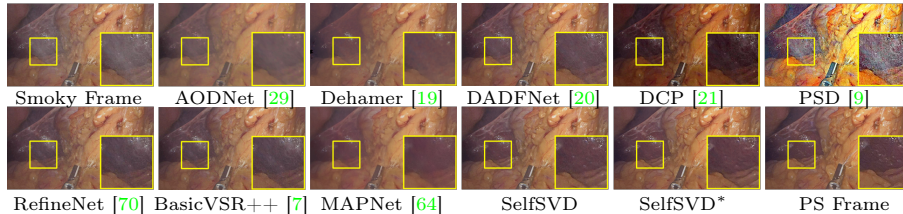


Fig. 6: Qualitative comparison on LSVD dataset. Our methods generate results more consistent with the PS frame.

Beyond conducting experiments on our dataset, we provide more comparisons on a synthetic dataset in the Suppl.

Quantitative Comparison. Tab. 1 summarizes the quantitative results. First, video processing methods generally perform better than single-image ones, which demonstrates the benefits of utilizing temporal clues for surgical smoke removal. Second, our methods outperform state-of-the-art ones by a large margin. Among all competing methods, BasicVSR++ [7] has achieved the best PSNR score of 23.35dB. Our SelfSVD and SelfSVD* achieve PSNR gains of 0.88dB and 1.23dB over BasicVSR++ [7], respectively. Although PSD [70] and DCP [21] get better FADE scores, they generally over-process smoky videos and result in over-saturated colors (as shown in Fig. 6), leading to poor NIQE and PI scores, as well as unsatisfactory visual effects. Third, with lower computation cost and fewer model parameters, the proposed SelfSVD-S and SelfSVD*-S still perform well, which further illustrates the effectiveness of our methods.

Qualitative Comparison. Due to space limitation, we only provide qualitative results of some methods with better quantitative scores in Fig. 6. It can be seen that the compared methods often introduce color distortion, over-smoothing, or smoke preservation in their results. Instead, our methods remove more smoke and restore more details that are consistent with PS frames. More qualitative comparison results can be seen in the Suppl.

6 Ablation Study

6.1 Effect of Taking PS Frame as Supervision

PS frame (*i.e.*, \mathbf{S}_{ps}) is regarded as unaligned supervision in our methods. To mitigate the adverse effects of misalignment between output and target supervision, we deploy a pre-trained PWC-Net [54] to align the desmoking result $\hat{\mathbf{I}}_i$ with \mathbf{S}_{ps} , then calculate reconstruction loss \mathcal{L}_{rec} , as shown in Eq. (5). Here we conduct experiments with different loss variants to validate the effectiveness of our method. The quantitative results are shown in Tab. 2. First, we can train the desmoking network without \mathcal{L}_{rec} , *i.e.*, only with the adversarial loss. In this case, the method degenerates to the unpaired learning manner, and it leads to a significant performance drop. Second, when we do not consider the misalignment issue and utilize the supervision with naive ℓ_1 loss, it still results in a severe performance drop. Third, we compare our deformation-based loss with

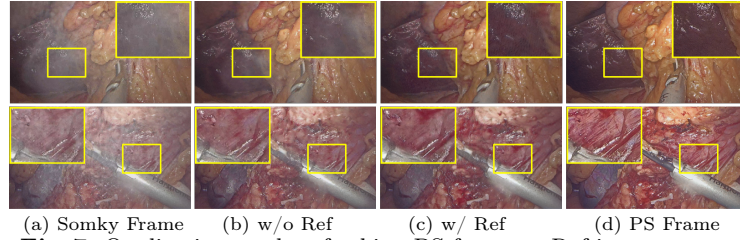


Fig. 7: Qualitative results of taking PS frame as Ref input or not.

alternative misalignment-tolerated methods, including the contextual bilateral (CoBi) loss [68] and reversed-order alignment (*i.e.*, ‘align \mathbf{S}_{ps} with $\hat{\mathbf{I}}_i$ ’). Although these two methods can also improve performance compared with naive ℓ_1 loss, our method achieves better results than theirs. Compared with the CoBi loss, our method exploits the motion prior in the pre-trained optical flow estimation network, achieving more accurate spatial matching. Compared with the reversed-order alignment strategy that partially destroys the information of PS frame, our method keeps PS frame unchanged, thus providing more accurate supervision.

6.2 Effect of Taking PS Frame as Reference Input

Note that PS frame is available during training and testing. We take it as an additional reference (Ref) input. We perform experiments to evaluate its effectiveness by removing the Ref. As shown in Tab. 3, the additional Ref enables 0.67dB PSNR improvements. Besides, Fig. 7 shows that it leads to cleaner smoke removal and finer-scale detail recovery, especially in areas with dense smoke.

We also conduct experiments to validate whether PS frame is a suitable choice to be used as Ref. For comparison, we replace Ref with the first smoky frame, as it generally has less smoke than subsequent smoky ones. As shown in Tab. 3, the replacement leads to 0.28dB PSNR drop, which illustrates the frames before the activations of high-energy devices are preferable to be taken as Ref.

