Benchmarking Endoscopic Surgical Image Restoration and Beyond
⟨ Supplementary Material ⟩

Jialun Pei[1] Diandian Guo[1] Donghui Yang[2] Zhixi Li[3]
Yuxin Feng[4] Long Ma[2*] Bo Du[5] Pheng-Ann Heng[1]

[1] The Chinese University of Hong Kong [2] Dalian University of Technology
[3] Southern Medical University [4] Xidian University [5] Wuhan University

A. Overview

We summarize the supplementary material from the following sections:

- **Section B: SurgClean Dataset.** More sample analysis and exhibitions for the proposed SurgClean dataset.
- **Section C: Visual Analysis with ESIR Datasets.** Data visual comparison with other open-source endoscopic surgical image restoration datasets.
- **Section D: More Benchmarking Results.** The comparative results on fine-grained subsets of SurgClean and additional visualization results.
- **Section E: Additional Ablation Studies.** We conducted additional ablation experiments to verify the impact of optical flow alignment on model performance.
- **Section F: Different Training Strategies.** The effect of different pre-training weights and training strategies on natural and surgical scene data.

B. SurgClean Dataset

B.1. Clarity for Label Alignment

As mentioned in Section 3.2, it is impossible to obtain perfectly aligned reference labels that match degraded images at the exact same moment in real-world endoscopic surgical videos. Therefore, we extract frames from the video just before the scene becomes degraded to obtain an approximate paired image, which inevitably introduces misalignment. The misalignment in our study is the same as the issue in LSVD [14]. Therefore, the optical flow alignment method based on PWC-Net [11] is equally applicable to our benchmark. Fig. 1 exhibits the visualization of reference labels before and after alignment on SurgClean dataset. It can be observed that the optical flow-based warping method effectively converts misaligned images to alignments and masks out the error regions.

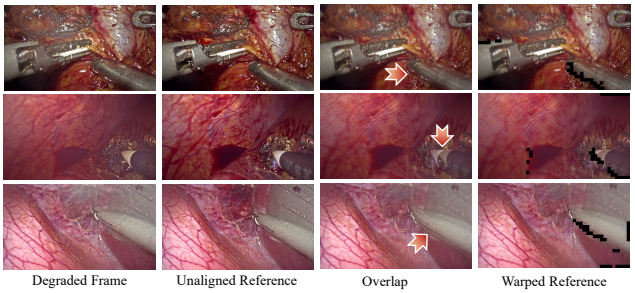

Figure 1. Visual comparison for reference label alignment on the SurgClean dataset.

B.2. More Data Visualization

To further illustrate the composition of SurgClean, we provide additional visualizations for Defogging, Desmoking, and Desplashing tasks in Fig. 2, Fig. 3, and Fig. 4. These plots reveal that degradation types are hierarchically distributed across diverse surgical scenarios, reflecting the intrinsic variability of real endoscopic environments. These fine-grained visualizations offer an intuitive understanding of both degradation diversity and clinical representativeness within SurgClean.

C. Visual Analysis with ESIR Datasets

We conducted a comprehensive visual analysis comparing the SurgClean dataset with existing ESIR benchmarks, specifically evaluating Desmoke-LAP [10] and Desmoke-Data [15]. (Note that Cyclic-DesmokeGAN [13] was excluded from this comparison due to its non-public availability.) As illustrated in Fig. 5, critical distinctions emerge between these datasets. Both Desmoke-LAP and Desmoke-Data exhibit inherent limitations: they are generated from single surgical sources, which restricts their applicability exclusively to the desmoking task. Furthermore, Desmoke-LAP suffers from the additional constraint of containing unpaired training samples. Meanwhile, Desmoke-LAP and

*Corresponding author. (malone94319@gmail.com)

