# Rethinking Surgical Smoke: A Smoke-Type-Aware Laparoscopic Video Desmoking Method and Dataset

**Qifan Liang[1,2], Junlin Li[3], Zhen Han[1,2*], Xihao Wang[1,2], Zhongyuan Wang[1,2], Bin Mei[4]**

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, China
[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
[3]School of Cyber Science and Engineering, Wuhan University, China
[4]Zhongnan Hospital, Wuhan University, China
{liangqifan, junlin_li, hanzhen, wangxihao, neuromei20}@whu.edu.cn, wzy_hope@163.com

## Abstract

Electrocautery or lasers will inevitably generate surgical smoke, which hinders the visual guidance of laparoscopic videos for surgical procedures. The surgical smoke can be classified into different types based on its motion patterns, leading to distinctive spatio-temporal characteristics across smoky laparoscopic videos. However, existing desmoking methods fail to account for such smoke-type-specific distinctions. Therefore, we propose the first **S**moke-**T**ype-**A**ware Laparoscopic Video Desmoking **Net**work (STANet) by introducing two smoke types: **Diffusion Smoke** and **Ambient Smoke**. Specifically, a smoke mask segmentation sub-network is designed to jointly conduct smoke mask and smoke type predictions based on the attention-weighted mask aggregation, while a smokeless video reconstruction sub-network is proposed to perform specially desmoking on smoky features guided by two types of smoke mask. To address the entanglement challenges of two smoke types, we further embed a coarse-to-fine disentanglement module into the mask segmentation sub-network, which yields more accurate disentangled masks through the smoke-type-aware cross attention between non-entangled and entangled regions. In addition, we also construct the first large-scale synthetic video desmoking dataset with smoke type annotations. Extensive experiments demonstrate that our method not only outperforms state-of-the-art approaches in quality evaluations, but also exhibits superior generalization across multiple downstream surgical tasks.

**Datasets** — https://simon-leong.github.io/STSVD/
**Extended version** — https://arxiv.org/abs/2512.02780

## Introduction

Laparoscopic videos can provide real-time visual feedback from surgical procedures, serving as indispensable guidance for high-precision surgery. However, high-energy surgical tools such as electrocautery or lasers will inevitably generate surgical smoke due to the cauterization of tissue components, which severely degrades the visibility of laparoscopic video. This degradation can obscure anatomical structures and hinder clinical decision-making, highlighting the importance of laparoscopic video desmoking methods.

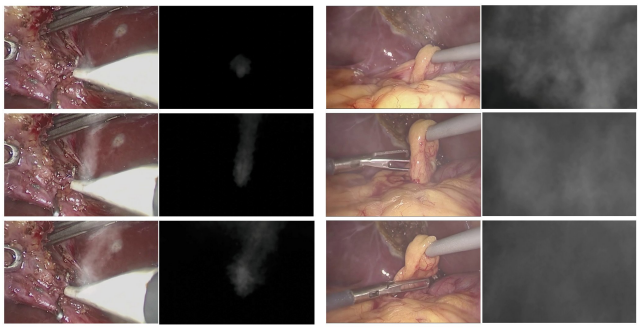


Figure 1: Visual comparison of two smoke types in surgical videos. From top to bottom, the rows represent smoky frames at successive time steps and their corresponding smoke masks.

The most relevant traditional technology for laparoscopic desmoking is unsupervised dehazing (Fan et al. 2024, 2025) and supervised dehazing (Xu et al. 2023; Song et al. 2023; Fu et al. 2025; Fang et al. 2025), which have demonstrated generalization capabilities in various scenarios. However, due to the lack of consideration for smoke spatio-temporal characteristics, dehazing methods fail to achieve satisfactory results when dealing with varying smoky motion patterns.

In recent years, some researchers have focused on laparoscopic image desmoking and proposed both unsupervised methods (Salazar-Colores et al. 2020; Pan et al. 2022) and supervised methods (Sidorov et al. 2020; Zhang et al. 2023; Liu et al. 2024). More recently, a laparoscopic video desmoking method (Wu et al. 2024) has been proposed, which introduces an unalignment masking strategy along with a temporal coherence regularization term to further enhance the smoke removal performance in video. However, smoke will regenerate turbulence and form a new motion pattern when it collides with the surgical cavity (Kim et al. 2008). Therefore, as shown in the Fig.1, surgical smoke can be classified into two types based on its motion patterns: **Diffusion smoke** and **Ambient Smoke**. The former is local

*Corresponding author.

and directional, appearing in the early stage (pre-collision) of surgical cauterizations, while the latter is global and directionless, appearing in the later stage (post-collision). These differences lead to distinctive spatio-temporal characteristics across smoky laparoscopic videos. However, existing desmoking methods have not considered the influence of smoke type, failing to achieve type-specific smoke removal.

To tackle this challenge, we introduce two smoke types into the desmoking method for the first time and propose the **S**moke-**T**ype-**A**ware Laparoscopic Video Desmoking **Net**work (STANet) that consists of three sub-networks. Specifically, the smoky feature perception sub-network first extracts and refines smoky video features through the lightweight non-rigid trajectory attention. And then, the smoke mask segmentation sub-network jointly conducts smoke mask and smoke type predictions at the local patch level, where local predictions are subsequently aggregated into global masks of two smoke types via the attention-weighted mask aggregation mechanism. On this basis, the smokeless video reconstruction sub-network introduces adaptive deformable convolution and dilated convolution into two branches respectively for smoke removal guided by two types of smoke masks. Moreover, due to multiple cauterizations in surgical procedures, the Ambient Smoke from the later stage of previous cauterization tends to be entangled with the Diffusion Smoke from the early stage of subsequent cauterization. To address the entanglement challenges of two smoke types, we further embed a coarse-to-fine mask disentanglement module into the smoke mask segmentation sub-network, which yields more accurate disentangled masks through the smoke-type-aware cross attention between non-entangled and entangled regions.

