2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Endoscopic Spine Surgical View Enhancement via Diffusion-Prior Contrastive and Physics-Informed Constraints for Robotic Navigation

Haojie Han[1†], Longfei Ma[1†], Kai Xu[3], Suxi Gu[3], Shipeng Zhang[1], Guochen Ning[1], Fang Chen[2*] and Hongen Liao[1,2*]

***Abstract*— In robot-assisted spinal endoscopy, intraoperative imaging is frequently degraded by bleeding, irrigation fluids, bubbles, smoke, and uneven illumination, which can severely compromise surgical precision, safety, and decision-making. Accurate identification of anatomical structures is particularly critical in spinal procedures, yet acquiring paired clean and degraded images in real clinical settings is infeasible. To address this challenge, we propose DCP-Net, an unpaired endoscopic image restoration framework tailored for robotic spinal surgery. DCP-Net integrates Diffusion-Prior Contrastive Learning (DPCL) to leverage generative priors and contrastive objectives for robust latent representations, and Physics-Informed Constraints (PIC) to ensure anatomically consistent restoration. Furthermore, we introduce Diffusion-Prior Uncertainty Estimation (DPUE), providing pixel-wise confidence maps that quantify restoration reliability and guide risk-aware robotic perception. We further constructed a dataset comprising 21,845 paired/unpaired samples of intraoperative visual degradations in spinal endoscopy, primarily involving bleeding, bubbles, and other artifacts. Extensive experiments show that DCP-Net outperforms existing methods in both quantitative metrics and perceptual quality, significantly improving visual clarity and supporting various robotic navigation tasks. Among these tasks, accurate bleeding point detection plays a particularly critical role in ensuring safe and precise navigation in clinical practice.**

I. INTRODUCTION

Robot-assisted Minimally Invasive Surgery (RMIS) critically relies on high-quality endoscopic imaging to guide precise surgical maneuvers [1]. In spinal endoscopy, intraoperative views are often severely degraded by bleeding, irrigation fluids, bubbles, smoke, and uneven illumination, which can compromise both the surgeon's visibility and the performance of robotic perception modules supporting downstream tasks, such as bleeding point detection, visual navigation, and autonomous assistance. These degradations are particularly consequential in RMIS, where surgeons rely on indirect vision rather than tactile feedback, making visual clarity essential for procedural safety and efficiency.

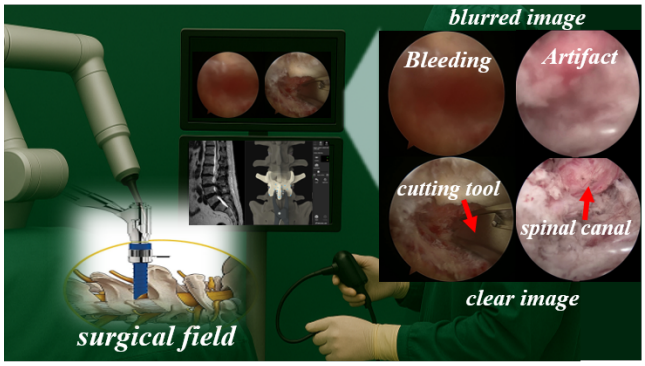

Fig. 1. The figure illustrates the blurred surgical view caused by intraoperative bleeding, artifact and irrigation fluidsduring spinal endoscopic surgery. After applying the deblurring model, the surgical view is significantly improved, thereby enhancing the safety of intraoperative procedures.

Acquiring paired clean and blurred images in real clinical settings is infeasible due to the complexity, variability, and rapid dynamics of operative scenes, further compounded by occlusions, instrument motion, and fluid interactions [2]. Unlike natural image restoration [3], surgical image restoration must preserve fine-grained anatomical structures to ensure accurate execution of critical intraoperative tasks, such as electrocautery hemostasis, vessel dissection, and precise instrument localization. Thus, effective restoration requires a careful balance between visual enhancement, structural fidelity, and task-oriented precision, directly impacting the reliability of robotic interventions.

Several studies have attempted to address laparoscopic smoke removal [4] by leveraging synthetic or real paired datasets for model training. Yang [5] proposed a method that integrates a U-Net backbone, a learnable Wiener filter, and a multi-objective loss function, achieving superior image clarity and objective metrics on a real paired dataset. Xia et al [6] constructed the first paired laparoscopic dataset with 961 real smoky/clear image pairs, and combined it with motion tracking to evaluate and compare existing de-smoking algorithms, providing valuable insights for future research. Chen et al [7]. proposed LighTDiff, a lightweight diffusion-based model that enables low-light image enhancement in surgical endoscopy while maintaining high computational efficiency. However, *few studies have focused on spinal endoscopy, where the surgical field is filled with fluid, and research*

*This work was supported in part by the National Key Research and Development Program of China (2024YFC2418101, 2022YFC2405200); in part by the National Natural Science Foundation of China (U22A2051); in part by the Science and Technology Commission of Shanghai Municipality (Nos.24511104100); and in part by the Natural Science Foundation of Beijing Municipality (L252126, L252201). *Corresponding authors: Hongen Liao and Fang Chen.*

[1] School of Biomedical Engineering, Tsinghua University, Beijing 100084, China. hanhj24@mails.tsinghua.edu.cn

[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. chen-fang@sjtu.edu.cn, liao@tsinghua.edu.cn

[3] Beijing Tsinghua Changgung Hospital, Orthopedics & Sports Medicine Center, Beijing, 102218, Beijing, China.

