Selective Confidence Fusion for Specular Highlight Detection in Medical Images

Daosong Hu, Zefeng Li✉, Shuai Liu, Mingyue Cui, Rihui Song, Yanchong Xie, and Kai Huang

Sun Yat-sen University, Guangzhou, China
lizf33@mail.sysu.edu.cn

Abstract. Specular highlights frequently occur in endoscopic imaging due to directed illumination and moist tissue surfaces, producing saturated regions that obscure anatomical details and degrade downstream tasks. Accurate highlight detection is therefore essential for reliable medical image analysis and restoration. We propose a confidence-aware multi-cue fusion framework for medical highlight detection. A Selective Confidence Fusion Module adaptively estimates the spatial reliability of structural and chromatic cues, including edge responses, color contrast, background suppression, and semantic consistency. Through per-pixel soft weighting, the model suppresses unreliable signals caused by illumination variation while enhancing anatomically consistent highlight regions, especially along fine boundaries and scattered reflections. Experiments on a public dataset and a clinically collected nasal endoscopy dataset demonstrate consistent improvements over recent methods. Cross-domain evaluation on CVC-ColonDB further verifies strong generalization. In addition, restoration results confirm that more accurate highlight localization leads to cleaner reflection removal and better structural preservation. These findings highlight the importance of confidence-aware cue modulation for robust medical highlight modeling.

Keywords: Highlight detection · Multi-cue feature fusion · Surgical image processing.

1 Introduction

Modern minimally invasive procedures, including endoscopy [8], laparoscopy [21], and ophthalmic microsurgery [1], rely heavily on intraoperative imaging to provide real-time visualization of internal anatomy and surgical instruments. These imaging modalities assist clinicians in localizing lesions, guiding instrument movements, and verifying therapeutic actions, thereby enhancing diagnostic accuracy and procedural safety. Consequently, high-quality visual perception is essential not only for manual surgical operations but also for downstream automated systems [9].

Daosong Hu and Zefeng Li contributed equally to this work.

In practice, intraoperative imaging suffers from adverse illumination and optical effects: the confined, moist environment inside the body forces the use of directed lighting, and smooth tissue or fluid-covered surfaces commonly generate specular reflections. These specular highlights appear as locally saturated bright spots with steep intensity gradients and altered chromatic responses. From an image-formation viewpoint, such effects correspond to a superposition of diffuse tissue reflectance and a specular component, a decomposition formalized in the dichromatic reflection model [13]. Specular highlights impede both human and algorithmic interpretation in multiple ways. Clinically, bright glare can occlude small but critical structures (vessels, nerve bundles, and lesion boundaries), increasing the risk of mis-localization during intervention [12]. For automated systems, highlights distort low-level cues (edges, color distributions, and texture statistics) that many perception algorithms rely on; as a result, segmentation, registration, and restoration modules may misclassify tissues, hallucinate edges, or produce visually unnatural inpainting results. Accurately detecting highlight regions is therefore a fundamental prerequisite for reliable localization, faithful restoration, and robust downstream analysis [23, 24].

Traditional solutions include optical suppression, thresholding-based segmentation and filtering methods [5, 19, 20], as well as decomposition-based methods [2,9,17,18]. These methods are computationally efficient, but they are often fragile under complex illumination and strong tissue textures. Although deep learning has achieved significant progress in specular highlight detection [6,7,16, 22,24]. Nonetheless, several recurring limitations remain: **(i)**many learning-based methods treat all structural or chromatic cues as equally reliable, which leads to over-confident enforcement of spurious edges produced by mirror-like reflections; **(ii)**chromatic priors degrade under local saturation or blood/fluids that alter color statistics; **(iii)**background or coarse semantic cues—if erroneous—can propagate false context through skip-connections and cause incorrect restoration near boundaries; **(iv)**annotated medical datasets remain scarce, which limits generalization and complicates evaluation. These observations motivate a principled fusion strategy that accounts for cue reliability rather than naïvely aggregating all priors.