6.3 Effect of Strategies to Avoid Trivial Solutions

A masking strategy and a regularization term are introduced to prevent trivial solutions. To validate the effectiveness, we conduct experiments with their different combinations, *i.e.*, ‘w/o Mask & w/o Reg’, ‘w/ Mask & w/o Reg’, ‘w/o Mask & w/ Reg’ and ‘w/ Mask & w/ Reg’. ‘Mask’ and ‘Reg’ denote the masking strategy and the regularization term respectively. Naively inputting PS frame as Ref easily leads to trivial solutions, as marked with yellow boxes in Fig. 8 (b). Fig. 8 (c) and (d) show that both ‘Mask’ and ‘Reg’ inhibit the trivial solutions, but using one of them alone does not achieve the best results. On the one hand, using ‘Mask’ alone can handle poorly aligned areas well, as they are relatively easy to detect. Nevertheless, there still remain artifacts around high-energy devices, as marked with yellow boxes in Fig. 8 (c). On the other hand, using ‘Reg’ alone generally suppresses trivial solutions. As the areas with significantly imperfect optical flow are not explicitly processed, it may leave some traces of high-energy

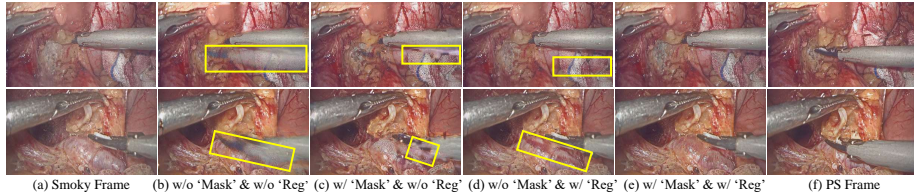


Fig. 8: Effect of strategies to avoid trivial solutions. ‘Mask’ and ‘Reg’ denote masking strategy and regularization term, respectively. Naively inputting PS frame as Ref leads to trivial solutions, as marked with yellow boxes in (b). Using ‘Mask’ alone may generate artifacts around high-energy devices, as marked with yellow boxes in (c). Using ‘Reg’ alone may leave trivial solution traces of high-energy devices from Ref, as marked with yellow boxes in (d). Their combination produces better results in (e).

Table 2: Ablation studies on reconstruction loss \mathcal{L}_{rec} .

\mathcal{L}_{rec}	PSNR \uparrow /SSIM \uparrow /FADE \downarrow /NIQE \downarrow /PI \downarrow
None	22.40/0.6030/0.4219/5.37/4.04
Naive ℓ_1 Loss	22.67/0.5880/0.4720/5.23/4.00
CoBi Loss [68]	22.82/0.6093/0.4721/4.86/3.86
Align S_{ps} with \hat{I}_i	24.12/0.6205/0.4663/5.05/3.94
Align \hat{I}_i with S_{ps}	24.23/0.6225/0.4626/4.85/3.87

Table 3: Quantitative results when inputting different Ref. ‘None’ denotes no Ref being input.

Input Ref	PSNR \uparrow /SSIM \uparrow /FADE \downarrow /NIQE \downarrow /PI \downarrow
None	23.56/0.6183/0.4772/4.87/3.85
First Smoky Frame	23.95/0.6216/0.4701/4.88/3.85
PS Frame	24.23/0.6225/0.4626/4.85/3.87

devices from Ref, as marked with yellow boxes in Fig. 8 (d). Instead, their combination generates improved results, as shown in Fig. 8 (e). Besides, we also provide the effect of regularization loss weight λ_{reg} in the Suppl.

7 Conclusion

Existing laparoscopic surgery desmoking works struggle with processing real-world smoke, especially dense smoke, due to the unavailable of real-world paired data and the severely ill-posed nature of single-image methods. To address the issue, we suggested leveraging the internal characteristics of real-surgery video for effective self-supervised video desmoking and propose SelfSVD to achieve this. Based on the observation that the pre-smoke (PS) frame has less smoke and similar contents as subsequent smoky ones, SelfSVD utilizes it as an unaligned supervision. Moreover, SelfSVD takes PS frame as a reference input to handle dense smoke better, and a masking strategy and a regularization term are introduced to prevent trivial solutions. Besides, we collect a real-world laparoscopic surgery video desmoking (LSVD) dataset, which can potentially be advantageous for future studies. Extensive experiments show that our SelfSVD outperforms the state-of-the-art methods both quantitatively and qualitatively.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62371164 and No. U22B2035.

Self-Supervised Video Desmoking for Laparoscopic Surgery (Supplementary Material)

Renlong Wu¹, Zhilu Zhang¹(✉), Shuohao Zhang¹, Longfei Gou²,
Haobin Chen², Lei Zhang³, Hao Chen²(✉), and Wangmeng Zuo¹

¹ Harbin Institute of Technology, China

² Southern Medical University, China

³ Hong Kong Polytechnic University, China

hirenlongwu@gmail.com, cszlzhang@outlook.com,
yhyzshrby@163.com, Calvin_smu@163.com, HaoBin_Chen@outlook.com,
cszlzhang@comp.polyu.edu.hk, chenhao.05@163.com, wmzuo@hit.edu.cn

The content of the supplementary material involves:

- Results on synthetic dataset in Sec. **A**.
- Networks details in Sec. **B**.
- Visualization of mask in Sec. **C**.
- Effect of regularization term in Sec. **D**.
- Examples from LSVD dataset in Sec. **E**.
- More result comparisons in Sec. **F**.
- Practical deployment in surgery in Sec. **G**.
- Limitation and social impact in Sec. **H**.

A Results on Synthetic Dataset

We also conduct experiments in a synthetic video dataset. We evaluate the compared methods with paired clean videos that are spatially aligned with the smoky videos. We collect 380 clean video clips from public Cholec80 dataset [56], where 280 clips are used for training and the remaining 100 ones are used for evaluation. We follow the surgery smoke simulation manner [25] and add synthetic smoke in the clean video clips, except the first frame (regarded as PS frame). The results are shown in Tab. **A**. As the PS frame in the synthetic dataset is clean, we do not perform experiments with SelfSVD* and SelfSVD*-S. Our results get the best PSNR scores, indicating the effectiveness of the proposed method. Moreover, the visual comparisons in Figs. **A** and **B** show that our results produce few visual artifacts, remove more clean smoke and are more consistent with the GT.

Table A: Comparison results on the synthetic dataset. The best results in each category are marked in **bold**.