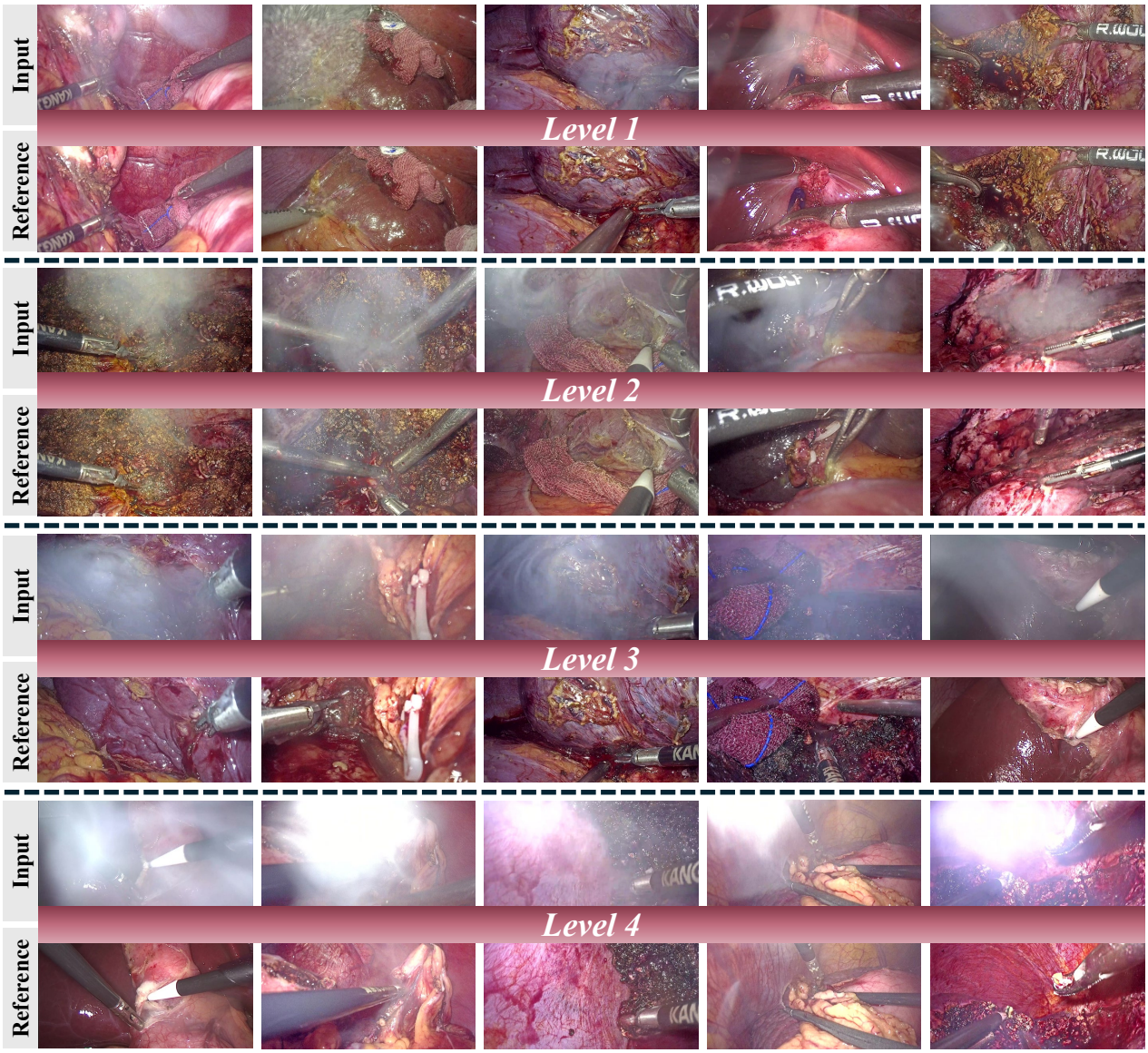

Figure 2. More fine-grained data exhibitions for desmoking in SurgClean dataset.