† Denotes equal contribution.

on image restoration in such scenarios remains scarce. Compared with laparoscopic surgery, spinal endoscopy poses greater challenges for image restoration: the surgical field is filled with irrigation fluid, where bleeding quickly diffuses and causes severe blur and occlusion. At the same time, the narrow working channel and limited field of view make clarity further compromised by bubbles and light scattering. In addition, small patient movements and fluid disturbances reduce scene stability. Moreover, the lack of large-scale real paired datasets also limits algorithm training and evaluation.

To address these challenges, we propose DCP-Net, an unpaired endoscopic image restoration framework specifically designed for robot-assisted spinal surgery. DCP-Net integrates three complementary components: (i) *Diffusion-Prior Contrastive Learning (DPCL)*, which leverages pre-trained *Stable Diffusion (SD)* priors [8] and a bidirectional translation backbone to align latent representations of unpaired clean and degraded surgical images, enabling robust generalization across diverse intraoperative conditions; and (ii) *Physics-Informed Constraints (PIC)*, which incorporate optical priors such as the *Dark Channel Prior (DCP)* [9] to enforce structural fidelity and maintain anatomical consistency; and (iii) *Diffusion-Prior Uncertainty Estimation (DPUE)*, this module leverages the stochastic sampling property of the diffusion model to generate pixel-level uncertainty maps, thereby quantifying the reliability of the restored images and incorporating them as weights into downstream tasks. This synergy allows DCP-Net to suppress intraoperative artifacts—including blood, bubbles, and illumination noise—while preserving critical anatomical details essential for safe robotic manipulation.

We also construct a large-scale spine endoscopy dataset comprising 21,845 paired and unpaired degraded images, with degradation patterns including bleeding, bubbles, and lighting variations. This dataset enables the development of a unified visual enhancement model that addresses the full complexity of spinal endoscopic procedures. Extensive experiments on real surgical data demonstrate that DCP-Net achieves state-of-the-art restoration performance and substantially enhances downstream robotic tasks. Among these, *accurate detection of bleeding points under intraoperative hemorrhage is particularly critical, as it enables the robotic system to precisely navigate to the target location for effective cauterization*. Therefore, we selected this task as a downstream validation to ensure that the model can achieve precise navigation, thereby facilitating safe and efficient hemostasis. By directly enhancing the reliability of perception modules in RMIS, our framework addresses practical clinical challenges, reduces intraoperative risks, and advances the integration of image restoration into intelligent robotic navigation systems. Our main contributions are summarized as follows:

(1) We are the first to construct a large-scale spine endoscopy dataset comprising 21,845 paired and unpaired samples for blood, bubble, and artifact removal, and to explore its applications in robotic navigation tasks for spinal endoscopy. *We plan to release both our dataset and code to the public, aiming to facilitate reproducibility and further research in this area.*

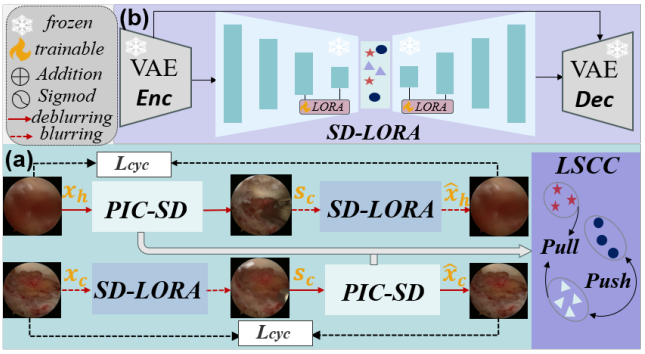

Fig. 2. The overall architecture of our proposed model, where (a) illustrates the backbone network and (b) depicts the SD-LORA module.

(2) We present DCP-Net, a novel unpaired endoscopic image restoration framework for RMIS. Our method integrates DPCL and PIC to achieve robust latent alignment and structural fidelity under diverse intraoperative degradations.

(3) DCP-Net substantially improves critical surgical tasks such as bleeding point detection, achieving a 16.31% increase in mAP with the introduction of DPUE.

(4) Extensive experiments on real surgical data demonstrate that DCP-Net delivers SOTA restoration performance and holds significant clinical value for robotic navigation.

II. METHODOLOGY

To tackle the challenges of unpaired endoscopic image restoration in robot-assisted surgery, we propose a unified framework named DCP-Net. As illustrated in Fig. 1, our model integrates *Diffusion-Prior Contrastive Learning (DPCL)* and *Physics-Informed Constraints (PIC)*, along with auxiliary losses to ensure both semantic accuracy and physical consistency. In addition, we introduce a *Diffusion-Prior Uncertainty Estimation (DPUE)* module that provides pixel-wise confidence maps, improving the trustworthiness of restored images for robotic perception. Our model is built upon CycleGAN and establishes a bidirectional "deblurring–blurring" cycle, enabling unpaired training on real-world data. During training, in the deblurring branch, a real blurred image x_h is first transformed into a synthetic clear image S_c, and then mapped back into a cycle-generated blurred image $\hat{x}_h$, as illustrated in Fig. 2. Conversely, in the blurring branch, the process is reversed.