	Methods	PSNR \uparrow	SSIM \uparrow
Unsupervised Image Processing	PSD [9]	13.94	0.7683
	DCP [21]	15.71	0.7413
Unpaired Image Processing	DCP-Pixel2Pixel [50]	20.01	0.5226
	DistentGAN [52]	20.88	0.8056
	Desmoke-LAP [43]	24.82	0.8966
	RefineNet [70]	26.94	0.9334
Self-Supervised Image Processing	UHD [63]	25.50	0.9410
	MSDesmoking [59]	26.77	0.9021
	Wang <i>et al.</i> [60]	30.76	0.9470
	MSBDN [11]	23.27	0.6251
	AODNet [29]	22.69	0.9002
	DADNet [20]	23.32	0.8264
Self-Supervised Video Processing	DeHamar [19]	29.06	0.9396
	BasicVSR [6]	31.33	0.9503
	BasicVSR++ [7]	31.84	0.9570
	MAPNet [64]	31.29	0.9242
	(Ours) SelfSVD-S	31.96	0.9520
	(Ours) SelfSVD	32.30	0.9611

B Network Details

We design the video desmoking network based on the unidirectional recurrent network [62]. It includes five modules, *i.e.*, feature encoder, masked-ref generator, alignment, fusion, and reconstruction module. When processing the i -th smoky frame \mathbf{S}_i , it is first fed into the encoder to obtain feature representations \mathbf{F}_i . \mathbf{S}_{ref} and \mathbf{S}_i are fed into the masked-ref generator to get the masked reference features $\tilde{\mathbf{F}}_{ref \rightarrow i}$. Then, we deploy the alignment module to align previous temporal features \mathbf{H}_{i-1} to \mathbf{F}_i , getting the warped ones $\mathbf{H}_{i-1 \rightarrow i}$. Next, the fusion module concatenates \mathbf{F}_i , $\tilde{\mathbf{F}}_{ref \rightarrow i}$ and $\mathbf{H}_{i-1 \rightarrow i}$ as inputs, producing the fused features. Finally, the fused features are fed into the reconstruction module to generate the restored clean component $\hat{\mathbf{I}}_i$.

Details of SelfSVD and SelfSVD*. For SelfSVD and SelfSVD*, the encoder includes two 3×3 convolutional layers with a stride of 2 for scale down-sampling and 5 residual blocks [22] for feature extraction. We deploy individual encoders for the smoky frame and reference input respectively. The alignment module is built upon a pre-trained optical flow network (*e.g.*, PWC-Net [54]). The fusion module consists of a 3×3 convolutional layer for channel reduction and 60 residual blocks for feature enhancement. The reconstruction module includes 5 residual blocks, two pixel-shuffle operations, and a final 3×3 convolutional layer.

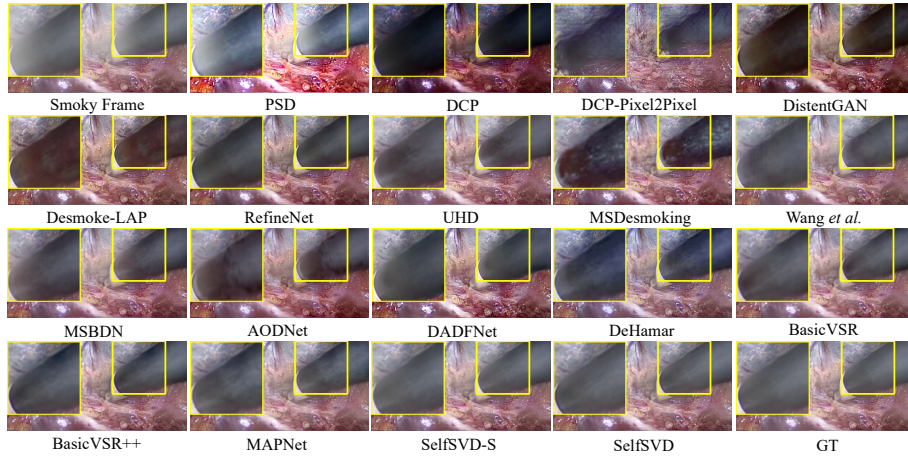


Fig. A: Qualitative comparisons on the synthetic dataset. Our results produce few visual artifacts and are more consistent with the GT.

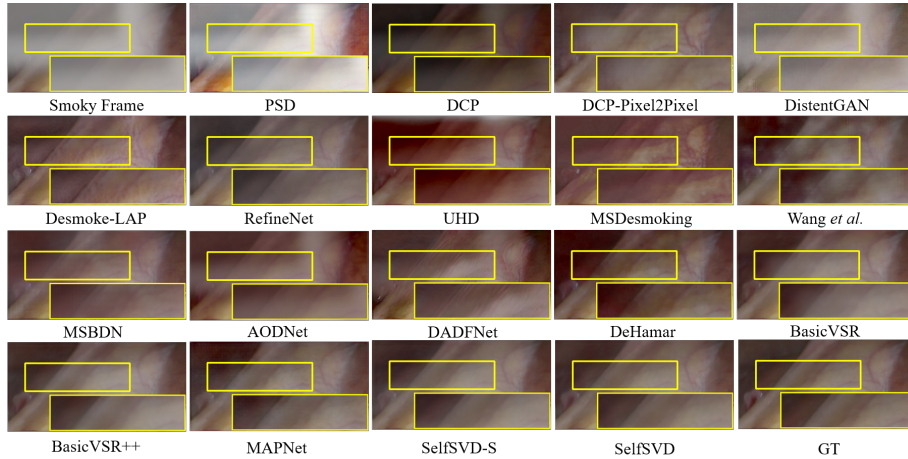


Fig. B: Qualitative comparisons on the synthetic dataset. Our results remove more clean smoke and are more consistent with the GT.