In addition, due to the lack of smoke type annotations in existing desmoking datasets, we also construct the first large-scale synthetic video desmoking dataset with smoke type labels. In this dataset, 120 smoky videos of 100 frames cover Diffusion Smoke, Ambient Smoke, and their entangled scenarios from 28 types of surgeries, along with the corresponding smoke-free videos and ground truth smoke masks. In summary, the contributions of this paper can be summarized as follows:

- We first classify surgical smoke into two types: Diffusion Smoke and Ambient Smoke, and propose the first smoke-type-aware laparoscopic video desmoking method to perform specially smoke removal guided by two smoke types.
- To address the entanglement challenges of two smoke types, we further propose a coarse-to-fine mask disentanglement module to yield more accurate disentangled masks.
- We construct the first large-scale synthetic video desmoking dataset with smoke type labels, including 120 paired clean-smoky videos (100 frames each) across 28 surgery types, along with the corresponding ground-truth smoke masks.
- Extensive experiments on synthetic and real datasets demonstrate that our method surpasses state-of-the-art dehazing and desmoking methods in quality evaluations, while exhibiting strong generalization across diverse downstream surgical tasks.

## Related Work

### Unsupervised/Supervised Dehazing

Dehazing is closely related to surgical desmoking, as both aim to restore clear visuals from degraded scenes. Existing dehazing methods based on deep learning can be broadly categorized into two groups: unsupervised and supervised.

**Unsupervised dehazing methods** eliminate the need for paired clean-hazy data by leveraging non-aligned constraints to learn effective dehazing mappings. Notably, DVD (Fan et al. 2024) introduces a non-aligned regularization strategy based on flow-guided attention, and its follow-up (Fan et al. 2025) proposes a depth-centric framework jointly modeling dehazing and depth estimation, significantly enhancing results under the real-world misalignment scenario. Despite their annotation-free advantage, unsupervised dehazing methods often suffer from unstable training and performance limitations, motivating the development of supervised alternatives.

**Supervised dehazing methods** rely on paired datasets to directly optimize dehazing model under explicit guidance, and achieve higher restoration accuracy by leveraging task-specific losses and deep network architectures. Among these, ASM (Atmospheric Scattering Model)-driven methods leverage physical scattering principles to guide supervised learning and bridge the synthetic-to-real domain gap, as exemplified by domain adaptation techniques embedding physical priors (Li et al. 2017, 2018; Guo et al. 2022), while ASM-free methods bypass explicit physical assumptions and instead rely on powerful model architectures to learn efficient dehazing mapping from paired data (Song et al. 2023; Yuan et al. 2023; Xu et al. 2023; Wu et al. 2023b). More recently, IPC-Dehaze (Fu et al. 2025) employs iterative coding-decoding architecture for progressive refinement, and SGDN (Fang et al. 2025) uses the superior structural properties of YCbCr features to guide RGB features for real-world dehazing. However, these state-of-the-art dehazing methods often overlook the diverse spatio-temporal characteristics of surgical smoke, which limits their performance of smoke removal under the surgical scenarios.

### Unsupervised/Supervised Desmoking

Deep learning has driven advances in unsupervised and supervised laparoscopic desmoking. **Unsupervised desmoking methods** typically employ unpaired domain translation frameworks such as CycleGAN (Salazar-Colores et al. 2020; Pan et al. 2022) to learn desmoking mappings without aligned training data. However, these methods often suffer from mode collapse and loss of details. Therefore, **Supervised desmoking methods** introduce smoke rendering engines (Chen et al. 2019; Holl et al. 2020; Sidorov et al. 2020; Zhang et al. 2023; Ma et al. 2025) to generate synthetic paired desmoking datasets, enabling supervised learning with more stable convergence and higher restoration fidelity. Recently, transformer-based methods like AALIDNet

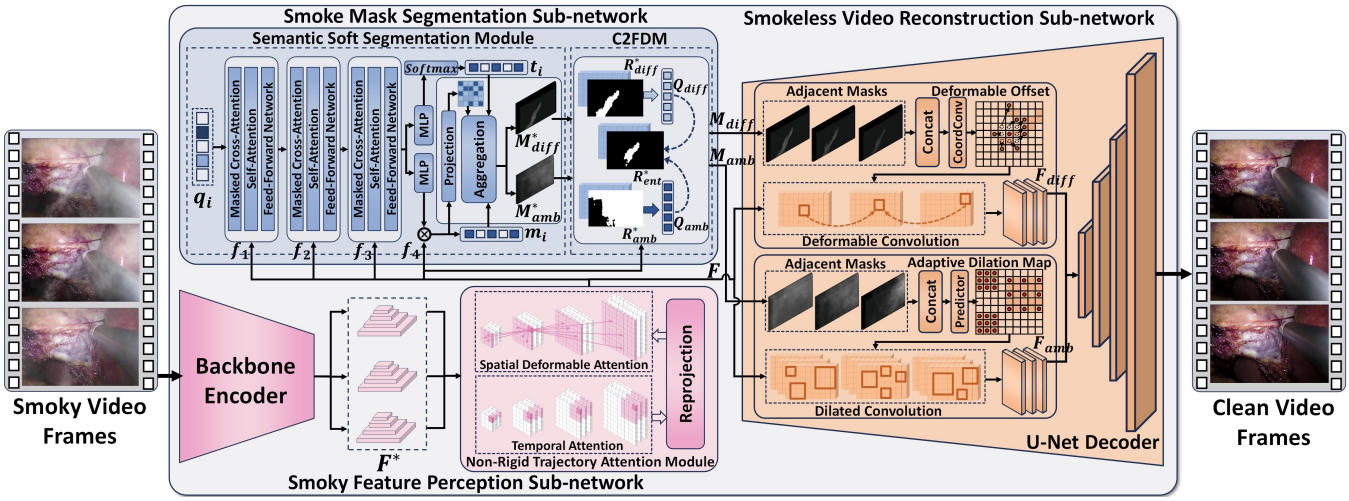


Figure 2: The overall framework of our STANet. The pink, blue and orange regions respectively denote the Smoky Feature Perception Sub-network, the Smoke Mask Segmentation Sub-network and the Smokeless Video Reconstruction Sub-network.

(Liu et al. 2024) introduce a two-stage framework that integrates smoke mask estimation with guided embedding for smokeless feature reconstruction. SelfSVD (Wu et al. 2024) introduces a video desmoking framework that leverages pre-smoke frames as pseudo ground truth to enable real-world supervision learning, and enhances desmoking performance through an unaligned masking strategy and a temporal coherence regularization term. However, existing desmoking methods still oversimplify smoke as a single type and neglect the diversity of its motion patterns, which ultimately results in unsatisfactory desmoking performance.