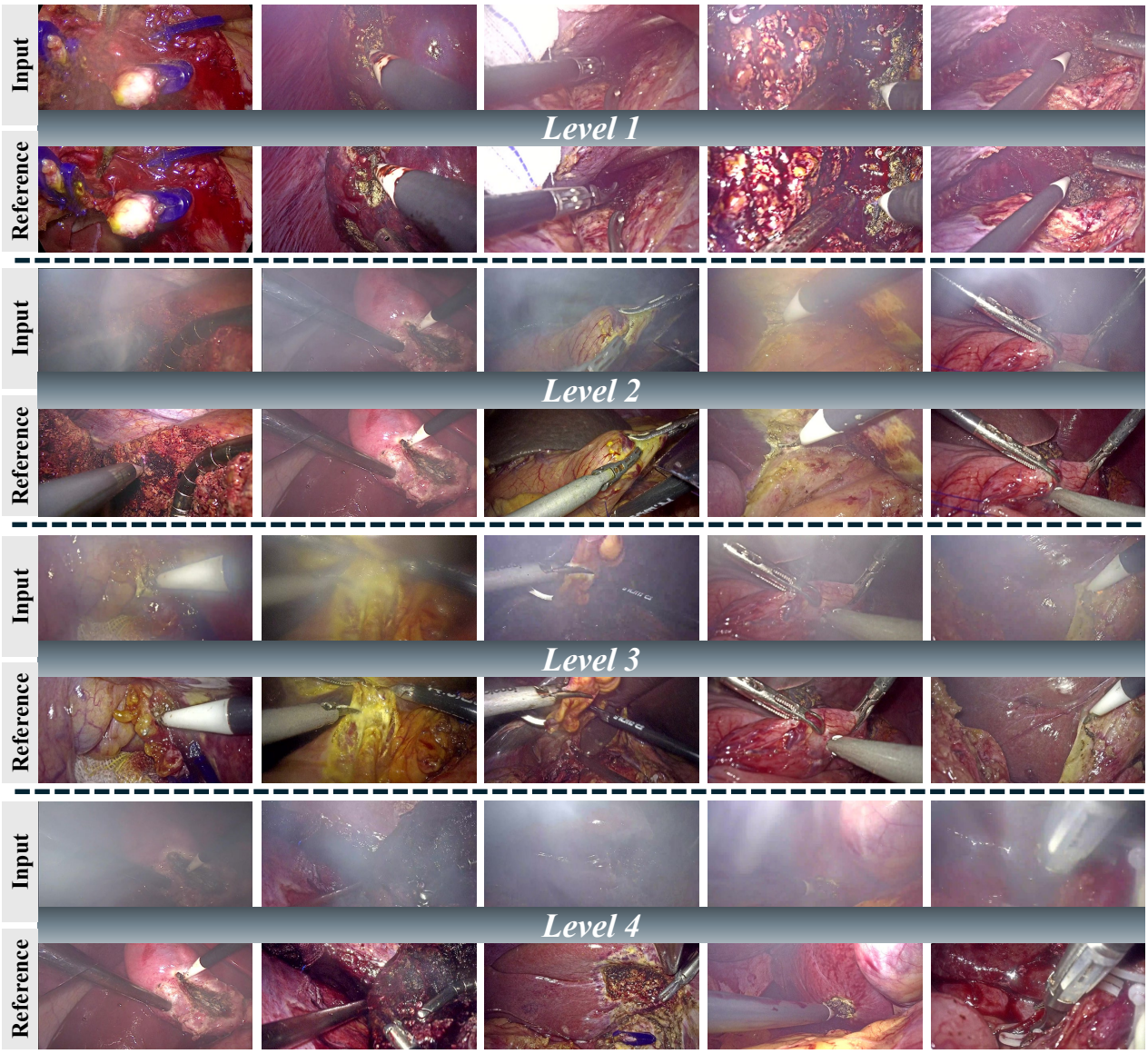

Figure 3. More fine-grained data exhibitions for defogging in SurgClean dataset.