Details of SelfSVD-S and SelfSVD*-S. The computation costs of SelfSVD and SelfSVD* are generally similar to the video processing methods. To make the computation cost consistent with single-image ones, we present the lightweight models SelfSVD-S and SelfSVD*-S. Specifically, we first replace PWC-Net [54] in the alignment module to a more lightweight optical flow network SpyNet [46]. Second, we reduce the number of residual blocks in the feature encoder, masked-ref generator, fusion, and reconstruction module from 5, 5, 60, 5 to 3, 3, 8, 3 respectively. Third, we reduce the channel numbers from 64 to 32. Benefiting

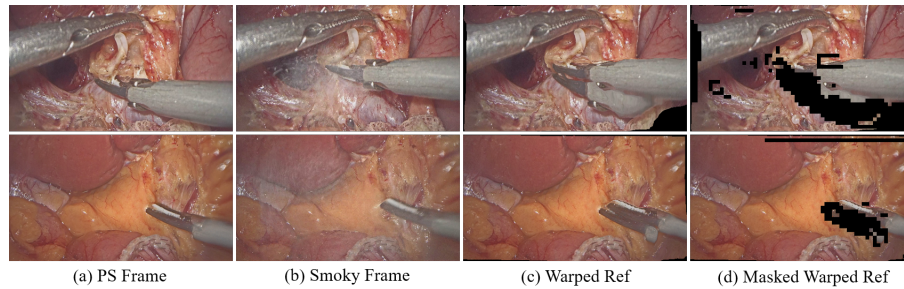


Fig. C: Mask examples. A mask is generated to indicate areas with imperfect optical flow to prevent trivial solutions. Masked regions are marked with black pixels in (d).

Table B: Structure configuration of the discriminator. The kernel size of all convolutional layers is 4×4 . ‘Stride’ denotes the stride of convolutional layer. ‘BN’ denotes the BatchNorm [27] operation.

Layer	Filter	Stride	Output size
Conv, LeakyReLU	$3 \rightarrow 64$	2	128×128
Conv, BN, LeakyReLU	$64 \rightarrow 128$	2	64×64
Conv, BN, LeakyReLU	$128 \rightarrow 256$	2	32×32
Conv, BN, LeakyReLU	$256 \rightarrow 512$	2	31×31
Conv, BN, LeakyReLU	$512 \rightarrow 1$	1	30×30

from the above simplification, SelfSVD-S and SelfSVD*-S significantly reduce the number of model parameters and computation costs, while keeping performance.

Details of Discriminator. PatchGAN [74] is employed as the discriminator to distinguish whether a patch is real or fake. Its structure is shown in Tab. B.

C Visualization of Mask

To prevent trivial solutions, we generate a mask to indicate areas with imperfect optical flow. Moreover, we process the smoky frame and warped reference (Ref) input with dark channel prior [21] (DCP) and large-kernel Gaussian blur (Blur) to mitigate smoke interference. Mask examples are provided in Figs. C and D. On the one hand, it successfully detects the areas with imperfect optical flow, as shown in Fig. C. On the other hand, DCP and Blur help to avoid detecting incorrect masked areas, as shown in Fig. D.

D Effect of Regularization Term

Here we conduct experiments to validate the effect of regularization term weight λ_{reg} . The results are shown in Fig. E and Tab. C. In general, a smaller λ_{reg} leads

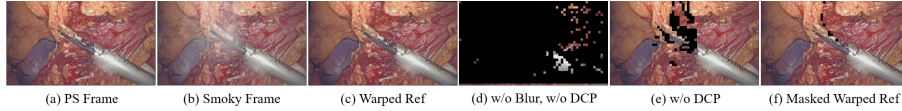


Fig. D: Effect of dark channel prior (DCP) and Gaussian blur (Blur) operations. DCP and Blur help to avoid detecting incorrect masked areas. The masked areas are marked with black pixels in (d), (e), and (f).

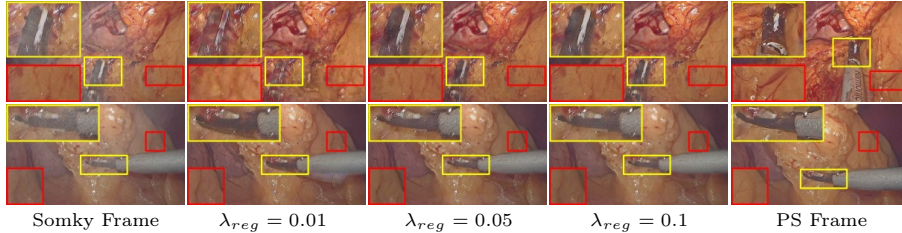


Fig. E: Effect of regularization term weight λ_{reg} . Please zoom in for better observation.

Table C: Effect of regularization term weight λ_{reg} .

λ_{reg}	PSNR \uparrow / SSIM \uparrow / FADE \downarrow / NIQE \downarrow / PI \downarrow
0.01	24.46 / 0.6254 / 0.4650 / 4.91 / 3.88
0.05	24.23 / 0.6225 / 0.4626 / 4.85 / 3.87
0.1	24.03 / 0.6200 / 0.4658 / 4.87 / 3.82

to a weaker suppression of trivial solutions, while a larger one leads to a weaker utilization of Ref. We set λ_{reg} to 0.05 for the trade-off.

E Examples from LSVD Dataset

We construct the laparoscopic surgery video desmoking (LSVD) dataset from professional hospitals. Some examples from the dataset are provided in Fig. F. It can be seen that the dataset contains diverse and complex surgery smoke, such as mist, droplets, and streaks.

F More Result Comparisons

In the main text, we utilize the processed PS frame as the target to evaluate desmoking methods, as smoke may remain in it. Here, we provide the comparison results when taking the original PS frame as the target, as shown in Tab. D. Due to interference from smoke in the PS frame, SelfSVD* and SelfSVD*-S get lower PSNR and SSIM values than SelfSVD and SelfSVD-S, respectively. Moreover, it shows that the proposed methods still outperform state-of-the-art ones.

Table D: Comparisons on LSVD dataset. The original PS frame is taken as the clean target for calculating metrics. The best results in each category are marked in **bold**.