Based on the above limitations, we propose a confidence-aware multi-cue enhancement framework designed to overcome the limitations above. Instead of hard concatenation of edge, chromatic, background and prediction cues, our Selective Confidence Fusion Module (SCFM) estimates spatially-varying confidence maps for each cue and performs soft (per-pixel) fusion. Concretely, edge priors provide high-frequency spatial information but may be down-weighted where specular reflection produces mirror-like gradients; chromatic priors capture local color deviations useful for distinguishing tissue from highlights but are suppressed when saturation or blood alters color distributions; background priors (from coarse predictions) focus attention away from unlikely foreground but are used conservatively to avoid propagating early errors; finally, prediction-reliability cues indicate semantic certainty and help gate other priors. By combining these signals through learned confidence estimation, softmax normalization

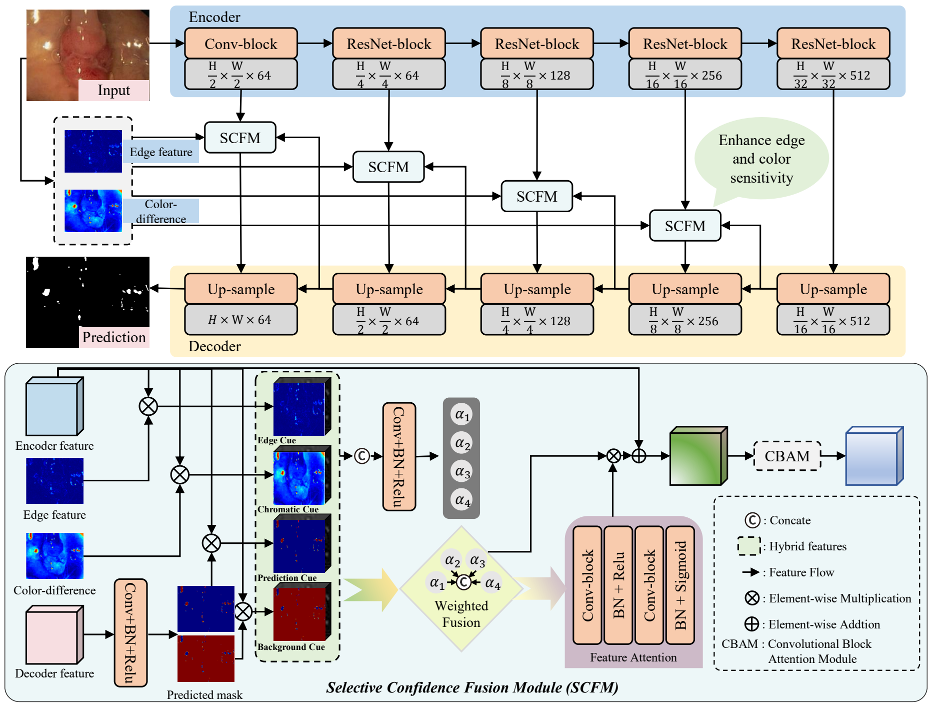

Fig. 1: The overview of our proposed method.

and residual attention reweighting, SCFM selectively reinforces anatomically-consistent highlight regions while avoiding over-constraining the segmentation or restoration stages. We demonstrate that accurate highlight detection produced by this module substantially benefits downstream localization and restoration tasks, yielding crisper boundaries and more faithful texture recovery compared to prior fusion strategies.

2 Method

2.1 Overall Framework

We propose a confidence-aware multi-cue enhancement network for medical highlight detection, as shown in Fig. 1. The framework follows a U-shaped encoder–decoder architecture. The encoder extracts hierarchical semantic representations, while the decoder progressively restores spatial resolution through skip connections.

In conventional U-Net structures, encoder features are directly concatenated with decoder features. However, in medical imaging, specular highlights often introduce unreliable structural and chromatic cues. For example, mirror-like reflections on moist organ surfaces produce sharp intensity transitions that resemble true anatomical boundaries. If edge features are blindly enforced, these reflections may be over-interpreted as structural discontinuities, leading to segmentation errors. Similarly, strong illumination variations distort color distributions, making chromatic contrast unreliable in certain regions.

To address these issues, we introduce a Selective Confidence Fusion Module (SCFM) before each skip connection. Instead of directly merging multiple cues, SCFM estimates the spatial reliability of each cue and performs soft, confidence-guided fusion.

Let E_i denote the encoder feature at scale i, and P_i denote the coarse prediction from the decoder. The refined feature is computed as

$$\widetilde{E}_i = \mathrm{SCFM}(E_i, C_i^e, C_i^c, C_i^b, P_i), \tag{1}$$

where C_i^e, C_i^c and C_i^b represent edge, chromatic, and background cues, respectively.