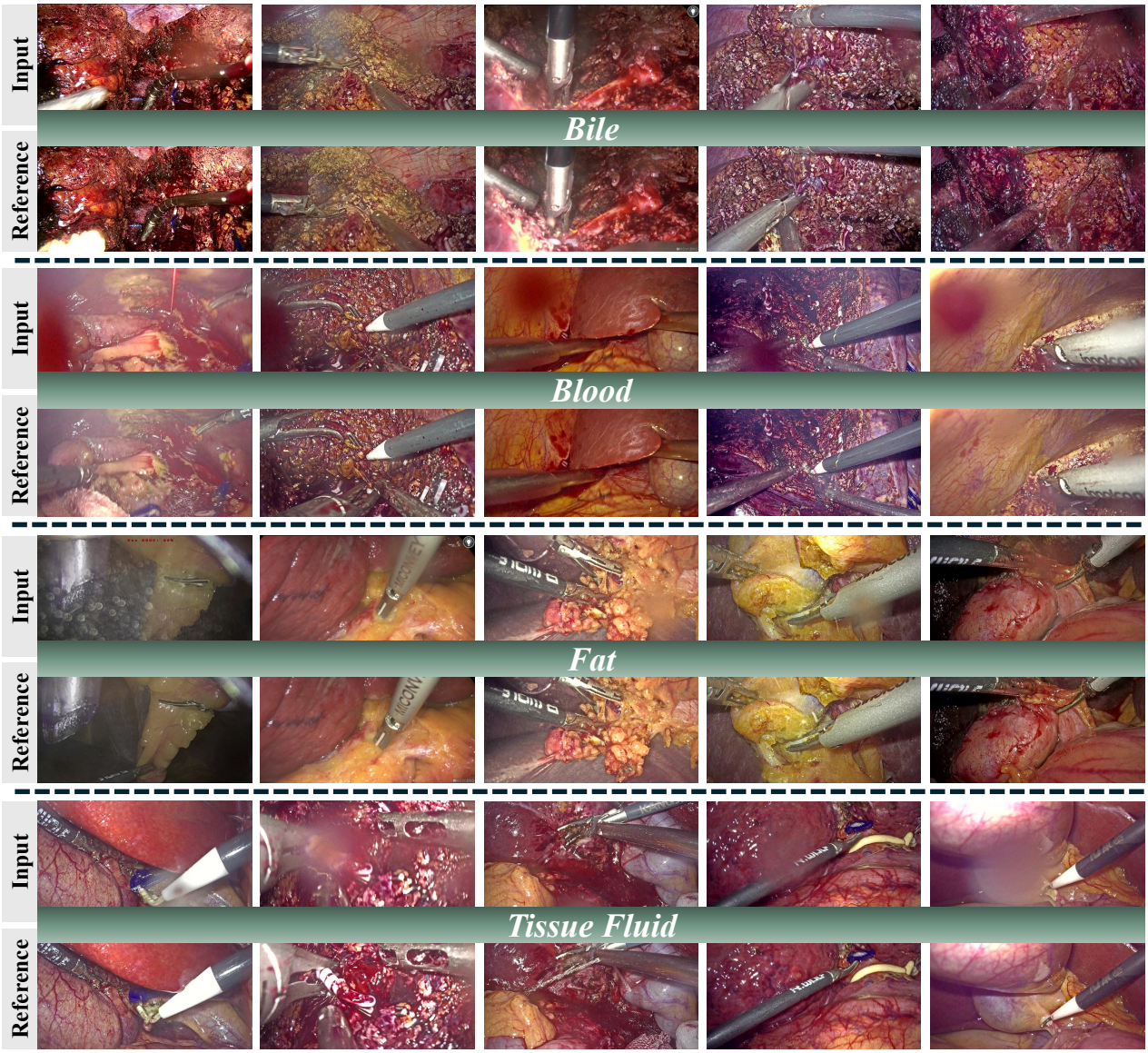

Figure 4. More fine-grained data exhibitions for desplashing in SurgClean dataset.

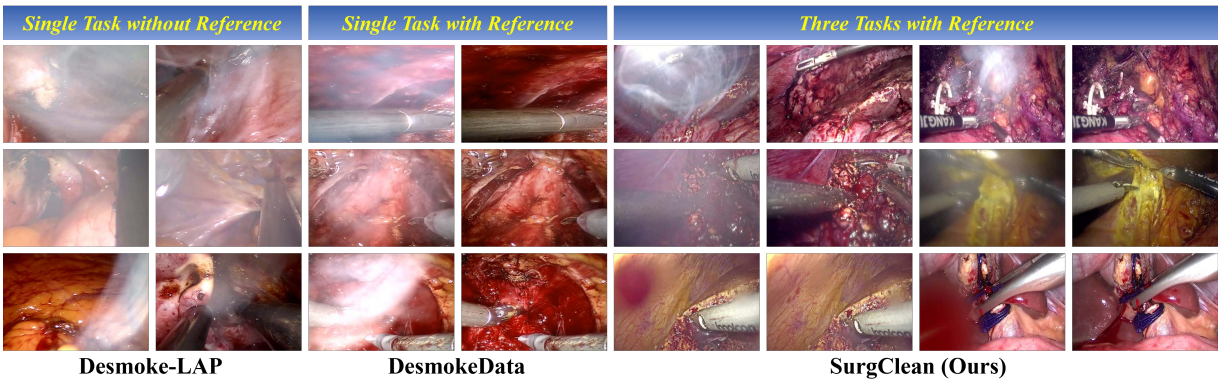

Figure 5. Visual comparison with ESIR datasets.

Table 1. Fine-grained comparison with generic restoration models for desmoking and defogging on sub-test sets of SurgClean dataset.