A. Diffusion-Prior Contrastive Learning

Traditional unpaired image-to-image translation methods employ shallow generators and pixel-wise cycle losses, which often blur fine anatomical details and disrupt semantic correspondence. To overcome these limitations, we propose a DPCL module that unifies high-fidelity generation with feature-level alignment in the latent space.

In this study, we adopt a distilled variant of Stable Diffusion (SD) [10] as the generative backbone,which primarily consists of the VAE encoder-decoder and a U-Net [11]. Two generators are defined: $G_{H\to C}$ translates hazy surgical

frames to clean counterparts, while $G_{C\to H}$ performs the reverse mapping. To fully leverage the latent prior knowledge embedded in SD, we update only the input layer of the backbone and the additional Low-Rank Adaptation (LoRA) adapters, while keeping all other parameters frozen. Specifically, the frozen SD enhanced with LoRA modules (SD-LoRA) injected into the cross-attention layers for efficient task-specific tuning. However, as the VAE encoder progressively downsamples image features and maps them into the latent space, a substantial amount of information is inevitably lost. This leads to a decrease in image fidelity, with noticeable discrepancies from the original image, particularly in local details and textures. To better preserve the fine details of the source image during the dehazing process, we introduce skip connections between the encoder and decoder of the VAE.

During training, the model performs bidirectional translation, and the cycle-consistency loss is defined as:

$$\begin{aligned}\mathcal{L}_{cyc} = \mathbb{E}_{x_h}\left[\|G_{C\to H}(G_{H\to C}(x_h)) - x_h\|_1\right] \\ + \mathbb{E}_{x_c}\left[\|G_{H\to C}(G_{C\to H}(x_c)) - x_c\|_1\right],\end{aligned} \tag{1}$$

which enforces coarse-level structural consistency without requiring paired supervision. However, pixel-level constraints alone cannot guarantee semantic correspondence in complex surgical scenes. To bridge this gap, we introduce a *Latent Space Contrastive Constraint* (LSCC) to regularize feature-level representations. We aim to apply contrastive constraints in the deep feature space to capture the shared characteristics of both the deblurring and reblurring processes, thereby learning feature distributions that are relevant to each domain. In particular, when image patch pairs from different domains have similar features (i.e., "positive samples"), their embeddings in the latent space should be close, whereas dissimilar pairs should be farther apart.

Specifically, we extract intermediate features from the L-th layer of the diffusion backbone, and project them through two-layer MLPs to obtain latent space embeddings. Here, we do not share weights, allowing the model to capture variability across the two domains and learn richer embeddings. To enforce mutual feature constraints in the embedding space, let the query code sampled from one domain be $\hat{f}$, the positive sample code be $\hat{f}^+$, and k negative sample codes from the other domain be $\hat{f}_i^-$, $i = 1, \ldots, k$. The LSCC enforces its constraint by pulling positive samples closer together while pushing negative samples farther apart, thereby providing proper guidance to the generators. Given a feature embedding f, its positive counterpart f^+ from a semantically aligned clean or hazy sample, and a set of negatives $\{f_i^-\}_{i=1}^K$, the contrastive loss is formulated as:

$$\mathcal{L}_{cc} = -\log \frac{\exp(\text{sim}(f, f^+)/\tau)}{\exp(\text{sim}(f, f^+)/\tau) + \sum_{i=1}^{K} \exp(\text{sim}(f, f_i^-)/\tau)}, \tag{2}$$

where sim denotes cosine similarity and τ is the temperature parameter. This formulation encourages anatomical correspondence between features extracted from clean and hazy domains.

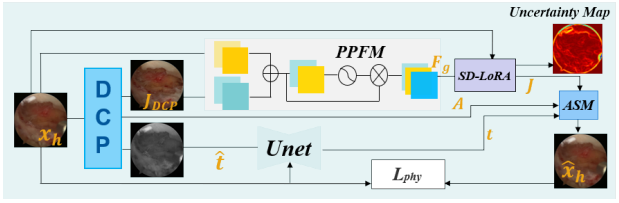

Fig. 3. This figure illustrates our Physics-Informed Constraint (PIC) module and Diffusion-Prior Uncertainty Estimation (DPUE) module. The diffusion module ultimately outputs a confidence map of the denoised image, which can be utilized for downstream tasks.

B. Physics-Informed Constraints (PIC)

While LSCC enhances semantic alignment, it does not guarantee the physical plausibility of restored outputs. Generative models, especially under unpaired settings, often hallucinate unrealistic structures that violate physical imaging principles. To mitigate this, we introduce PIC module based on the Atmospheric Scattering Model (ASM) [12]:

$$I = J \cdot t + A \cdot (1 - t), \tag{3}$$

where I denotes the observed degraded image, J is the underlying clean image, t is the transmission map, and A represents the global atmospheric light.

As shown in Figure 3, we leverage the clear images restored by classic priors Dark Channel Prior (DCP) to extract physics-aware guidance features, which are fused with the original degraded input and provided as conditional signals to the SD model. In addition, we use the DCP-restored results as reference images for reconstructing degraded inputs, encouraging the model to better capture the physical characteristics of degradations in real surgical scenarios. More specifically, We estimate intermediate clean images $\hat{J}$ and transmission maps $\hat{t}$ using DCP.