	Methods	PSNR \uparrow	SSIM \uparrow
Unsupervised Image Processing	PSD [9]	14.49	0.3811
	DCP [21]	17.54	0.5819
Unpaired Image Processing	DCP-Pixel2Pixel [50]	20.31	0.5409
	DistGAN [52]	22.58	0.6429
	Desmoke-LAP [43]	23.70	0.6548
	RefineNet [70]	24.06	0.6559
Self-Supervised Image Processing	UHD [63]	21.78	0.6198
	MSDesmoking [59]	22.42	0.6415
	Wang <i>et al.</i> [60]	23.63	0.6496
	MSBDN [11]	23.69	0.6515
	AODNet [29]	23.45	0.6555
	DADFNet [20]	23.73	0.6516
	DeHamar [19]	24.12	0.6595
Self-Supervised Video Processing	BasicVSR [6]	24.06	0.6560
	BasicVSR++ [7]	24.38	0.6592
	MAPNet [64]	24.25	0.6567
	(Ours) SelfSVD-S	24.84	0.6551
	(Ours) SelfSVD*-S	24.25	0.6556
	(Ours) SelfSVD	25.08	0.6564
	(Ours) SelfSVD*	24.62	0.6548

More qualitative comparison results are provided in Figs. G to J. Our methods can remove more smoke, recover more photo-realistic details, and produce results more consistent with PS frames than state-of-the-art ones.

G Practical Deployment in Surgery

In the main text, we have introduced how SelfSVD processes a single smoky video clip. As a practical surgery video contains multiple smoky clips, here we illustrate how to handle it. When processing the i -th smoky frame \mathbf{S}_i , an additional reference (Ref, \mathbf{S}_{ref}) should be fed into the model to help smoke removal. There are two possible ways to select it. One is to always adopt the starting frame, and another one is to select it dynamically as the surgery proceeds. The latter is more reasonable in surgical scenarios, as the changing video contents lead to starting frames providing insufficient information for long-distance ones.

Specifically, we prefer to select \mathbf{S}_{ref} from previous neighboring frames that are clearer than \mathbf{S}_i . As shown in Algorithm A, we first deploy a Ref detector according to the residual between the current smoky input and the desmoking output. \mathbf{S}_{ref} is updated when next clearer frame occurs. Then we feed \mathbf{S}_{ref} and

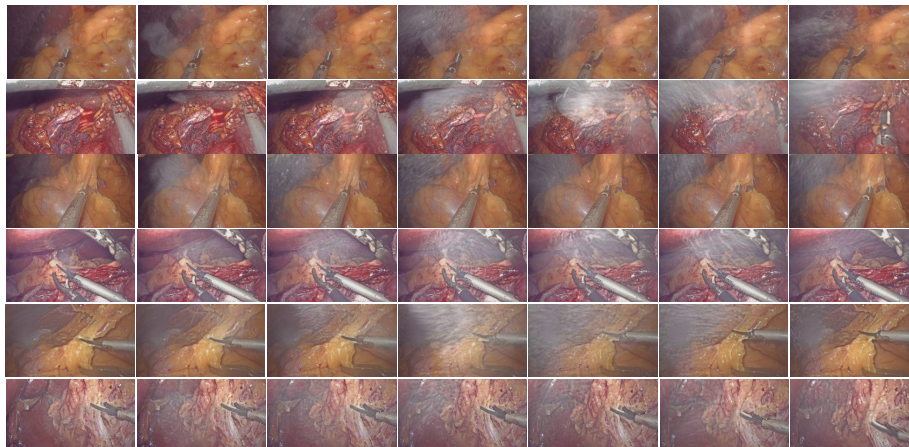


Fig. F: Examples from LSVD dataset. Each row shows several frames from a video clip. It can be seen that the dataset contains diverse and complex surgery smoke, such as mist, droplets, and streaks. Please zoom in for better observation.

Algorithm A Pseudo code about the practical deployment of SelfSVD in surgery

Require: $\{\mathbf{S}_i\}_{i=1}^N$: N surgery video frames,
1: **for** i from 1 to N with stride L **do**
2: **if** $\|\mathbf{S}_i - \text{SelfSVD}(\mathbf{S}_i, \mathbf{S}_i)\|_1 < \epsilon$ **then** ▷ utilize \mathbf{S}_i as Ref for itself
3: $\mathbf{S}_{ref} = \mathbf{S}_i$ ▷ detect the additional reference input \mathbf{S}_{ref}
4: **end if**
5: $\{\hat{\mathbf{I}}_k\}_{k=i}^{i+L} = \text{SelfSVD}(\{\mathbf{S}_k\}_{k=i}^{i+L}, \mathbf{S}_{ref})$ ▷ remove smoke for $\{\mathbf{S}_k\}_{k=i}^{i+L}$
6: **return** $\{\hat{\mathbf{I}}_k\}_{k=i}^{i+L}$.
7: **end for**

the smoky video clip $\{\mathbf{S}_k\}_{k=i}^{i+L}$ into a SelfSVD model, generating the clean results $\{\hat{\mathbf{I}}_k\}_{k=i}^{i+L}$. L is the frame number of the current video clip and we set it to 5. Please see some visualization examples at the <https://github.com/ZcsrenlongZ/SelfSVD>.

H Limitation and Social Impact

This work is still limited in processing complex surgery smoke droplets. The droplets influence the accuracy of the alignment module, making it hard to effectively utilize the complementary information from input frames. It leads to some droplet traces in the desmoking results.

As for the social impact, this work is promising to be applied to laparoscopic surgery for observing the surgical sites more clearly. It has no foreseeable negative influence. Besides, the images utilized in this work are from professional hospitals and have been authorized to be public. There is no personally identifiable information about patients or offensive content in the experimental data.