Tasks	Methods	Publications	Level-1		Level-2		Level-3		Level-4	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Desmoking	Restormer [17]	CVPR'22	20.7850	0.6757	17.5388	0.6370	15.0176	0.5631	13.7141	0.5944
	FocalNet [3]	ICCV'23	21.5067	0.6898	18.5554	0.6141	16.9457	0.5807	15.3117	0.5666
	ConvIR [4]	TPAMI'24	21.4525	0.6818	18.5778	0.6066	17.0121	0.5695	15.2747	0.5515
	Fourmer [18]	ICML'24	20.6496	0.6988	18.5020	0.6637	16.4806	0.5908	15.3344	0.6167
	MambaIR [5]	ECCV'24	21.0877	0.6840	18.5940	0.6078	17.2621	0.5745	15.8857	0.5603
	Histoformer [12]	ECCV'24	18.4626	0.3977	16.4624	0.3367	15.1055	0.3095	13.9213	0.2765
	RAMiT [2]	CVPR'24	20.4156	0.7056	16.6750	0.6591	14.3269	0.5828	12.8964	0.6116
	AMIR [16]	MICCAI'24	21.6867	0.6887	18.6949	0.6098	17.2324	0.5831	15.5010	0.5609
	AST [19]	CVPR'24	21.1211	0.6776	18.6088	0.6281	16.7118	0.5289	15.8765	0.5850
	X-Restormer [1]	ECCV'24	21.0098	0.7119	17.4058	0.6601	15.2322	0.5550	13.4543	0.6074
	SFHformer [7]	ECCV'24	20.5770	0.6904	17.1439	0.6444	14.6852	0.5722	13.5559	0.5982
	MambaIRv2 [6]	CVPR'25	21.1561	0.6857	18.6954	0.6099	17.3127	0.5772	15.8400	0.5615
Defogging	Restormer [17]	CVPR'22	21.3809	0.6331	18.2525	0.5626	16.5593	0.5362	14.6312	0.5076
	FocalNet [3]	ICCV'23	21.0877	0.7093	17.9446	0.6654	15.6601	0.5945	14.7192	0.6244
	ConvIR [4]	TPAMI'24	21.1809	0.7072	18.4453	0.6686	15.8753	0.5943	14.4820	0.6168
	Fourmer [18]	ICML'24	20.7729	0.6775	18.4023	0.6042	17.2244	0.5723	16.0295	0.5591
	MambaIR [5]	ECCV'24	20.8506	0.7085	18.5537	0.6701	16.4459	0.5945	15.2686	0.6221
	Histoformer [12]	ECCV'24	18.8211	0.4534	17.1464	0.4405	15.2224	0.3950	14.0211	0.3948
	RAMiT [2]	CVPR'24	21.6109	0.6878	18.6471	0.6096	16.8567	0.5731	15.2260	0.5556
	AMIR [16]	MICCAI'24	21.0818	0.7096	17.8085	0.6664	15.3175	0.5906	13.9828	0.6180
	AST [19]	CVPR'24	17.5711	0.5887	16.6786	0.5636	16.4697	0.6076	15.3710	0.5840
	X-Restormer [1]	ECCV'24	19.5498	0.6528	18.7151	0.6179	17.8697	0.6079	16.7165	0.5986
	SFHformer [7]	ECCV'24	21.3376	0.6794	18.8253	0.6051	17.1736	0.5723	15.8657	0.5618
	MambaIRv2 [6]	CVPR'25	21.0578	0.7095	18.2369	0.6670	15.8925	0.5921	14.5339	0.6168

DesmokeData provide lower-resolution images that may hinder model performance in fine-grained surgical scenes.

In contrast, SurgClean demonstrates significant advancements in dataset design. First, it incorporates meticulously paired data collected across four distinct surgical procedures, including Cholecystectomy, Hepatectomy, Pancreatectomy, and Splenectomy, ensuring broader clinical relevance. Second, our dataset supports three advanced vision restoration tasks (*i.e.*, desmoking, defogging, and desplashing), surpassing the single-task focus of existing alternatives. Third, SurgClean provides relatively high-resolution imaging data that preserves critical details, enhancing key limitations of prior datasets.

D. More Benchmarking Results

D.1. Fine-grained evaluation on SurgClean

To further validate the performance of representative methods, we also report PSNR and SSIM values of sub-test sets for three restoration tasks on SurgClean in Table 1 and Table 2, *i.e.*, Level 1–4 for desmoking and defogging, and bile, blood, fat, and tissue fluid for desplashing. The fine-grained evaluation highlights several important findings. First, in the desmoking and defogging tasks, as the degradation level increased, the restoration performance of evaluation models consistently declined, indicating that existing methods exhibit insufficient restoration robustness when the degra-

Table 2. Fine-grained comparison with generic restoration models for desplashing on sub-test sets of SurgClean dataset.

Tasks	Methods	Publications	Bile		Blood		Fat		Tissue Fluid	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Desplashing	Restormer [17]	CVPR'22	20.4006	0.7009	20.0409	0.6629	22.7995	0.7331	21.6047	0.7295
	FocalNet [3]	ICCV'23	20.7119	0.7188	19.8308	0.6731	23.1669	0.7402	21.6969	0.7322
	ConvIR [4]	TPAMI'24	20.6809	0.7255	19.5713	0.6686	23.1313	0.7404	21.6604	0.7330
	Fourmer [18]	ICML'24	20.6454	0.7249	19.7375	0.6657	23.0745	0.7333	21.6263	0.7251
	MambaIR [5]	ECCV'24	20.7461	0.7291	19.7011	0.6743	23.2169	0.7442	21.7519	0.7372
	Histoformer [12]	ECCV'24	19.5254	0.6060	18.5349	0.5361	21.5869	0.5840	20.3437	0.5790
	RAMiT [2]	CVPR'24	20.7680	0.7228	19.8221	0.6719	23.1750	0.7410	21.7022	0.7338
	AMIR [16]	MICCAI'24	20.5077	0.7187	19.8158	0.6705	23.0523	0.7415	21.5080	0.7337
	AST [19]	CVPR'24	20.2490	0.6984	21.4021	0.6829	23.8603	0.7588	21.6277	0.7197
	X-Restormer [1]	ECCV'24	20.2838	0.7095	21.2440	0.6867	23.8657	0.7366	21.7215	0.7216
	SFHformer [7]	ECCV'24	20.7504	0.7169	19.7931	0.6728	23.2287	0.7403	21.7073	0.7357
	MambaIRv2 [6]	CVPR'25	20.7246	0.7135	21.4479	0.6888	23.0982	0.7403	21.8318	0.7264

dation level increases. For example, while AMIR [16] achieves the best PSNR (21.6867) at Level-1 desmoking, its performance drops sharply to 15.5010 at Level-4. Second, for fine-grained desplashing, there are significant differences between different splash substances. MambaIR [5] and MambaIRv2 [6] show relatively stable restoration results across different sub-test sets. However, no single model dominates all levels and types of degradation scenarios. Overall, the fine-grained experimental results indicate that endoscopic surgical image restoration still faces significant challenges under diverse and severe degradations.