## Methodology

As illustrated in Fig.2, we propose the **S**moke-**T**ype-**A**ware Laparoscopic Video Desmoking **Net**work (STANet) for end-to-end laparoscopic video desmoking, which consists of three sub-networks. In the following sections, we provide a detailed description of each sub-network in sequence.

### Smoky Feature Perception Sub-network

As shown in the pink region of Fig.2, a sequence of smoky video frames is first fed into an interchangeable ResNet-18 Backbone Encoder to extract multi-scale spatial features $F^*$. Considering the non-rigid nature of surgical smoke, we draw inspiration from the reversed temporal-spatial formulation in SODA (Liu et al. 2023) and introduce a lightweight non-rigid trajectory attention module to capture the deformable characteristics of $F^*$. Specifically, while preserving the core attention and reprojection mechanism of SODA, we introduce shared projection layers, a window-based attention scheme, and reductions in both the number of attention heads and the dimensionality of vectors to compress the computational load of temporal attention and spatial deformable attention.

### Smoke Mask Segmentation Sub-network

**Semantic Soft Segmentation Module** To enable soft segmentation of different types of surgical smoke, we introduce a Semantic Soft Segmentation Module ($S^3M$) within the smoke mask segmentation sub-network, as illustrated in the blue region of Fig.2.

Inspired by recent work (Cheng et al. 2022), we formulate the soft segmentation of different types of smoke as a set prediction paradigm, in which $N$ learnable queries $q_i (i = 1 \cdots N)$ are iteratively refined to act as $N$ localized smoke-type-aware experts for distinct regions of interest, and each query $q_i$ predicts a corresponding local smoke mask $m_i$ and smoke type $t_i$. Specifically, $q_i$ is first input into three cascaded segmentation blocks, each of which consists of masked cross-attention (Cheng et al. 2022), self-attention and a feedforward network. In three blocks, $q_i$ sequentially interacts with the first three smaller scales of spatio-temporal features $F = \{f_l\}_{l=1}^{4}$ outputted by the non-rigid trajectory attention module, and enables the iterative incorporation of multi-scale smoky features into the smoke queries representation. After passing through all three blocks, each smoke query not only interacts with the fourth larger-scale feature $f_4$ via an MLP to generate the smoke mask $m_i$, but is also transformed by another MLP and a softmax operation into the smoke type $t_i$.

To enable the aggregation of all local masks $m_i$ into two smoke-type-specific global masks $M^*_{diff}$ and $M^*_{amb}$, each $m_i$ is further passed through a CNN-based attention projection to generate a local-global attention weight $w_i$. And then, under the guidance of smoke type $t_i$, the smoke-type-specific global smoke mask $M^*_{tpy}$ is computed through the attention-weighted local-global mask aggregation mechanism, as defined by the following equation:

$$M^*_{typ} = \sum_i \frac{w_i}{\sum_i w_i} \cdot m_i, \quad \{i \mid t_i = typ\} \tag{1}$$

where the value of $typ$ is $diff$ or $amb$, representing Dif-

fusion Smoke or Ambient Smoke respectively.

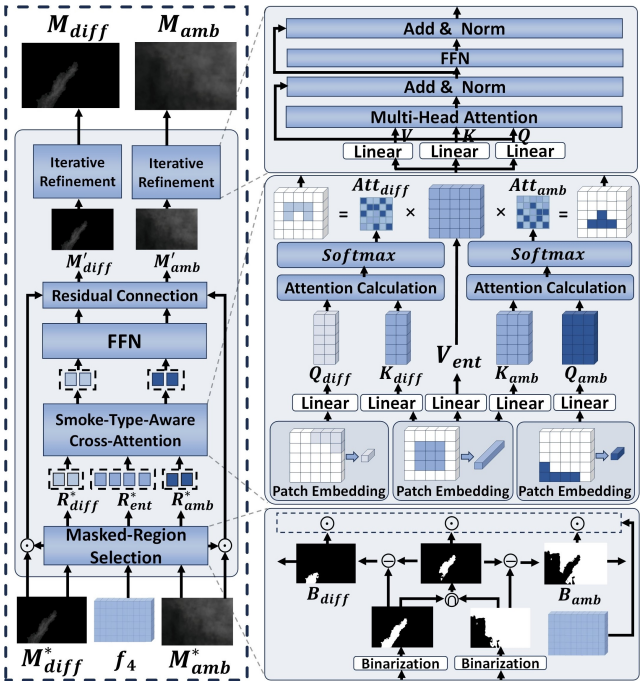


Figure 3: The illustration of the Coarse-to-Fine Disentanglement Module(C2FDM).

**Coarse-to-Fine Disentanglement Module** Multiple cauterizations during surgery will lead to entanglement of Diffusion Smoke and Ambient Smoke generated from different cauterizations. The S$^3$M does not consider this challenge, so we further embed a Coarse-to-Fine Disentanglement Module (C2FDM) after S$^3$M to refine $M^*_{typ}$ and separate the entangled regions of two smoke types.

Specifically, as shown in Fig.3, two coarse smoke masks without disentanglement $M^*_{diff}$ and $M^*_{amb}$ produced by S$^3$M, along with the fourth-scale feature $f_4$, are first fed into the masked-region selection sub-module of C2FDM. In this sub-module, three mutually exclusive region masks are generated from $M^*_{diff}$ and $M^*_{amb}$ via binarization and set operations, and then multiplied by $f_4$ to obtain coarse region features $R^*_{diff}$, $R^*_{amb}$ and $R^*_{ent}$ for Diffused Smoke, Ambient Smoke and entangled regions, respectively. In the following smoke-type-aware cross attention sub-module, three region features are first passed through a patch embedding layer and a linear projection to produce smoke-type-specific queries $Q_{diff}$ and $Q_{amb}$, as well as smoke-type-specific keys $K_{diff}$ and $K_{amb}$, and a shared entangled value $V_{ent}$. On this basis, the feature of entangled region is disentangled into two smoke types through two cross attentions between non-entangled and entangled regions. Subsequently, these two smoke-type-specific features are reconstructed into two masks of entangled regions through a feedforward network $FFN$, which are fused with two masks of non-entangled regions in $M^*_{diff}$ and $M^*_{amb}$ via residual connections to produce two smoke-type-specific masks of all regions $M'_{diff}$ and $M'_{amb}$.The above computation process is as follows:

$$M'_{typ} = FFN\left(\text{Softmax}\left(\frac{Q_{typ}K^{\top}_{typ}}{\sqrt{d}}\right)V_{ent}\right) + M^*_{typ}\cdot B_{typ}, \tag{2}$$

where $B_{typ}$ represents the smoke-type-specific binary mask corresponding to non-entangled region, which is generated from the masked-region selection sub-module and used to extract the non-entangled region in $M^*_{typ}$.