To effectively incorporate physical priors, we input both the DCP-dehazed image $\hat{J}$ and the original hazy image x_h into the *Physical Prior Fusion Module (PPFM)*. This design not only generates guidance features enriched with physical priors but also compensates for the potential information loss introduced by the DCP dehazing process, thereby providing precise control for Stable Diffusion in surgical video frame dehazing. Specifically, latent features are first extracted from J_{DCP} and x_h using pointwise convolution layers, ReLU activation, and an MLP:

$$F_J, F_x = \text{MLP}\Big(\text{ReLU}\big(\text{Conv}_{1\times1}(J_{\text{DCP}}, x_h)\big)\Big), \tag{4}$$

Then, these features are fused through pointwise convolution, ReLU, Sigmoid, and residual connections:

$$F_{\text{fused}} = F_J + \sigma\Big(\text{ReLU}\big(\text{Conv}_{1\times1}(F_x)\big)\Big), \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid function. Finally, global average pooling, an MLP, and a Softmax operation are applied to concatenate and refine the features, yielding the final guidance feature F_g:

$$F_g = \text{Softmax}\Big(\text{MLP}\big(\text{GAP}(F_{\text{fused}})\big)\Big). \tag{6}$$

Subsequently, the generated guidance features F_g are employed as conditional inputs to Stable Diffusion, steering the deblurring process. By feeding the hazy image into Stable Diffusion, we obtain a realistic deblurred result J. Meanwhile, for the estimated transmission map $\hat{t}$, we concatenate it with the corresponding original hazy image x_h and jointly feed them into the U-Net to obtain the refined transmission map t, thereby enhancing the reliability of the estimation:

$$t = \text{U-Net}\big([\hat{t}, x_h]\big), \tag{7}$$

where x_h denotes the original hazy image and $[\cdot]$ represents the concatenation operation. A pseudo-degraded image $\hat{x}_h$ is then generated via ASM, serving as a self-supervised reference to enforce physical plausibility. To better capture differences in image structure, texture, and perceptual quality, and to ensure that the generated images are closer to the "realistic effect" perceived by the human eye, we introduce the LPIPS loss [13]. The physics-informed loss is defined as:

$$\mathcal{L}_{phy} = \alpha \|I - \hat{I}_{phy}\|_1 + (1-\alpha) \cdot \text{LPIPS}(I, \hat{I}_{phy}), \tag{8}$$

where α balances pixel-wise accuracy and perceptual similarity.

C. Diffusion-Prior Uncertainty Estimation

In clinical applications, the trustworthiness of restored images is as critical as their visual quality, since robotic navigation systems depend on reliable perception for safety-critical tasks. To quantify restoration confidence, we leverage the stochastic sampling property of diffusion priors to estimate pixel-wise uncertainty.

Specifically, for each input x_h, we perform M stochastic reconstructions $\{\hat{J}^m\}_{m=1}^{M}$ using diffusion sampling with injected Gaussian noise. The mean image $\bar{J}$ is taken as the restored output, while the per-pixel variance serves as the uncertainty map:

$$U(p) = \frac{1}{M} \sum_{m=1}^{M} \big(\hat{J}^m(p) - \bar{J}(p)\big)^2, \tag{9}$$

where p indexes pixel locations.

The uncertainty map U is further integrated into downstream robotic tasks (e.g., bleeding point detection, instrument localization) as confidence weights, enabling the perception system to discount low-confidence regions and prioritize reliable visual cues. This mechanism directly improves safety and robustness in robot-assisted surgery.

D. Overall Objective

In the absence of ground truth annotations, we introduce a hybrid loss function to effectively constrain the training of our model, thereby facilitating improved restoration of spinal endoscopic surgical views. This design ensures that the network learns to enhance visual clarity while maintaining structural consistency in the absence of explicit supervision. The final training objective combines all components:

$$\mathcal{L}_{total} = \mathcal{L}_{cyc} + \lambda_{cc}\mathcal{L}_{cc} + \lambda_{phy}\mathcal{L}_{phy}. \tag{10}$$

Here, λ_{cc} and λ_{phy} control the balance of semantic alignment and physical consistency. The weighting factors λ_{cc} and λ_{phy} are both set to 0.5, balancing the contributions of the corresponding loss terms during network training.

III. EXPERIMENTS

A. Dataset Construction

In spinal endoscopic surgery, the surgical field is continuously filled with irrigation fluid, making the construction of paired datasets extremely challenging. For instance, during intraoperative bleeding, blood rapidly diffuses throughout the entire surgical view, which not only increases the risk of operational errors but may also prolong the surgical time. *This rapid dispersion of blood in fluid prevents us from capturing nearly stationary image sequences as is possible in laparoscopic scenarios.* To address this issue, we designed a dataset construction strategy tailored for spinal endoscopy. Specifically, the endoscope was fixed to a robotic arm to maintain a relatively stable view over an extended period. We then selected bleeding and non-bleeding video segments and applied 5× downsampling to eliminate redundant frames. Following method [6], we further performed deformable registration based on the red channel prior for motion correction, thereby constructing a paired dataset for model evaluation.