D.2. Additional Visual Results

We provide more qualitative results of representative image restoration models with superior quantitative results for desmoking, defogging, and desplashing on the SurgClean test set, including ConvIR [4], FocalNet [3], AST [19], Restormer [17], RAMiT [2], AMIR [16], MambaIR [5], SFHformer [7]. As shown in Fig. 6 and Fig. 7, the restoration models are relatively effective at lower degradation levels. When the smoke and fog levels reach 3 or even 4, the restoration visualization effect is no longer noticeable. It indicates that the noisy samples in the SurgClean dataset pose significant challenges for surgical image restoration. As shown in Fig. 8, for the brand-new desplashing task, all evaluated models did not show obvious responses to degraded regions. It reflects that desplashing is more challenging than the other two restoration tasks, which drives us to explore more specific algorithms in follow-up research to solve multi-type splashing problems in the surgical restoration domain.

E. Ablation Study for Optical Flow Alignment

We conducted a quantitative ablation experiment to validate the effectiveness of optical flow alignment for the desmoking task on SurgClean test set. As shown in Table 3, disabling optical flow consistently leads to a performance drop

Table 3. Comparisons for w/ & w/o optical flow alignment.

Methods	w/ Optical Flow		w/o Optical Flow	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Restormer [17]	18.470	0.634	17.905	0.610
FocalNet [3]	18.347	0.655	17.863	0.611
ConvIR [4]	18.464	0.649	18.005	0.624
Fourmer [18]	18.208	0.642	17.568	0.582
MambaIR [5]	18.357	0.649	17.922	0.615
Histoformer [12]	16.687	0.376	16.251	0.342
RAMiT [2]	18.348	0.647	17.842	0.621
AMIR [16]	18.097	0.637	17.787	0.604
AST [19]	18.630	0.617	18.249	0.579
X-Restormer [1]	17.570	0.647	17.179	0.618
SFHformer [7]	17.965	0.637	17.483	0.601
MambaIRv2 [6]	18.572	0.638	17.943	0.593

across all models, *e.g.*, MambaIRv2 shows a decrease from 18.572 to 17.943 in PSNR and from 0.638 to 0.593 in SSIM. This is mainly due to the lack of precise spatial correspondence caused by unpaired supervision, which introduces pixel-level misalignment and leads to structural blurring in the reconstructed outputs. We will further explore better alignment strategies for surgical image restoration in follow-up research.

F. Results of Different Training Strategies

While the image restoration task has been extensively studied in natural scenes, its application to surgical scenarios presents unique challenges due to the distinct characteristics of endoscopic imaging. To systematically evaluate cross-domain generalization capabilities, we design a two-stage experimental scheme to validate the effect of natural and surgical scene data on the performance of restoration models in surgical defogging. In the first stage, we implement 12 representative restoration algorithms pretrained on Haze4K [9] and RESIDE [8], two datasets optimized for haze removal in natural scenes with both synthetic and realistic images, and directly deploy them on

Table 4. Comparing different training strategies for defogging on the SurgClean test sets.