Subsequently, the coarse masks $M'_{diff}$ and $M'_{amb}$ are further passed into an iterative refinement block for optimization. In detail, the refinement block consists of a multi-head attention layer with four self-attention heads, along with normalization layers and feedforward networks, which aim to progressively refine the smoke features and ultimately produce fine masks $M_{diff}$ and $M_{amb}$.

## Smokeless Video Reconstruction Sub-network

Considering the locality and directionality of Diffusion Smoke, as well as the globality and non-directionality of Ambient Smoke, we employ a dual-branch desmoking process guided by smoke-type-specific masks within the smokeless video reconstruction sub-network.

For Diffusion Smoke, as illustrated in the upper branch of the orange region in Fig.2, we adopt deformable convolution to perform feature desmoking along the smoke diffusion path. To leverage the temporal coherence of smoke, we first concatenate Diffusion Smoke masks from adjacent frames to form a temporal composite mask $\bar{M}_{diff}$. $\bar{M}_{diff}$ is subsequently processed by a lightweight CoordConv layer (Liu et al. 2018) (with 2 added coordinate channels) and a channel-reduction attention $\mathcal{A}_{attn}$ to produce an 18-channel offset field that adaptively emphasizes informative spatial displacements:

$$\Delta_{offset} = \mathcal{A}_{attn}\left(CoordConv\left(\bar{M}_{diff}\right)\right), \tag{3}$$

Subsequently, a 3×3 deformable convolution layer is applied using the offset $\Delta_{offset}$ to extract aligned features along the smoke diffusion trajectories.

$$F_{diff} = DeformConv\left(F, \Delta_{offset}\right). \tag{4}$$

For Ambient Smoke, as illustrated in the lower branch of the orange region in Fig.2, we adopt an adaptive dilated convolution to accommodate the global and non-uniform smoke distribution. Specifically, we first concatenate Ambient Smoke masks from adjacent frames to obtain a temporal composite mask $\bar{M}_{amb}$, which is then fed into a CNN-based predictor to estimate $K$ adaptive dilation sampling position maps $map_k$ corresponding to $K$ dilation rates $rate_k$. Then, we apply parallel 3×3 dilated convolutions $DilatConv$ with all dilation rates and fuse their outputs:

$$F_{amb} = \sum_{k=1}^{K} DilatConv\left(F, rate_k, map_k\right). \tag{5}$$

where $K = 3$ and $rate_1, rate_2, rate_3$ are set to 1, 2, 3 respectively. Finally, the features $F_{diff}$ and $F_{amb}$ from two

branches are routed to a U-Net decoder, which reconstructs the smokeless video frame by fusing multi-scale features through progressive upsampling. In addition, it is worth noting that when only one type of smoke appears in the video, the corresponding branch will be activated exclusively to save unnecessary computational cost.

### Overall Training Loss

Following recent works (Cheng et al. 2022; Liu et al. 2024), our STANet is optimized using a multi-task loss $\mathcal{L}_{mul}$ to jointly supervise smoke mask segmentation, smoke type classification, and smokeless video reconstruction tasks.

To tackle the challenge of smoke mask details prediction, we introduce an additional Smoke High-frequency Wing Loss (SHWL) $\mathcal{L}_{shwl}$ to emphasize high-frequency information of masks via adaptive gradient modulation. Specifically, we first extract the high-frequency of normalized ground truth mask $M_{GT}$ and predicted mask $M_{typ}$ using a $3 \times 3$ high-pass filter, and then compute the absolute high-frequency error $\epsilon$ between them. Finally, we refer to the wing loss (Bédard et al. 2025) to optimize $\epsilon$, which can better focus on the overlooked small error of smoke mask.

To further adapt to varying smoke density, an exponential modulation factor is introduced to adjust $\mathcal{L}_{shwl}$:

$$\phi = 1 + \lambda_g(e^{M_{GT}} - 1) \tag{6}$$

where $\lambda_g = 2.0$ is empirically set to balance gradient sensitivity, and $\phi$ increases the penalty in dense smoke regions ($M_{GT} \rightarrow 1$) while reducing overfitting in sparse regions ($M_{GT} \rightarrow 0$). The total loss is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{mul} + \phi\mathcal{L}_{shwl}. \tag{7}$$

## Smoke-Type-Specific Desmoking Dataset

Due to the lack of smoke type annotations in existing desmoking datasets, we construct a large-scale **S**moke-**T**ype-**S**pecific **V**ideo **D**esmoking (STSVD) dataset, including 120 videos (100 frames each, 720×1080 resolution) with semantic labels for Diffusion Smoke, Ambient Smoke, and their entanglement across 28 types of surgical scenarios.

In order to ensure a realistic smoke simulation, STSVD considers five dimensions: surgical scenario, smoke type, smoke attribute, environment parameter and render setting, which are divided into 27 sub-dimensions. In addition, STSVD further illustrates the subdivisions of three major smoke attributes: density, orientation and velocity, which follow an empirically established distribution of real-world surgical scenarios (Hong et al. 2023). More details of STSVD are provided in the extended version and website.

## Experiments

### Datasets and Metrics

**Training Datasets** In comparison experiments, we use STSVD as the training dataset. In ablation experiments, we use PSv2rs (Liu et al. 2024) as an additional training dataset to analyze the impact of different training datasets, thereby demonstrating the superiority of STSVD. PSv2rs is the largest existing open-source synthetic desmoking dataset, which consists of 54420 smoky frames with 256×256 resolution. It should be noted that due to the lack of smoke type labels, PSv2rs can only be used to train the baseline model without awareness of smoke type.