2.2 Selective Confidence Fusion

Motivation: In endoscopic and surgical images, highlights are primarily caused by specular reflections on moist tissue surfaces. These reflections introduce strong intensity gradients that are not associated with true anatomical boundaries, locally saturated color responses, and ambiguous foreground–background separation. Under such conditions, fixed-weight fusion of structural and chromatic priors may over-constrain the model. For example, edge responses around specular regions can exhibit high gradient magnitude but low anatomical reliability, which may mislead boundary inference. Therefore, the contribution of each cue should be adaptively modulated to reduce unreliable responses and ensure stable highlight localization.

Confidence-aware soft fusion: Given encoder feature $X = E_i$, we construct cue-conditioned features

$$B^{(k)} = X \odot M^{(k)}, \quad k = 1, \ldots, K, \tag{2}$$

where $M^{(k)}$ denotes the k-th cue map.

Instead of directly summing these features, we estimate a spatial confidence score $s^{(k)}$ for each cue:

$$s^{(k)} = \sigma(\phi_k(X, M^{(k)}, P_i)). \tag{3}$$

These scores reflect how consistent each cue is with current semantic predictions. For instance, if a strong edge response appears in a region with low prediction certainty, the model learns to reduce its confidence.

The normalized fusion weights $\alpha^{(k)} \in \mathbb{R}^{H \times W}$ are obtained through softmax: $\alpha^{(k)} = \frac{\exp(s^{(k)})}{\sum_j \exp(s^{(j)})}$. The fused representation is then

$$F = \sum_k \alpha^{(k)} B^{(k)}. \tag{4}$$

This formulation ensures that unreliable cues are automatically down-weighted rather than fully suppressed, preserving flexibility while avoiding hard decisions.

A residual refinement step is applied:

$$X^{\mathrm{enh}} = X + A \odot F, \tag{5}$$

where A is a learned attention map.

2.3 Cue Modeling

Edge cue: Specular highlights are characterized by abrupt intensity transitions. However, such transitions do not necessarily correspond to anatomical boundaries. To extract structural information while suppressing low-frequency illumination effects, we construct a multi-scale Laplacian representation:

$$\mathcal{L}_s = I_{s-1} - U(G_{\sigma_s} * I_s) \tag{6}$$

The edge cue is obtained by aggregating multi-scale responses and normalizing them to $[0, 1]$. This cue enhances potential highlight boundaries but may be unreliable near reflective tissues, motivating confidence reweighting.

Chromatic cue: Highlights often exhibit abnormal color distributions due to saturation and specular reflection. We compute local chromatic deviation in Lab space:

$$C_i^c(x) = \frac{1}{|\Omega_r|} \sum_{y \in \Omega_r(x)} \|I_{\mathrm{Lab}}(x) - I_{\mathrm{Lab}}(y)\|_2. \tag{7}$$

This measure captures local spectral inconsistency. However, in homogeneous tissues with subtle texture variation, chromatic contrast may be weak and less informative. Therefore, its influence should be reduced when confidence is low.

Background cue: The coarse prediction P_i provides semantic localization guidance. The complementary background cue is defined as $C_i^b = 1 - P_i$. This prior helps suppress non-highlight regions but may propagate early-stage prediction errors. Confidence estimation mitigates this risk.

Prediction reliability cue: We further measure prediction certainty as $C_i^p = P_i$. Regions with uncertain prediction (values near 0.5) indicate ambiguity in appearance, often occurring near reflective boundaries. This cue guides the fusion module to reduce over-reliance on unstable priors.

2.4 Loss Function

To supervise highlight detection, we adopt a hybrid segmentation objective that jointly considers pixel-wise classification accuracy and regional overlap consistency. The overall loss is defined as a weighted combination of binary cross-entropy (BCE) and Dice loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \alpha \mathcal{L}_{\mathrm{Dice}} \tag{8}$$

where α balances the contribution of the two terms. In this work, α is set to 1.

3 Experiments

Datasets: We conduct experiments using two datasets for training, including the public CVC-Clinic [3, 11] and our private Nasal-Endo Dataset (NED), and introduce CVC-ColonDB [4] as an external test set to evaluate cross-domain generalization. NED contains 10,000 frames sampled from real-time human nasal endoscopic surgery videos under authentic intraoperative conditions, with an original resolution of 1920×1080. Each frame is annotated by three experienced physicians, where only specular highlight regions are labeled to generate binary masks. For model development, the training datasets are randomly split into 80% training and 20% testing, while CVC-ColonDB is used exclusively for external evaluation.