Methods	Haze4K [9]										REISDE [8]									
	Trained on Natural Fog					*Finetuned on Surgical Fog*					*Trained on Natural Fog*					*Finetuned on Surgical Fog*				
	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	PI↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	PI↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	PI↓	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	PI↓
Restormer [17]	17.42	0.62	0.46	5.95	27.98	17.99	0.63	0.41	6.16	16.69	16.86	0.59	0.46	6.10	28.61	17.43	0.62	0.42	6.31	17.32
FocalNet [3]	16.17	0.55	0.42	4.79	21.02	18.78	0.62	0.44	5.97	32.33	15.63	0.57	0.46	6.09	31.31	17.92	0.63	0.45	7.30	39.21
ConvIR [4]	15.77	0.54	0.42	4.74	19.26	18.16	0.59	0.42	5.03	17.94	17.33	0.61	0.48	6.82	31.19	19.56	0.66	0.47	7.76	31.80
Fourmer [18]	17.25	0.60	0.45	5.80	30.84	17.61	0.62	0.47	6.66	33.99	17.09	0.60	0.45	6.17	30.79	17.45	0.62	0.48	7.03	33.94
MambaIR [5]	17.24	0.58	0.42	4.96	20.76	19.08	0.63	0.43	6.23	34.15	16.98	0.59	0.47	6.47	31.52	18.29	0.63	0.50	8.23	51.82
Histoformer [12]	18.01	0.59	0.41	4.61	11.01	15.23	0.31	0.64	6.22	16.20	16.58	0.59	0.47	6.02	26.34	15.68	0.54	0.48	5.74	18.52
RAMiT [2]	16.50	0.59	0.47	6.05	29.30	18.13	0.62	0.48	7.31	33.94	17.27	0.60	0.47	6.28	30.76	18.90	0.63	0.48	7.54	35.40
AMIR [16]	18.04	0.60	0.41	4.88	18.09	18.94	0.62	0.44	5.99	36.85	17.01	0.62	0.46	6.22	30.71	17.94	0.63	0.47	6.83	42.77
AST [19]	16.70	0.58	0.42	5.65	15.95	18.34	0.57	0.46	8.80	1.60	16.27	0.57	0.47	6.12	10.99	17.91	0.56	0.52	9.27	8.36
X-Restormer [1]	16.57	0.59	0.47	5.99	27.01	18.19	0.61	0.48	7.63	5.91	17.24	0.60	0.47	6.70	30.81	18.86	0.62	0.48	8.34	19.71
SFHformer [7]	16.06	0.57	0.47	5.93	23.74	18.00	0.61	0.48	7.01	5.44	17.06	0.60	0.44	6.07	22.09	19.00	0.64	0.47	7.15	23.79
MambaIRv2 [6]	18.19	0.61	0.40	5.27	23.28	19.13	0.63	0.43	6.16	38.42	16.59	0.59	0.47	6.28	33.94	17.57	0.63	0.48	7.17	53.04

the SurgClean sub-test set for defogging without any adaptation. As shown in Table 4, most models exhibit substantial performance degradation, revealing significant domain discrepancies between natural and surgical fog patterns.

To investigate domain adaptation potential, the second stage involves fine-tuning the pre-trained model using the defogging training set of SurgClean while maintaining their architectural configurations. The result confirms that surgical-specific data are indispensable for bridging domain gaps through targeted retraining. These findings further emphasize the necessity of developing specialized benchmarks for endoscopic scene restoration, as generic computer vision solutions prove insufficient for meeting the complex demands of clinical applications.

References

[1] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *ECCV*, pages 74–91, 2024. 5, 6, 7

[2] Haram Choi, Cheolwoong Na, Jihyeon Oh, Seungjae Lee, Jinseop Kim, Subeen Choe, Jeongmin Lee, Taehoon Kim, and Jihoon Yang. Reciprocal attention mixing transformer for lightweight image restoration. In *IEEE CVPR*, pages 5992–6002, 2024. 5, 6, 7

[3] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *IEEE ICCV*, pages 13001–13011, 2023. 5, 6, 7

[4] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Revitalizing convolutional network for image restoration. *IEEE TPAMI*, 2024. 5, 6, 7

[5] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, pages 222–241, 2024. 5, 6, 7

[6] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. *IEEE CVPR*, 2025. 5, 6, 7

[7] Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *ECCV*, pages 381–402, 2024. 5, 6, 7

[8] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE TIP*, 28(1):492–505, 2019. 6, 7

[9] Ye Liu, Lei Zhu, Shunda Pei, Huazhu Fu, Jing Qin, Qing Zhang, Liang Wan, and Wei Feng. From synthetic to real: Image dehazing collaborating with unlabeled real data. In *Proceedings of the 29th ACM international conference on multimedia*, pages 50–58, 2021. 6, 7

[10] Yirou Pan, Sophia Bano, Francisco Vasconcelos, Hyun Park, Taikyeong Ted Jeong, and Danail Stoyanov. Desmoke-lap: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *IJCARS*, 17(5):885–893, 2022. 1

[11] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE CVPR*, pages 8934–8943, 2018. 1

[12] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *ECCV*, pages 111–129, 2025. 5, 6, 7

[13] Vishal Venkatesh, Neeraj Sharma, Vivek Srivastava, and Munendra Singh. Unsupervised smoke to desmoked laparoscopic surgery images using contrast driven cyclic-desmokegan. *Computers in Biology and Medicine*, 123: 103873, 2020. 1

[14] Renlong Wu, Zhilu Zhang, Shuohao Zhang, Longfei Gou, Haobin Chen, Lei Zhang, Hao Chen, and Wangmeng Zuo. Self-supervised video desmoking for laparoscopic surgery. In *ECCV*, pages 307–324, 2024. 1