In total, we retrospectively collected 60 real surgical videos. Among them, 54 cases were utilized to construct an unpaired training dataset, while 6 cases were reserved for building a paired dataset. To further evaluate the effectiveness of our deblurring model in supporting downstream tasks, we annotated bleeding points within the dataset. This enabled us to evaluate whether applying the deblurring model improves the accuracy and efficiency of bleeding point recognition, thereby enhancing surgical safety and supporting intraoperative visual navigation. Ultimately, **we compiled 18,140 unpaired blurry/clear image pairs for Dataset 1 and 3,705 paired images for Dataset 2.**

B. Implementation Details

Our experiments were implemented using Python 3.12 and PyTorch 2.6.0, with training performed on two NVIDIA GeForce RTX A6000 GPUs (48 GB memory each). We employed the Adam optimizer with parameters set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and trained the model for a total of 40 epochs. Specifically, the first 20 epochs were trained from scratch with a fixed learning rate of 0.0001, followed by another 20 epochs where the learning rate was linearly decayed to zero. In the LSCC, each query was associated with 255 internal negative samples and 256 external negative samples, and the temperature parameter was fixed at $\tau = 0.07$. All training images were randomly cropped into 256×256 patches in an unpaired learning manner. To further enhance the effectiveness of contrastive learning, we applied task-specific data augmentation strategies to the benchmark datasets. For the evaluation metrics, we adopted peak signal-to-noise ratio (**PSNR**), structural similarity index (**SSIM**), and the CIE 2000 color difference formula (**CIEDE-2000**).

TABLE I

COMPARISON OF PERFORMANCE WITH 7 STATE-OF-THE-ART MODELS. **BOLD FONT** INDICATES THE BEST PERFORMANCE, "_" INDICATES THE SECOND-BEST PERFORMANCE. OUR ABLATION STUDY RESULTS ARE ALSO PRESENTED.

Type	Method	PSNR↑	SSIM↑	CIEDE-2000↓
Prio-based	DCP	20.98	0.719	6.98
	BCCR	17.72	0.578	7.04
Unpaired	CycleGAN	23.05	0.692	4.33
	RR-GAN	22.36	0.763	3.81
	DerainCycleGAN	23.54	0.801	3.97
	DCD-GAN	24.74	0.812	**2.79**
	CSUD	24.41	0.826	3.36
	w/o SD	24.38	0.796	4.01
	w/o LSCC	25.55	0.842	3.47
	w/o PIC	<u>25.76</u>	<u>0.856</u>	3.31
	Ours	**26.11**	**0.879**	<u>3.04</u>

C. Experiments and Ablation Studies

To validate the effectiveness of the proposed DCP-Net, we conducted experiments on real-world datasets from the following two perspectives: *(1) evaluating the performance of unpaired deblurring on the large-scale unpaired Dataset 1; (2) comparing the differences between unpaired training and paired training using 2,000 pairs of paired data from Dataset 2. The remaining 1,705 pairs in Dataset 2 were used exclusively for testing, and the same test dataset was employed across both experiments.* In Experiment (1), the unpaired model was trained on unpaired datasets and evaluated on the 1,705 paired samples. In Experiment (2), a supervised model was trained exclusively on the paired dataset, while the unpaired model was trained by randomly shuffling the same paired dataset into an unpaired format. Both models were then tested on the same set of 1,705 paired samples.

We conduct a comprehensive evaluation of our proposed method against five unsupervised deep learning baselines, namely CycleGAN [14], RR-GAN [15], DerainCycleGAN [16], DCD-GAN[17], and CSUD [18]. In addition, we include five paired-supervised models—AINDNet [19], DeamNet[20], ScaoedNet [21], Restormer[22] and DiffBIR [23]—together with two prior-based methods: DCP [9] and BCCP [24], as references. To ensure fairness, each model is tested under identical experimental settings, and performance is assessed using three widely adopted image quality metrics: PSNR, SSIM, and CIEDE-2000.

Experiment (1): The quantitative results are summarized in Table I, focusing on comparisons under the unpaired setting. Our method consistently delivers the highest performance in both PSNR and SSIM, demonstrating its superior ability to restore structural details and preserve perceptual quality. Although the CIEDE-2000 score is marginally lower than that of DCD-GAN by 0.25, this gap is relatively small compared to the substantial improvements achieved in the other two metrics, where our approach exceeds DCD-GAN by 1.37 dB in PSNR and 6.7% in SSIM. These results highlight that, even without access to paired supervision,