Implementation details: Our network is implemented in PyTorch and trained on a single NVIDIA GeForce RTX 4070 Ti SUPER GPU. All images and masks are processed with a batch size of 4 for 50 training epochs. We use the Adam optimizer (weight decay 5e-4) with an initial learning rate of 1e-5.

Comparison with existing methods: Table 1 summarizes the quantitative results on CVC-Clinic and the private Nasal-Endo Dataset (NED), which contains fine and scattered reflections from nasal examinations. Our method achieves the best performance on both datasets, reaching 0.848 Dice and 0.758 IoU on CVC-Clinic, and 0.954 Dice and 0.928 IoU on NED, outperforming recent approaches [6, 10, 11, 14, 15]. While most methods detect dominant highlight regions, the improvement of our model mainly comes from more precise boundary delineation and better identification of subtle reflections, which are crucial for structural consistency.

Fig. 2 presents representative visual comparisons. In challenging boundary regions and small reflective spots, our method produces clearer contours and fewer omissions due to confidence-aware cue integration. On CVC-Clinic, where block-wise and point-like highlights often coexist, the dashed boxes show that our model more accurately delineates highlight boundaries while preserving small isolated reflections. On NED, HFA-Net and CM-UNet tend to generate boundary artifacts in narrow anatomical regions, whereas our method maintains stable and consistent predictions. Although the metric gains are moderate, they are important for reliable downstream processing. In addition, our implementation runs at 89.16 FPS (11.22 ms/frame) on our GPU, satisfying real-time clinical requirements.

Ablation analysis: Table 2 reports the effect of each component. Introducing SCFM improves performance over the baseline, demonstrating the benefit of confidence-aware fusion. Incorporating the edge cue further enhances results on both datasets, confirming the importance of structural guidance for boundary delineation and fine reflection modeling. The color cue brings additional gains on CVC-Clinic but shows limited improvement on NED, where narrow anatomy and saturated responses reduce chromatic reliability. The full model achieves the best performance, indicating that adaptive soft fusion effectively integrates complementary cues while avoiding over-reliance on any single prior.

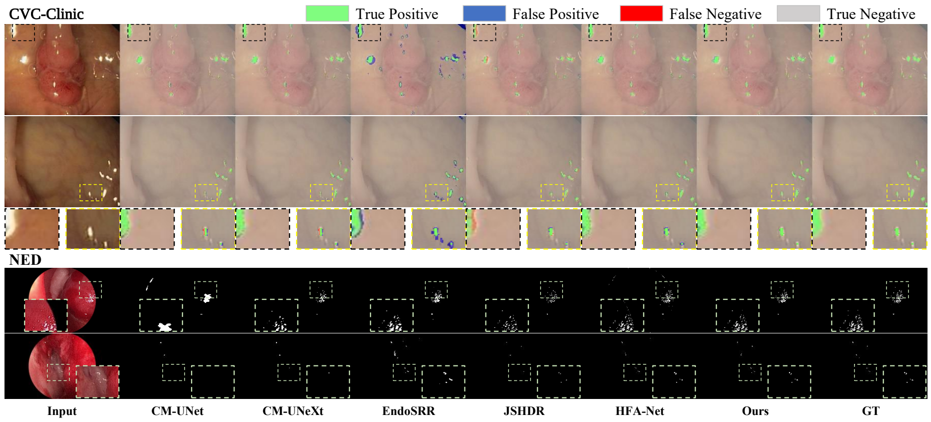

Fig. 2: Visualization results comparing the results with existing methods.

Table 1: Quantitative comparison of highlight detection performance on two datasets.

Method	CVC-Clinic		NED	
	Dice	IoU	Dice	IoU
CM-UNet [15]	0.832	0.733	0.921	0.899
CM-UNeXt [14]	0.829	0.731	0.918	0.902
EndoSRR [10]	0.824	0.700	0.912	0.839
JSHDR [6]	0.823	0.700	0.916	0.844
HFA-Net [11]	0.833	0.729	0.936	0.891
Ours	**0.848**	**0.758**	**0.954**	**0.928**

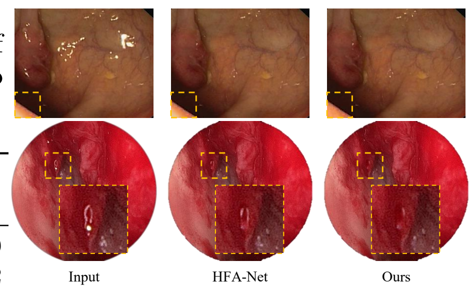

Fig. 3: Highlight restoration results based on predicted masks. The first row shows examples from the CVC dataset, and the second row from the NED.