**Testing Datasets** To evaluate the generalization of different desmoking methods in real-world scenarios, our testing dataset includes two real-world desmoking datasets Vivo and STSVD-R in addition to synthetic dataset STSVD. Vivo (Xia et al. 2024) is the only existing real-world paired desmoking dataset, which consists of 961 pairs of smoky frames extracted from 63 laparoscopic prostatectomies and is aligned using optical flow supervision, allowing for paired quantitative assessment. Considering that Vivo does not cover diverse smoke types, we have specifically constructed an new unpaired real-world dataset STSVD-R by resampling smoky videos from three surgical datasets: Cholec80 (Twinanda et al. 2016), M2CAI16 (Stauder et al. 2016), and Hamlyn (Giannarou et al. 2014), which contains 24 videos (100 frames each, 720×1080) categorized into diffusion, ambient and entangled smoke scenarios.

**Metrics** For synthetic dataset with ground truth, we use PSNR, SSIM, and LPIPS (Zhang et al. 2018) as the video reconstruction quality metrics. For real-world datasets without ground truth, we use four no-reference metrics: TOPIQ (Chen et al. 2024), Q-Align (Wu et al. 2023a), MANIQA (Yang et al. 2022), and MUSIQ (Ke et al. 2021), which respectively evaluate semantic consistency, cross-modal alignment, local statistics, multi-scale quality, and provide a comprehensive perceptual assessment.

### Implementation Details

Each input frame is resized to 720×1080 and augmented with random cropping and photometric distortion after normalization. The number of queries $N$ in the smoke mask segmentation sub-network is set to 100. Adam with an initial learning rate of $1 \times 10^{-4}$, weight decay 0.05, gradient clipping 0.01, and a polynomial decay schedule over 90K iterations are adopted. The batch size is set to 4, and PyTorch is used to implement our model with RTX 3090 GPUs.

### Comparison with State-of-the-arts

**Quantitative Comparisons** Tab.1 demonstrates the objective metrics of 8 dehazing methods, 6 desmoking methods, and our method on three datasets. Compared to the best-performing dehazing methods SGDN (Fang et al. 2025) and DehazeFormer (Song et al. 2023), our method respectively improves reference metrics PSNR, SSIM, LPIPS and no-reference metrics TOPIQ, Q-align, MANIQA, MUSIQ by an average of 0.7770 dB, 0.0156, 0.0120 and 0.0202, 0.0445, 0.0050, 0.3564, due to the precise perception and representation for spatio-temporal characteristics of surgical smoke. Compared to the best-performing desmoking method SelfSVD (Wu et al. 2024), our method also exhibits superior performance across all metrics due to the advantage of smoke-type-specific modeling under diverse surgical scenarios. Specifically, it improves reference metrics by an average of 2.8649 dB, 0.0291, 0.0487, and no-reference metrics by 0.0117, 0.0133, 0.0130, 1.7906, respectively.

| Settings | Methods | Venue | **STSVD**(synthetic dataset) | | | **Vivo**(paired real-world dataset) | | | **STSVD-R**(unpaired real-world dataset) | | | | **Complexity** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | TOPIQ↑ | Q-align↑ | MANIQA↑ | MUSIQ↑ | Params | FLOPS | Times |
| Unsupervised Dehazing | DVD | CVPR'24 | 26.1833 | 0.8945 | 0.0436 | 22.3465 | 0.7820 | 0.3028 | 0.2805 | 3.0083 | 0.1512 | 36.7322 | 15.37M | 101.30G | 0.73s |
| | DCL | AAAI'25 | 27.8133 | 0.9362 | 0.1132 | 21.6424 | 0.8341 | 0.1493 | 0.2985 | 3.0115 | 0.1813 | 38.7317 | 11.38M | 73.40G | 0.47s |
| Supervised Dehazing | AODNet | ICCV'17 | 26.6464 | 0.9520 | 0.0775 | 22.1000 | 0.8480 | 0.1820 | 0.3058 | 2.9501 | 0.1700 | 37.1496 | 0.02M | 1.36G | 0.32s |
| | MapNet | CVPR'23 | 26.7672 | 0.9156 | 0.1014 | 22.5223 | 0.8607 | 0.1653 | 0.3068 | 3.0669 | 0.1068 | 40.0033 | 28.75M | 261.20G | 0.49s |
| | RIDCP | CVPR'23 | 29.0814 | 0.9581 | 0.0629 | 21.4168 | 0.8363 | 0.2100 | 0.3025 | 3.0106 | 0.1812 | 40.1816 | 28.72M | 182.69G | 0.72s |
| | DeFormer | TIP'23 | 32.5400 | 0.9517 | 0.0463 | 23.2910 | 0.8648 | 0.1569 | 0.3082 | 3.1042 | 0.1874 | 39.5138 | 28.98M | 11.16G | 0.07s |
| | IPCDehaze | CVPR'25 | 29.8004 | 0.9629 | 0.0506 | 22.1459 | 0.8504 | 0.1823 | 0.3037 | 3.0472 | 0.1725 | 38.8480 | 43.16M | 665.42G | 4.16s |
| | SGDN | AAAI'25 | 32.6625 | 0.9674 | 0.0483 | 23.1528 | 0.8631 | 0.1662 | 0.3056 | 3.0725 | 0.1855 | 40.1860 | 11.09M | 53.40G | 0.84s |
| Unsupervised Desmoking | DCP-P2P | Access'20 | 23.1149 | 0.8937 | 0.2218 | 21.4891 | 0.8409 | 0.2399 | 0.2202 | 2.5458 | 0.1847 | 30.9309 | 54.42M | 194.00G | 0.82s |
| | Desmokelap | IJCARS'22 | 26.4311 | 0.9037 | 0.2217 | 23.2277 | 0.8470 | 0.2431 | 0.2273 | 2.8237 | 0.1776 | 32.5820 | 11.38M | 1571.00G | 0.65s |
| Supervised Desmoking | SSIM-PFAN | PMLR'20 | 27.4094 | 0.9579 | 0.0841 | 22.0537 | 0.8031 | 0.1961 | 0.2844 | 2.9750 | 0.1773 | 38.2428 | 51.65M | 180.15G | 0.35s |
| | PFAN | PRCV'23 | 26.0183 | 0.8827 | 0.2031 | 21.0142 | 0.8218 | 0.2391 | 0.2458 | 2.7544 | 0.1692 | 32.3478 | 54.41M | 71.80G | 0.07s |
| | AALIDNet | ROBIO'24 | 26.5064 | 0.9226 | 0.2125 | 21.9999 | 0.8171 | 0.3410 | 0.2780 | 2.4094 | 0.1895 | 34.0062 | 33.11M | 7.88G | 0.03s |
| | SelfSVD | ECCV'24 | 29.5132 | 0.9588 | 0.0842 | 22.1343 | 0.8376 | 0.1980 | 0.3154 | 3.1195 | 0.1784 | 38.4157 | 27.70M | 996.00G | 0.18s |
| | **Ours** | – | **33.5345** | **0.9733** | **0.0402** | **23.8427** | **0.8813** | **0.1446** | **0.3271** | **3.1328** | **0.1914** | **40.2063** | 25.15M | 174.13G | 0.21s |