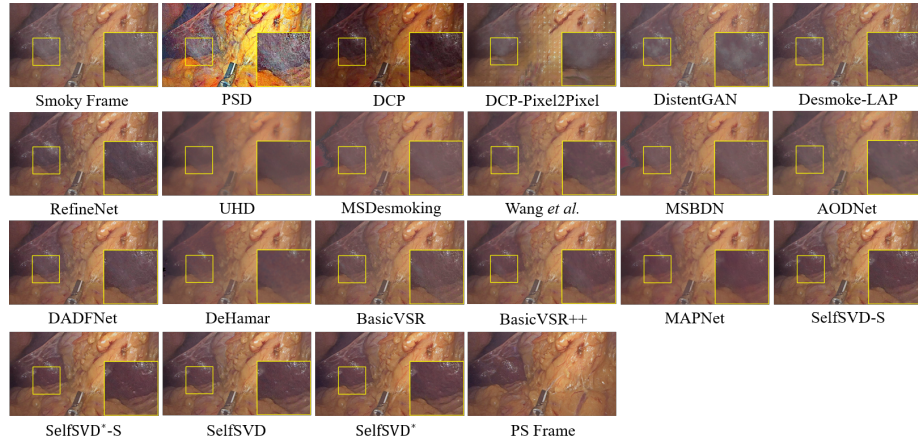


Fig. G: Qualitative comparisons on LSVD dataset. Our methods generate results more consistent with the PS frame. Please zoom in for better observation.

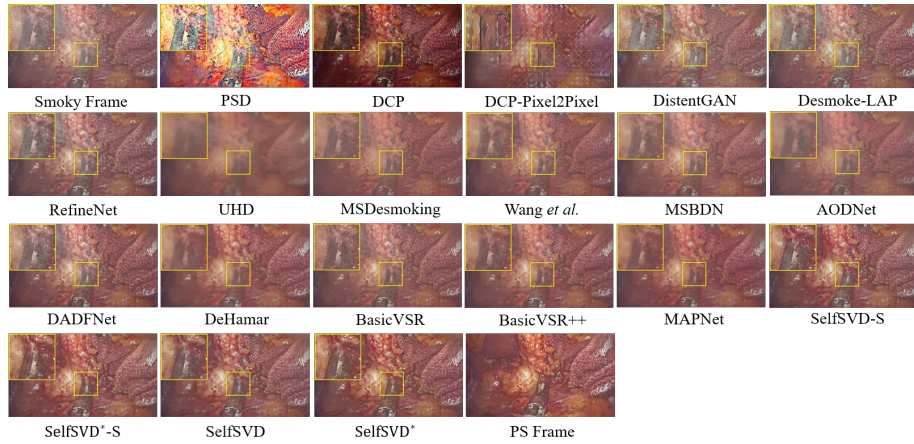


Fig. H: Qualitative comparisons on LSVD dataset. Our methods remove more clean smoke and recover more realistic details. Please zoom in for better observation.

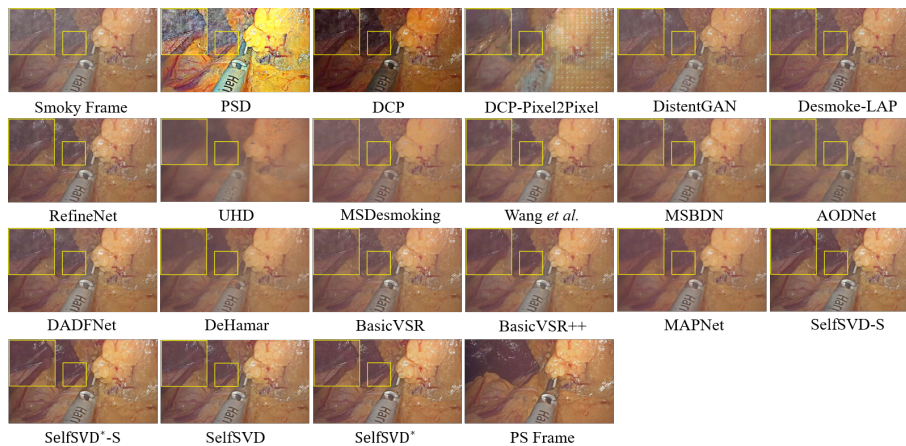


Fig. I: Qualitative comparisons on LSVD dataset. Our methods generate results with fewer artifacts and are more consistent with the PS frame. Please zoom in for better observation.

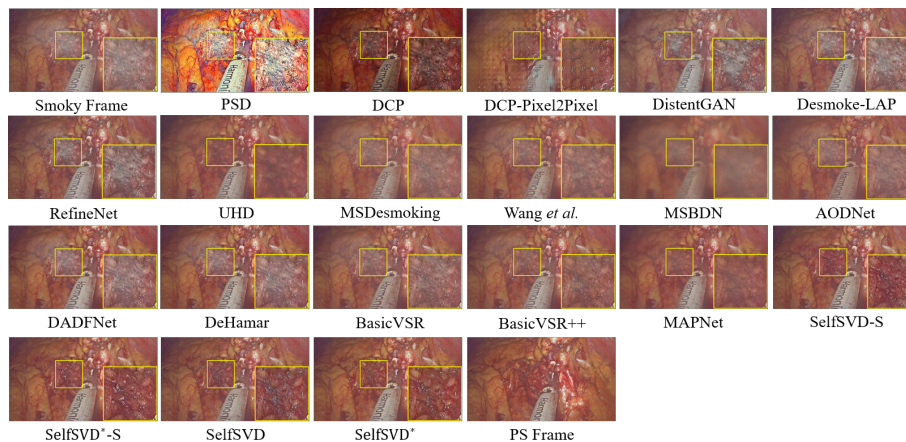


Fig. J: Qualitative comparisons on LSVD dataset. Our methods remove more clean smoke and recover more fine-scale details. Please zoom in for better observation.