[15] Wenyao Xia, Victoria Fan, Terry Peters, and Elvis CS Chen. A new benchmark in vivo paired dataset for laparoscopic image de-smoking. In *MICCAI*, pages 3–13, 2024. 1

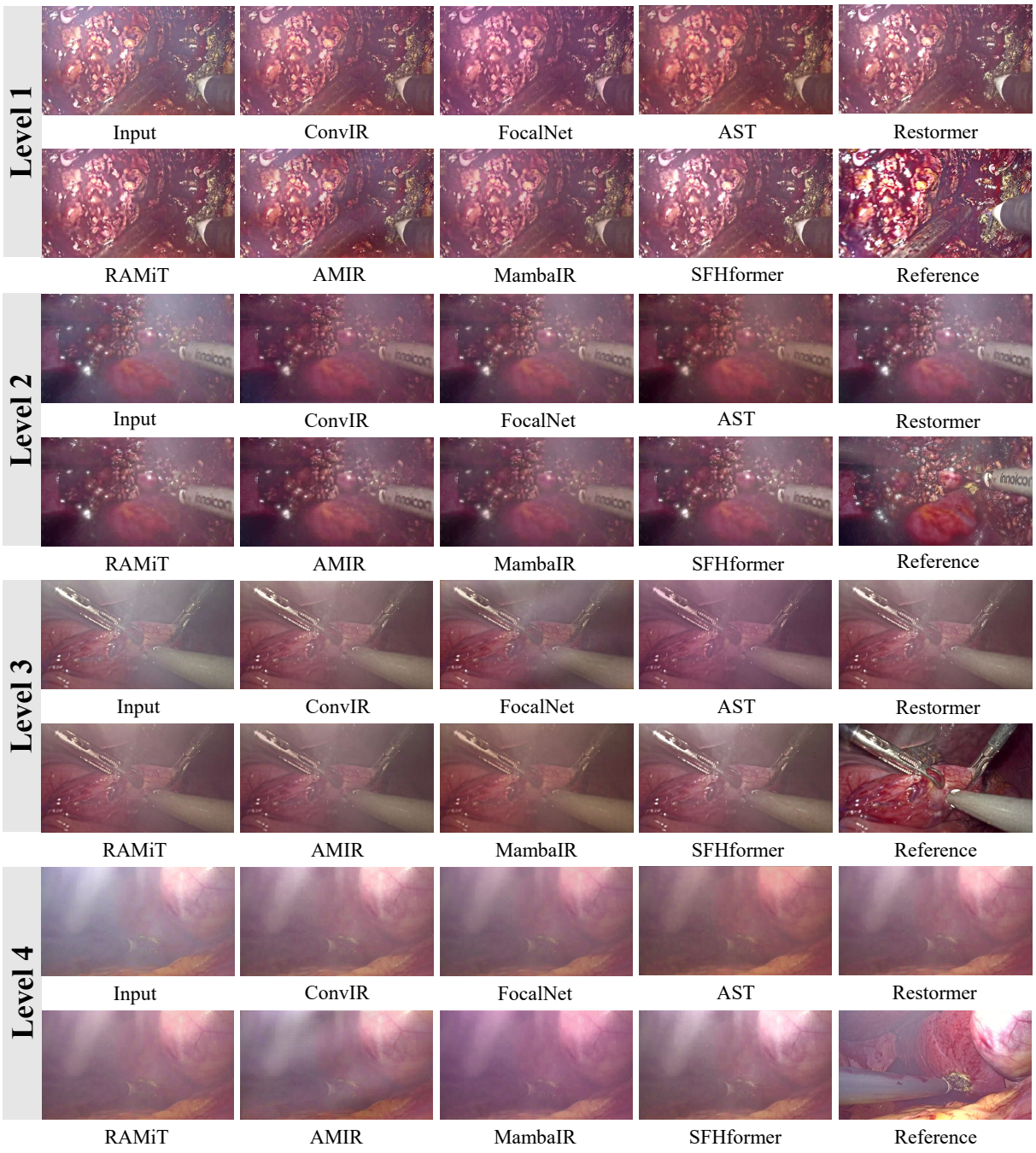

Figure 6. Additional visualizations of representative image restoration models for desmoking on SurgClean test set.

[16] Zhiwen Yang, Haowei Chen, Ziniu Qian, Yang Yi, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. All-in-one medical image restoration via task-adaptive routing. In *MICCAI*, pages 67–77, 2024. 5, 6, 7

[17] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE CVPR*, pages 5728–5739, 2022. 5, 6, 7

[18] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *ICML*, pages 42589–42601, 2023. 5, 6, 7

[19] Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *IEEE CVPR*, 2024. 5, 6, 7

Figure 7. Additional visualizations of representative image restoration models for defogging on SurgClean test set.

Figure 8. Additional visualizations of representative image restoration models for desplashing on SurgClean test set.