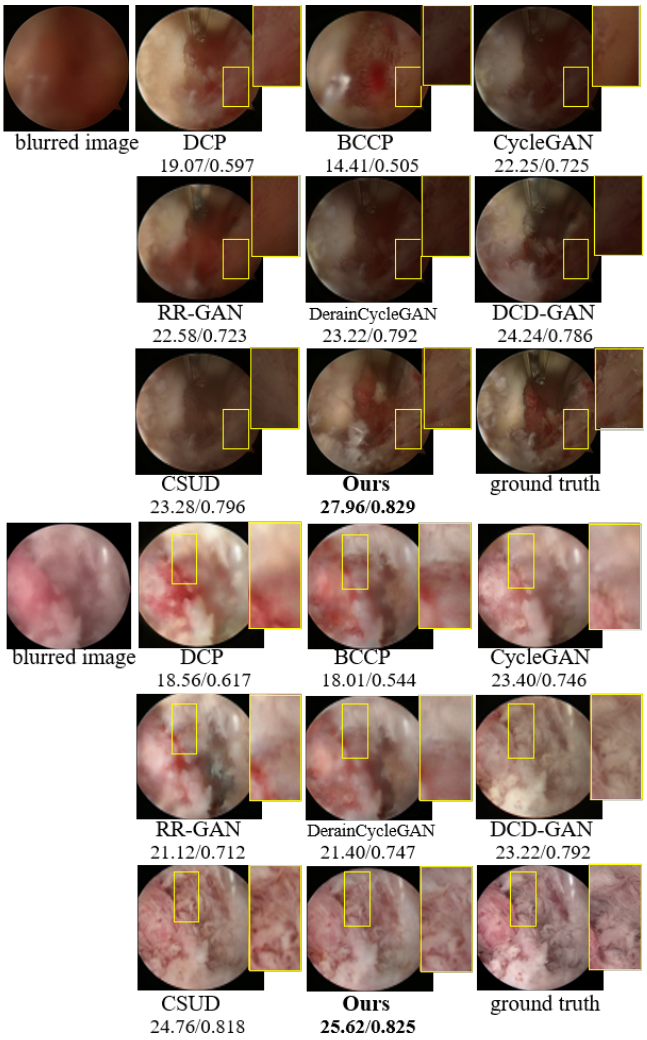

Fig. 4. The results of Experiment 1 are shown in the figure. To better highlight the key anatomical structures, yellow boxes are used for magnified visualization.

our model effectively enhances both fidelity and perceptual similarity, achieving a more balanced performance across different evaluation criteria than existing unpaired methods. Representative visual examples are illustrated in Fig 4, showing that our model restores clearer textures, sharper edges, and more natural color tones compared with competing methods.

To further understand the contribution of individual components, we conduct a series of ablation studies by selectively removing or replacing SD, LSCC, and PIC modules. The ablation results suggest that: SD is critical for global structural fidelity, with its removal causing a noticeable drop of 1.73 dB in PSNR. LSCC improves perceptual consistency, reducing artifacts and boosting SSIM by approximately 3.7%. PIC enhances color and texture alignment, yielding a clear advantage in the CIEDE-2000 metric. These findings highlight the complementary roles of the three modules in driving the overall performance gain.

Experiment (2): In this setting, we directly compared our

TABLE II

THE RESULTS OF EXPERIMENT (2) SHOW A PERFORMANCE COMPARISON OF OUR METHOD AGAINST 10 STATE-OF-THE-ART MODELS, INCLUDING PAIRED-SUPERVISED, AND UNPAIRED MODELS. **BOLD FONT** INDICATES THE BEST PERFORMANCE, WHILE " " DENOTES THE SECOND-BEST PERFORMANCE.

Type	Method	PSNR↑	SSIM↑	CIEDE-2000↓
Paired	AINDNet	21.27	0.657	5.32
	DeamNet	24.21	0.851	3.83
	ScaoedNet	22.52	0.710	4.66
	Restormer	23.95	0.728	4.04
	DiffBIR	**24.81**	0.831	3.76
Unpaired	CycleGAN	23.16	0.726	4.60
	RR-GAN	21.65	0.769	4.73
	DerainCycleGAN	20.23	0.679	6.05
	DCD-GAN	23.05	0.692	3.79
	CSUD	23.17	0.736	4.78
	ours	24.39	**0.863**	**3.55**

TABLE III

QUANTITATIVE RESULTS ON BLEEDING POINT DETECTION IN DOWNSTREAM TASKS.

Method	mAP(%)↑	Recall (%)↑
YOLO	70.14	81.63
YOLO+DCP-Net (**w/o DPUE**)	82.69	89.91
YOLO+DCP-Net (**w DPUE**)	86.45	92.23

method with both paired-supervised and unpaired-trained models using 2,000 pairs of paired data from Dataset 2.Although our model exhibits only a minor difference in PSNR compared to the strongest supervised competitor (24.39 dB vs. 24.81 dB), it clearly outperforms in terms of SSIM (86.3% vs. 83.1%) and CIEDE-2000 (3.55 vs. 3.76), indicating superior perceptual quality and more accurate color reproduction.This is particularly important for clinical applications, where accurate color and structural cues directly influence diagnostic reliability.

Moreover, based on the results of Experiments 1 and 2, when scaled up to a large unpaired training dataset, our method demonstrates remarkable generalization capability, surpassing the best supervised model by a margin of +1.72 dB in PSNR and +1.6% in SSIM, while further reducing the CIEDE-2000 error by 0.51. This result demonstrates the practical value of our approach in real-world clinical scenarios, where collecting high-quality paired data is often infeasible, yet large amounts of unpaired data can be readily obtained. Taken together, the experimental evidence strongly suggests that our method not only achieves state-of-the-art quantitative performance but also ensures high perceptual quality and clinical applicability. By effectively leveraging unpaired training data and introducing well-designed structural and perceptual modules, our framework bridges the gap between supervised and unsupervised paradigms, offering a scalable and clinically valuable solution for medical image restoration.