References

1. Azam, M.A., Khan, K.B., Rehman, E., Khan, S.U.: Smoke removal and image enhancement of laparoscopic images by an artificial multi-exposure image fusion method. *Soft Computing* **26**(16), 8003–8015 (2022) [2](#)
2. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Deep burst super-resolution. In: *CVPR* (2021) [10](#)
3. Bishop, C.M., Nasrabadi, N.M.: *Pattern recognition and machine learning*, vol. 4. Springer (2006) [8](#)
4. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. pp. 0–0 (2018) [10](#)
5. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. *TIP* (2016) [2, 4](#)
6. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: *CVPR* (2021) [11, 16, 20](#)
7. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: *CVPR* (2022) [10, 11, 12, 16, 20](#)
8. Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J.: De-smokegcn: generative cooperative networks for joint surgical smoke detection and removal. *T-MI* (2019) [2, 4](#)
9. Chen, Z., Wang, Y., Yang, Y., Liu, D.: Psd: Principled synthetic-to-real dehazing guided by physical priors. In: *CVPR* (2021) [2, 4, 11, 12, 16, 20](#)
10. Choi, L.K., You, J., Bovik, A.C.: Referenceless prediction of perceptual fog density and perceptual image defogging. *TIP* (2015) [10](#)
11. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: Multi-scale boosted dehazing network with dense feature fusion. In: *CVPR* (2020) [4, 11, 16, 20](#)
12. Donoho, D.L.: Compressed sensing. *IEEE Transactions on information theory* **52**(4), 1289–1306 (2006) [8](#)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020) [4](#)
14. Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., Yang, M.H.: Burst image restoration and enhancement. In: *CVPR* (2022) [10](#)
15. Engin, D., Genç, A., Kemal Ekenel, H.: Cycle-dehaze: Enhanced cyclegan for single image dehazing. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. pp. 825–833 (2018) [4](#)
16. Fan, J., Guo, F., Qian, J., Li, X., Li, J., Yang, J.: Non-aligned supervision for real image dehazing. *arXiv preprint arXiv:2303.04940* (2023) [2, 4](#)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *NeurIPS* (2014) [2, 4](#)
18. Gu, L., Liu, P., Jiang, C., Luo, M., Xu, Q.: Virtual digital defogging technology improves laparoscopic imaging quality. *Surgical innovation* **22**(2), 171–176 (2015) [2](#)
19. Guo, C.L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C.: Image dehazing transformer with transmission-aware 3d position embedding. In: *CVPR* (2022) [2, 4, 11, 12, 16, 20](#)

20. Guo, Y., Liu, R.W., Nie, J., Lyu, L., Xiong, Z., Kang, J., Yu, H., Niyato, D.: Dadfnet: Dual attention and dual frequency-guided dehazing network for video-empowered intelligent transportation. arXiv preprint arXiv:2304.09588 (2023) [4](#), [11](#), [12](#), [16](#), [20](#)
21. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. TPAMI (2010) [4](#), [8](#), [11](#), [12](#), [16](#), [18](#), [20](#)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) [10](#), [16](#)
23. Holl, P., Koltun, V., Um, K., Thuerey, N.: phiflow: A differentiable pde solving framework for deep learning via physical simulations. In: NeurIPS workshop. vol. 2 (2020) [2](#), [10](#)
24. Hong, T., Huang, P., Zhai, X., Gu, C., Tian, B., Jin, B., Li, D.: Mars-gan: Multilevel-feature-learning attention-aware based generative adversarial network for removing surgical smoke. IEEE Transactions on Medical Imaging **42**(8), 2299–2312 (2023). <https://doi.org/10.1109/TMI.2023.3245298> [2](#), [4](#), [5](#)
25. Hong, T., Huang, P., Zhai, X., Gu, C., Tian, B., Jin, B., Li, D.: Mars-gan: Multilevel-feature-learning attention-aware based generative adversarial network for removing surgical smoke. IEEE Transactions on Medical Imaging (2023) [15](#)
26. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters (2008) [10](#)
27. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015) [18](#)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) [10](#)
29. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. In: ICCV (2017) [2](#), [4](#), [11](#), [12](#), [16](#), [20](#)
30. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: End-to-end united video dehazing and detection. In: AAAI (2018) [4](#)
31. Li, B., Gou, Y., Gu, S., Liu, J.Z., Zhou, J.T., Peng, X.: You only look yourself: Unsupervised and untrained single image dehazing neural network. International Journal of Computer Vision **129**, 1754–1767 (2021) [4](#)
32. Li, B., Gou, Y., Liu, J.Z., Zhu, H., Zhou, J.T., Peng, X.: Zero-shot image dehazing. IEEE Transactions on Image Processing **29**, 8457–8466 (2020) [4](#)
33. Li, J., Li, Y., Zhuo, L., Kuang, L., Yu, T.: Usid-net: Unsupervised single image dehazing network via disentangled representations. IEEE Transactions on Multimedia (2022) [4](#)
34. Li, Y., Ren, D., Shu, X., Zuo, W.: Learning single image defocus deblurring with misaligned training pairs. In: AAAI (2023) [6](#)
35. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCV (2021) [4](#)
36. Lin, J., Jiang, M., Pang, Y., Wang, H., Chen, Z., Yan, C., Liu, Q., Wang, Y.: A desmoking algorithm for endoscopic images based on improved u-net model. Concurrency and Computation: Practice and Experience **33**(22), e6320 (2021) [2](#), [3](#), [4](#)
37. Liu, Y., Wan, L., Fu, H., Qin, J., Zhu, L.: Phase-based memory network for video dehazing. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5427–5435 (2022) [4](#)
38. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) [10](#)