Table 1: The qualitative comparison between different methods. ↓ indicates that lower values are better, while ↑ indicates that higher values are better. The best results are highlighted in bold, and the second-best results are underlined.

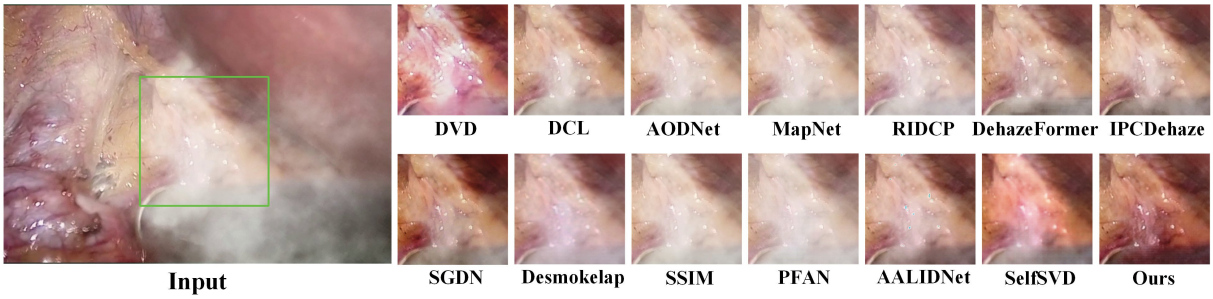


Figure 4: The qualitative comparison between different methods. More results are given on the paper website.

**Qualitative Comparisons** Fig.4 illustrates the qualitative results of different methods. It can be observed that the best-performing dehazing methods SGDN and DehazeFormer suffer from large areas of residual haze and lack color details of organ tissue, because they are difficult to utilize motion pattern information of surgical smoke and temproal complementarity between smoky frames. Meanwhile, compared to our method, the best-performing desmoking method SelfSVD tends to struggle with recovering clear textures in dense and entangled smoke regions due to the lack of smoke-type consideration, which limits its adaptability to complex smoky scenarios. Overall, our method can significantly remove varying smoke and reconstruct faithful details, which demonstrates the effectiveness and generalization of the smoke-type-aware desmoking framework.

**Complexity Comparison** As shown in Tab.1, our method has certain advantages in some dimensions of complexity compared to methods with the best desmoking performance. Specifically, our method outperforms SGDN in terms of time cost, DehazeFormer in terms of parameters, and SelfSVD in terms of FLOPS. It is worth noting that our method enables a dynamic activation strategy based on smoke-type guidance to avoid redundant computations as much as possible, when only one type of smoke appears in the video.

## Ablation Study

Tab.2 shows the objective results of 6 ablation states, which cover various ablation components such as training dataset, network module and loss function. It can be observed that compared to M1, M2 achieves significant improvements across all reference and no-reference metrics. This is attributed to the proposed STSVD dataset, which surpasses existing PSv2rs dataset in terms of smoke fidelity and scene diversity. Based on M2, M3 further improves reference metrics PSNR, SSIM, LPIPS and no-reference metrics TOPIQ, Q-align, MANIQA, MUSIQ by an average of 0.3791 dB, 0.0088, 0.0082 and 0.0015, 0.0339, 0.0027, 1.1203, respectively. This demonstrates that the $S^3M$ (semantic soft segmentation module) can jointly predict accurate smoke masks and types to provide valuable guidance for desmoking. Compared with M3, M5 improves reference metrics by an average of 0.7836 dB, 0.0059, 0.0064, and no-reference metrics by 0.0130, 0.0387, 0.0042, 0.5270, respectively. This is because the C2FDM (coarse-to-fine disentanglement module), along with the $\mathcal{L}_{shwl}$ loss function, addresses the entanglement challenges of two smoke types and yields more accurate disentangled masks. After incorporating the SVRS (smokeless video reconstruction sub-network), M6 achieves further improvements across all metrics by leveraging two types of smoke masks to guide special desmoking on video

| Index | Ablation Component | | | | | **STSVD**(synthetic dataset) | | | **Vivo**(paired real-world dataset) | | | **STSVD-R**(unpaired real-world dataset) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | $S^3M$ | C2FDM | SHWL | SVRS | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | TOPIQ↑ | Q-align↑ | MANIQA↑ | MUSIQ↑ |
| M1 | PSv2rs | | | | | 29.2978 | 0.9104 | 0.0835 | 22.9287 | 0.8515 | 0.1971 | 0.2449 | 2.9621 | 0.1806 | 34.6328 |
| M2 | STSVD | | | | | 31.1898 | 0.9577 | 0.0589 | 23.0744 | 0.8634 | 0.1639 | 0.3099 | 3.0540 | 0.1822 | 38.2087 |
| M3 | STSVD | ✓ | | | | 31.8980 | 0.9632 | 0.0545 | 23.1243 | 0.8755 | 0.1520 | 0.3114 | 3.0879 | 0.1849 | 39.3290 |
| M4 | STSVD | ✓ | ✓ | | | 32.7462 | 0.9683 | 0.0491 | 23.3514 | 0.8784 | 0.1498 | 0.3201 | 3.1207 | 0.1886 | 39. 5179 |
| M5 | STSVD | ✓ | ✓ | ✓ | | 33.0871 | 0.9716 | 0.0445 | 23.5024 | 0.8789 | 0.1492 | 0.3244 | 3.1266 | 0.1891 | 39.8560 |
| **M6** | STSVD | ✓ | ✓ | ✓ | ✓ | **33.5345** | **0.9733** | **0.0402** | **23.8427** | **0.8813** | **0.1446** | **0.3271** | **3.1328** | **0.1914** | **40.2063** |

Table 2: The quantitative comparisons of ablation study. The checkmark (✓) indicates whether an ablation component is included.