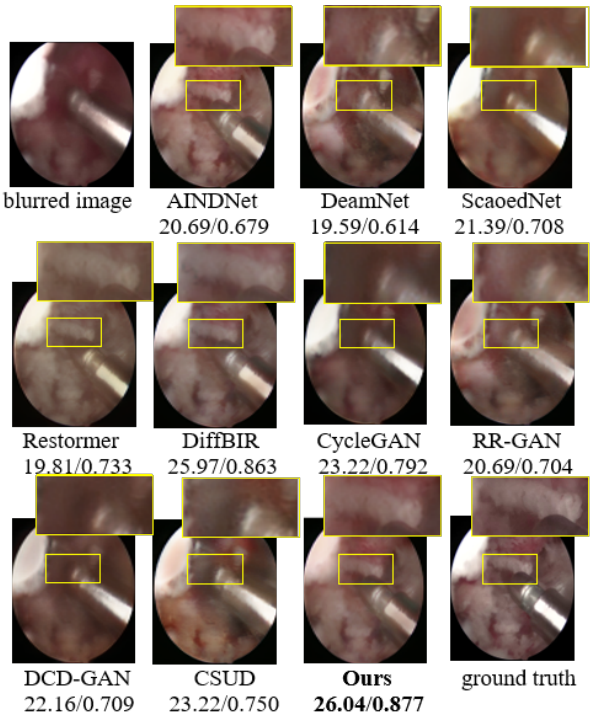

Fig. 5. The results of Experiment 2 are shown in the figure. To better highlight the key anatomical structures, yellow boxes are used for magnified visualization.

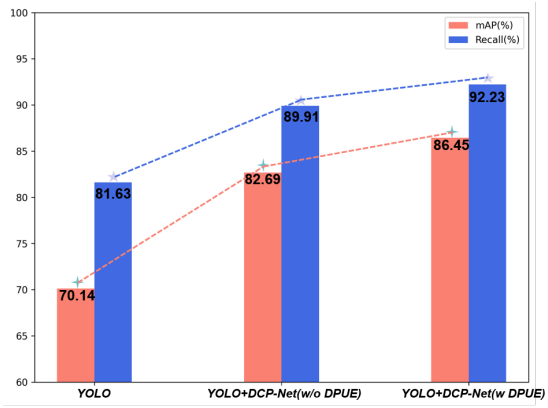

Fig. 6. Comparative results of hemorrhage detection. The proposed model substantially enhances detection performance, yielding a 16.31% increase in mAP compared with the baseline methods.

D. Visual Enhancement Facilitates Bleeding Point Detection

To further validate the clinical significance of our proposed deblurring framework, particularly in scenarios where intraoperative bleeding severely impairs endoscopic visibility, we designed an auxiliary hemorrhage point detection task. Accurate identification of bleeding points is essential in robotic vision navigation, as it enables the system to precisely navigate to the target site for timely cauterization, ensuring both surgical safety and effectiveness. By incorporating this task, we can directly assess whether the deblurred visual feedback improves the robot's ability to perform critical interventions under challenging conditions. In endoscopic

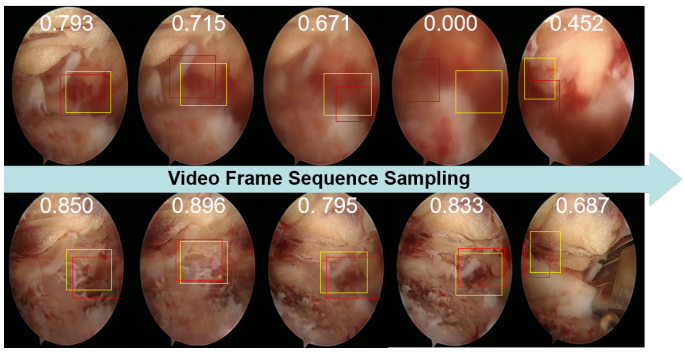

Fig. 7. Comparison of bleeding point detection results on sampled frames from video sequences. The first row shows the detection results on blurred image sequences, while the second row illustrates the results guided by the uncertainty maps generated by our model. Yellow boxes indicate the ground truth, red boxes represent the detected results, and the corresponding IoU values are displayed in white text above each image.

surgery, timely and accurate localization of bleeding sites is critical for enabling immediate electrocauterization, thereby reducing intraoperative risks and ensuring patient safety. However, in practice, the presence of blurred frames caused by bleeding often prevents surgeons from rapidly identifying the hemorrhage source. We employed the deblurred outputs of our model to generate corresponding uncertainty maps, which highlight ambiguous regions caused by motion blur and liquid scattering. These uncertainty maps were fused with the original blurred frames and fed into a YOLO-based [25] detection network, forming the input for training a bleeding-point recognition model. For comparison, we also trained a baseline detection model using only the original blurred frames as input, without deblurring or uncertainty enhancement. Both models were trained under identical conditions ensuring a fair evaluation.

As shown in Table III, our approach achieves superior detection performance across multiple evaluation metrics. Specifically, our method yields a mean Average Precision (mAP) of 86.45% and recall of 92.23%, clearly outperforming the baseline detection model, which achieves only 70.14% mAP and 81.63% recall. To more clearly illustrate the performance improvements of our method, we present the specific gains in Figure 6 using a bar chart. The improvement in recall indicates that our deblurring model helps recover key structural features of bleeding points, reducing the number of missed detections. Representative visualizations are shown in Fig 7, where the baseline model often fails to localize bleeding points under severe blur, leading to either missed detections or false bounding boxes. In contrast, our method consistently produces accurate and stable detection, even in challenging scenarios with heavy bleeding and specular reflections. The uncertainty-guided input allows the detector to better differentiate between actual bleeding points and diffuse blood regions.