39. Loukas, C.: Video content analysis of surgical procedures. *Surgical endoscopy* **32**, 553–568 (2018) [1](#)
40. Ma, L., Song, H., Zhang, X., Liao, H.: A smoke removal method based on combined data and modified u-net for endoscopic images. In: *EMBC (2021)* [2](#), [3](#), [4](#), [5](#)
41. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: *ICCV (2017)* [9](#)
42. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012) [10](#)
43. Pan, Y., Bano, S., Vasconcelos, F., Park, H., Jeong, T.T., Stoyanov, D.: Desmoke-lap: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *IJCARS (2022)* [2](#), [4](#), [5](#), [9](#), [11](#), [16](#), [20](#)
44. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *NeurIPS (2019)* [10](#)
45. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., Jin, Z.: Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12802–12813 (2023) [4](#)
46. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4161–4170 (2017) [17](#)
47. Ren, W., Zhang, J., Xu, X., Ma, L., Cao, X., Meng, G., Liu, W.: Deep video dehazing with semantic segmentation. *TIP (2018)* [4](#)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI (2015)* [4](#)
49. Salazar-Colores, S., Jimenez, H.M., Ortiz-Echeverri, C.J., Flores, G.: Desmoking laparoscopy surgery images using an image-to-image translation guided by an embedded dark channel. *Access (2020)* [2](#)
50. Salazar-Colores, S., Alberto-Moreno, H., Ortiz-Echeverri, C.J., Flores, G.: Desmoking laparoscopy surgery images using an image-to-image translation guided by an embedded dark channel (2020) [2](#), [4](#), [5](#), [11](#), [16](#), [20](#)
51. Sengar, V., Seemakurthy, K., Gubbi, J., P, B.: Multi-task learning based approach for surgical video desmoking. In: *Proceedings of the twelfth Indian conference on computer vision, graphics and image processing*. pp. 1–9 (2021) [2](#), [10](#)
52. Shyam, P., Yoon, K.J., Kim, K.S.: Towards domain invariant single image dehazing. In: *AAAI (2021)* [11](#), [16](#), [20](#)
53. Su, X., Wu, Q.: Multi-stages de-smoking model based on cyclegan for surgical de-smoking. *International Journal of Machine Learning and Cybernetics* pp. 1–14 (2023) [2](#), [4](#), [5](#)
54. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: *ICCV (2018)* [3](#), [10](#), [12](#), [16](#), [17](#)
55. Tchaka, K., Pawar, V.M., Stoyanov, D.: Chromaticity based smoke removal in endoscopic images. In: *Medical Imaging 2017: Image Processing (2017)* [3](#), [5](#)
56. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. *T-MI (2016)* [15](#)
57. Venkatesh, V., Sharma, N., Srivastava, V., Singh, M.: Unsupervised smoke to desmoked laparoscopic surgery images using contrast driven cyclic-desmokegan. *Comput. Biol. Med (2020)* [2](#), [4](#), [5](#), [9](#)
58. Wang, C., Alaya Cheikh, F., Kaaniche, M., Beghdadi, A., Elle, O.J.: Variational based smoke removal in laparoscopic images. *BEO (2018)* [3](#), [5](#)

59. Wang, C., Mohammed, A.K., Cheikh, F.A., Beghdadi, A., Elle, O.J.: Multiscale deep desmoking for laparoscopic surgery. In: *Medical Imaging 2019: Image Processing*. vol. 10949, pp. 505–513. SPIE (2019) [2](#), [3](#), [4](#), [5](#), [11](#), [16](#), [20](#)
60. Wang, F., Sun, X., Li, J.: Surgical smoke removal via residual swin transformer network. *IJCARS* (2023) [2](#), [3](#), [4](#), [5](#), [11](#), [16](#), [20](#)
61. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *TIP* (2004) [8](#), [10](#)
62. Wu, R., Zhang, Z., Zhang, S., Zhang, H., Zuo, W.: Rbsr: Efficient and flexible recurrent network for burst super-resolution. In: *PRCV* (2023) [9](#), [10](#), [16](#)
63. Xiao, B., Zheng, Z., Chen, X., Lv, C., Zhuang, Y., Wang, T.: Single uhd image dehazing via interpretable pyramid network (2022) [4](#), [11](#), [16](#), [20](#)
64. Xu, J., Hu, X., Zhu, L., Dou, Q., Dai, J., Qiao, Y., Heng, P.A.: Video dehazing via a multi-range temporal alignment network with physical prior. In: *CVPR* (2023) [4](#), [11](#), [12](#), [16](#), [20](#)
65. Yang, X., Xu, Z., Luo, J.: Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In: *AAAI* (2018) [4](#)
66. Yang, Y., Wang, C., Liu, R., Zhang, L., Guo, X., Tao, D.: 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 (2022) [4](#)
67. Zhang, X., Dong, H., Pan, J., Zhu, C., Tai, Y., Wang, C., Li, J., Huang, F., Wang, F.: Learning to restore hazy video: A new real-world dataset and a new method. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9239–9248 (2021) [4](#), [6](#)
68. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3762–3770 (2019) [13](#), [14](#)
69. Zhang, Z., Wang, R., Zhang, H., Chen, Y., Zuo, W.: Self-supervised learning for real-world super-resolution from dual zoomed observations. In: *ECCV* (2022) [6](#)
70. Zhao, S., Zhang, L., Shen, Y., Zhou, Y.: Refinednet: A weakly supervised refinement framework for single image dehazing. *TIP* (2021) [4](#), [11](#), [12](#), [16](#), [20](#)
71. Zheng, Q., Yang, R., Ni, X., Yang, S., Jiang, Z., Wang, L., Chen, Z., Liu, X.: Development and validation of a deep learning-based laparoscopic system for improving video quality. *IJCARS* (2023) [2](#), [3](#), [4](#), [5](#)
72. Zheng, Y., Zhan, J., He, S., Dong, J., Du, Y.: Curricular contrastive regularization for physics-aware single image dehazing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5785–5794 (2023) [4](#)
73. Zhou, Y., Hu, Z., Xuan, Z., Wang, Y., Hu, X.: Synchronizing detection and removal of smoke in endoscopic images with cyclic consistency adversarial nets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* pp. 1–12 (2022). <https://doi.org/10.1109/TCBB.2022.3204673> [2](#), [4](#)
74. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *ICCV* (2017) [2](#), [4](#), [9](#), [18](#)
75. Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. *TIP* (2015) [3](#), [5](#)