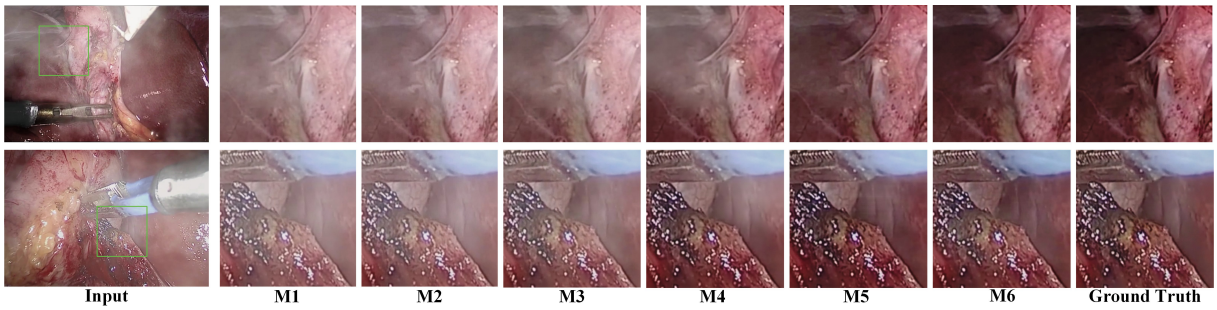


Figure 5: The qualitative comparisons of ablation study. More results are given on the paper website.

features.

Corresponding to objective results, subjective results in Fig.5 also reflect the improvement effect of different ablation components in our method. From M1 to M6, the visual quality of smoke removal shows progressive refinement, as each component incrementally contributes to more accurate mask prediction of different smoke types and more faithful color details restoration of organ tissue.

| Method | **Detection** | | | | | | **Segmentation** | |
|---|---|---|---|---|---|---|---|---|
| | CVC-ClinicDB | | CVC-ColonDB | | Kvasir-SEG | | EndoVis18 | |
| | DSC↑ | IoU↑ | DSC↑ | IoU↑ | DSC↑ | IoU↑ | IoU↑ | mcIoU↑ |
| Smoky | 0.8133 | 0.7813 | 0.6597 | 0.5846 | 0.8494 | 0.8054 | 58.512 | 38.588 |
| DeFormer | 0.8881 | 0.8203 | 0.6694 | 0.5916 | 0.8830 | 0.8194 | 72.768 | 49.336 |
| SGDN | 0.8933 | 0.8315 | 0.6828 | 0.6034 | 0.8822 | 0.8190 | 74.227 | 51.795 |
| SelfSVD | 0.8878 | 0.8254 | 0.6932 | 0.6152 | 0.8843 | 0.8205 | 68.403 | 45.463 |
| **Ours** | **0.9024** | **0.8394** | **0.7138** | **0.6331** | **0.8857** | **0.8240** | **74.814** | **51.899** |

Table 3: Quantitative results on downstream tasks (polyp detection and instrument segmentation).

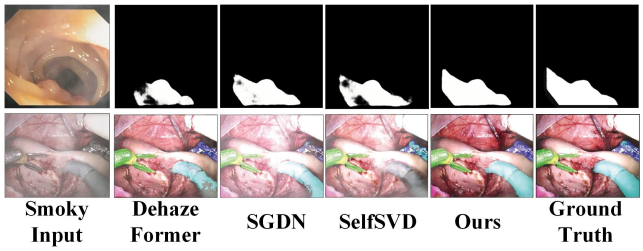


Figure 6: Qualitative results on downstream tasks (polyp detection and instrument segmentation).

## Evaluation on Downstream Tasks

To investigate the broader applicability of our method, we further evaluate its desmoking performance on two downstream tasks: polyp detection (Wei, Jiang, and Xu 2025) and surgical instrument segmentation (Yue et al. 2024), which include three detection task testing datasets CVC-ClinicDB, CVC-ColonDB, Kvasir-SEG and one segmentation task testing dataset EndoVis18. Specifically, we respectively apply our method and three best existing methods in comparative experiments to generate desmoked inputs for downstream evaluation. As shown in Table 3, our method achieves the best performance on all datasets, surpassing the second-best method by 0.0122 in DSC and 0.0118 in IoU on the detection task, while by 0.587 in IoU and 0.104 in mcIoU on the segmentation task. Qualitative comparisons in Fig.6 further reveal that our method produces more accurate detection targets and clearer segmentation boundaries. This demonstrates that our method can effectively facilitate downstream medical tasks in smoky conditions.

## Conclusion

In this paper, we propose the first smoke-type-aware laparoscopic video desmoking framework that explicitly differentiates Diffusion Smoke and Ambient Smoke. In addition, a coarse-to-fine mask disentanglement module is designed to improve mask accuracy for two types of smoke in entangled regions. To facilitate future research, we release a large-scale synthetic video desmoking dataset with smoke type annotations. Extensive experiments on synthetic and real datasets demonstrate superior performance of our method, benefiting downstream surgical tasks.

## Acknowledgments

This work was supported in part by the Hubei Provincial Science and Technology Plan Project (No. 2025CSA057), and the Transformation Fund Project of Scientific and Technological Achievements of Zhongnan Hospital of Wuhan University (No. 2023CGZH-ZD006).