These findings strongly support the clinical utility of our framework. By restoring visual clarity and providing more reliable bleeding point detection, our approach enables timely electrocauterization during endoscopic procedures, which is critical for reducing intraoperative complications. Importantly, the ability to leverage blurred frames with deblurring guidance eliminates the need for additional specialized imaging equipment, making our solution both cost-effective and easily integrable into existing surgical workflows [26]. Overall, the results confirm that our deblurring-based detection pipeline not only enhances quantitative detection accuracy but also directly addresses one of the most pressing clinical needs in robot-assisted spinal endoscopy—rapid and precise localization of bleeding sites under impaired visibility.

IV. CONCLUSIONS

This study proposes an unpaired diffusion-prior restoration framework that offers substantial clinical value, particularly in robotic navigation. By effectively removing bleeding and artifacts while preserving structural fidelity, the method significantly enhances surgical field clarity, enabling robotic systems and surgeons to execute critical tasks with greater speed and precision, while reducing cognitive load. In addition, the pixel-wise confidence maps derived from uncertainty estimation provide interpretable visual cues for intraoperative risk assessment and real-time decision-making, directly supporting dynamic adjustments in robotic maneuvers and strengthening autonomous robotic perception. The framework not only reduces reliance on paired data and annotation costs but also presents a novel approach to improving surgical safety and advancing intelligent robotic-assisted systems. Nonetheless, certain limitations remain, as current experiments are primarily conducted on offline datasets; future work will involve large-scale prospective validation in real surgical environments.

REFERENCES

[1] M. Fan, Y. Fang, Q. Zhang, J. Zhao, B. Liu, and W. Tian, "A prospective cohort study of the accuracy and safety of robot-assisted minimally invasive spinal surgery," *BMC surgery*, vol. 22, no. 1, p. 47, 2022.

[2] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Science robotics*, vol. 7, no. 62, p. eabj2908, 2022.

[3] L. Ma, Y. Feng, Y. Zhang, J. Liu, W. Wang, G.-Y. Chen, C. Xu, and Z. Su, "Coa: Towards real image dehazing via compression-and-adaptation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 197–11 206.

[4] J. Pei, D. Guo, D. Yang, Z. Li, Y. Feng, L. Ma, B. Du, and P.-A. Heng, "Benchmarking laparoscopic surgical image restoration and beyond," *arXiv preprint arXiv:2505.19161*, 2025.

[5] C. Yang and C. Liu, "Laparoscopic image desmoking using the u-net with new loss function and integrated differentiable wiener filter," *arXiv preprint arXiv:2505.21634*, 2025.

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

[7] T. Chen, Q. Lyu, L. Bai, E. Guo, H. Gao, X. Yang, H. Ren, and L. Zhou, "Lightdiff: surgical endoscopic image low-light enhancement with t-diffusion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 369–379.

[8] G. Parmar, T. Park, S. Narasimhan, and J.-Y. Zhu, "One-step image translation with text-to-image models," *arXiv preprint arXiv:2403.12036*, 2024.

[9] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[10] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[11] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[12] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 617–624.

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

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[15] H. Zhu, X. Peng, J. T. Zhou, S. Yang, V. Chanderasekh, L. Li, and J.-H. Lim, "Singe image rain removal with unpaired information: A differentiable programming perspective," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9332–9339.

[16] Y. Wei, Z. Zhang, Y. Wang, M. Xu, Y. Yang, S. Yan, and M. Wang, "Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking," *IEEE Transactions on Image Processing*, vol. 30, pp. 4788–4801, 2021.

[17] X. Chen, J. Pan, K. Jiang, Y. Li, Y. Huang, C. Kong, L. Dai, and Z. Fan, "Unpaired deep image deraining using dual contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2017–2026.

[18] G. Dong, T. Zheng, Y. Cao, L. Qing, and C. Ren, "Channel consistency prior and self-reconstruction strategy based unsupervised image deraining," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 7469–7479.

[19] C. Ren, X. He, C. Wang, and Z. Zhao, "Adaptive consistency prior based deep network for image denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8596–8606.

[20] Y. Kim, J. W. Soh, G. Y. Park, and N. I. Cho, "Transfer learning from synthetic to real-noise denoising with adaptive instance normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3482–3492.

[21] C. Ren, Y. Pan, and J. Huang, "Enhanced latent space blind model for real image denoising via alternative optimization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 386–38 399, 2022.

[22] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[23] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, and C. Dong, "Diffbir: Toward blind image restoration with generative diffusion prior," in *European conference on computer vision*. Springer, 2024, pp. 430–448.

[24] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 617–624.

[25] Ultralytics, https://github.com/ultralytics/ultralytics, 2023.

[26] A. Shahi, G. Bajaj, R. GolharSathawane, D. Mendhe, and A. Dogra, "Integrating robot-assisted surgery and ai for improved healthcare outcomes," in *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*. IEEE, 2024, pp. 1–5.