## References

Bédard, S.; Karthik, E. N.; Tsagkas, C.; Pravatà, E.; Granziera, C.; Smith, A.; Weber II, K. A.; and Cohen-Adad, J. 2025. Towards contrast-agnostic soft segmentation of the spinal cord. *Medical Image Analysis*, 101: 103473.

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33: 2404–2418.

Chen, L.; Tang, W.; John, N. W.; Wan, T. R.; and Zhang, J. J. 2019. De-smokeGCN: generative cooperative networks for joint surgical smoke detection and removal. *IEEE transactions on medical imaging*, 39(5): 1615–1625.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Fan, J.; Wang, K.; Yan, Z.; Chen, X.; Gao, S.; Li, J.; and Yang, J. 2025. Depth-centric dehazing and depth-estimation from real-world hazy driving video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2852–2860.

Fan, J.; Weng, J.; Wang, K.; Yang, Y.; Qian, J.; Li, J.; and Yang, J. 2024. Driving-video dehazing with non-aligned regularization for safety assistance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26109–26119.

Fang, W.; Fan, J.; Zheng, Y.; Weng, J.; Tai, Y.; and Li, J. 2025. Guided real image dehazing using ycbcr color space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2906–2914.

Fu, J.; Liu, S.; Liu, Z.; Guo, C.-L.; Park, H.; Wu, R.; Wang, G.; and Li, C. 2025. Iterative Predictor-Critic Code Decoding for Real-World Image Dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12700–12709.

Giannarou, S.; Stoyanov, D.; Noonan, D.; Mylonas, G.; Clark, J.; Visentini-Scarzanella, M.; Mountney, P.; and Yang, G. 2014. Hamlyn centre laparoscopic/endoscopic video datasets.

Guo, C.-L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and Li, C. 2022. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5812–5820.

Holl, P.; Koltun, V.; Um, K.; and Thuerey, N. 2020. Phiflow: A differentiable pde solving framework for deep learning via physical simulations. In *NeurIPS workshop*, volume 2.

Hong, T.; Huang, P.; Zhai, X.; Gu, C.; Tian, B.; Jin, B.; and Li, D. 2023. MARS-GAN: multilevel-feature-learning attention-aware based generative adversarial network for removing surgical smoke. *IEEE Transactions on Medical Imaging*, 42(8): 2299–2312.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5148–5157.

Kim, T.; Thürey, N.; James, D.; and Gross, M. 2008. Wavelet turbulence for fluid simulation. *ACM Transactions on Graphics (TOG)*, 27: 1–6.

Li, B.; Peng, X.; Wang, Z.; Xu, J.; and Feng, D. 2017. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, 4770–4778.

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2018. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28(1): 492–505.

Liu, L.; Prost, J.; Zhu, L.; Papadakis, N.; Liò, P.; Schönlieb, C.-B.; and Aviles-Rivero, A. I. 2023. Scotch and soda: A transformer video shadow detection framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10449–10458.

Liu, R.; Lehman, J.; Molino, P.; Petroski Such, F.; Frank, E.; Sergeev, A.; and Yosinski, J. 2018. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31.

Liu, Z.; Gao, W.; Zhu, J.; Liu, B.; and Fu, Y. 2024. Smoke Attention Based Laparoscopic Image Desmoking Network with Hybrid Guided Embedding. In *2024 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1018–1023. IEEE.

Ma, K.; Fang, Y.; Weibel, J.-B.; Tan, S.; Wang, X.; Xiao, Y.; Fang, Y.; and Xia, T. 2025. Phys-Liquid: A Physics-Informed Dataset for Estimating 3D Geometry and Volume of Transparent Deformable Liquids. arXiv:2511.11077.

Pan, Y.; Bano, S.; Vasconcelos, F.; Park, H.; Jeong, T. T.; and Stoyanov, D. 2022. DeSmoke-LAP: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 17(5): 885–893.

Salazar-Colores, S.; Jiménez, H. M.; Ortiz-Echeverri, C. J.; and Flores, G. 2020. Desmoking Laparoscopy Surgery Images Using an Image-to-Image Translation Guided by an Embedded Dark Channel. *IEEE Access*, 8: 208898–208909.

Sidorov, O.; et al. 2020. Generative smoke removal. In *Machine Learning for Health Workshop*, 81–92. PMLR.

Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32: 1927–1941.

Stauder, R.; Ostler, D.; Kranzfelder, M.; Koller, S.; Feußner, H.; and Navab, N. 2016. The TUM LapChole dataset for the M2CAI 2016 workflow challenge. *arXiv preprint arXiv:1610.09278*.

Twinanda, A. P.; Shehata, S.; Mutter, D.; Marescaux, J.; De Mathelin, M.; and Padoy, N. 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1): 86–97.

Wei, S.; Jiang, J.; and Xu, X. 2025. UniNet: A Contrastive Learning-guided Unified Framework with Feature Selection for Anomaly Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9994–10003.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023a. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Wu, R.; Zhang, Z.; Zhang, S.; Gou, L.; Chen, H.; Zhang, L.; Chen, H.; and Zuo, W. 2024. Self-supervised video desmoking for laparoscopic surgery. In *European Conference on Computer Vision*, 307–324. Springer.

Wu, R.-Q.; Duan, Z.-P.; Guo, C.-L.; Chai, Z.; and Li, C. 2023b. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22282–22291.

Xia, W.; Fan, V.; Peters, T.; and Chen, E. C. 2024. A new benchmark in vivo paired dataset for laparoscopic image de-smoking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 3–13. Springer.

Xu, J.; Hu, X.; Zhu, L.; Dou, Q.; Dai, J.; Qiao, Y.; and Heng, P.-A. 2023. Video dehazing via a multi-range temporal alignment network with physical prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18053–18062.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1191–1200.

Yuan, S.; Chen, J.; Li, J.; Jiang, W.; and Guo, S. 2023. Lhnet: A low-cost hybrid network for single image dehazing. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7706–7717.

Yue, W.; Zhang, J.; Hu, K.; Xia, Y.; Luo, J.; and Wang, Z. 2024. Surgicalsam: Efficient class promptable surgical instrument segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6890–6898.

Zhang, J.; Huang, W.; Liao, X.; and Wang, Q. 2023. Progressive frequency-aware network for laparoscopic image desmoking. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 479–492. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.