Impact on highlight restoration: Accurate highlight detection directly affects reflection removal quality. Following HFA-Net [11], we perform restoration using the predicted masks. As shown in Fig. 3, reflections in the lower-left region on CVC are heavily mixed with tissue color, causing incomplete recovery for HFA-Net due to inaccurate localization. Our color-guided modeling improves sensitivity to chromatic deviation and yields more complete restoration. On NED, insufficient boundary modeling in HFA-Net introduces artifacts after inpainting. With enhanced edge guidance and confidence-aware fusion, our method produces cleaner and structurally consistent results, which are critical for reliable downstream analysis.

Cross-domain generalization: Due to the scarcity of annotated medical data, robust generalization across anatomical domains is essential. We therefore train all models on NED and directly evaluate them on CVC-ColonDB, which contains colonoscopic images with markedly different structures and illumination condi-

Table 2: Ablation study on CVC-Clinic and the private Nasal-Endo Dataset (NED). SCFM denotes the Selective Confidence Fusion Module. Edge and Color indicate inclusion of the corresponding cue branches.

Exp ID	SCFM	Edge	Color	CVC-Clinic		NED	
				Dice	IoU	Dice	IoU
A1				0.791	0.677	0.941	0.922
A2	✓			0.814	0.711	0.943	0.913
A3	✓	✓		0.826	0.723	0.951	0.927
A4	✓		✓	0.821	0.723	0.938	0.906
A5	✓	✓	✓	**0.848**	**0.758**	**0.954**	**0.928**

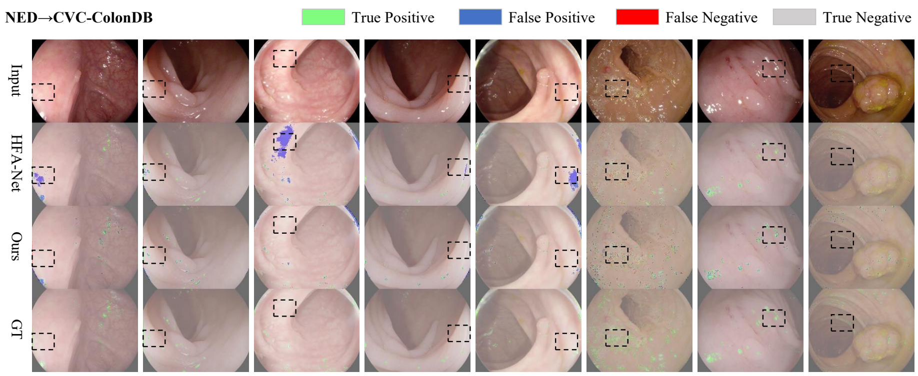

Fig. 4: Cross-domain evaluation from the Nasal-Endo Dataset (NED) to CVC-ColonDB. Models are trained on NED and directly tested on CVC-ColonDB without fine-tuning.

tions. As shown in Fig. 4, the pale mucosal surfaces and widespread reflections in colonoscopy images pose significant challenges. HFA-Net produces false positives on non-specular tissue regions and misses dense, scattered highlights. In contrast, our method maintains stable performance under domain shift. By adaptively balancing structural and chromatic cues through confidence-aware fusion, the proposed framework reduces both over-detection and omission, demonstrating improved cross-domain robustness.

4 Conclusion

In this paper, we present a confidence-aware highlight detection framework for endoscopic imaging. Unlike conventional fusion strategies that treat multiple cues equally, our method introduces a Selective Confidence Fusion Module to adaptively integrate edge, color, background, and semantic guidance within a unified architecture. This design enhances boundary modeling and improves the identification of subtle and scattered reflections under complex illumination conditions. Extensive experiments on public and private datasets demonstrate consistent quantitative gains, improved restoration quality, and stronger cross-domain generalization. These results indicate that precise highlight modeling is essential not only for segmentation accuracy but also for reliable downstream clinical applications.

Acknowledgments. This work was supported by the Guangdong–Hong Kong Joint Funding Program under Grant 2025A0505080012, the Science and Technology Development Fund, Macau SAR under Grant 0010/2023/AKP, and the Shenzhen Medical Research Funds under Grant SZMRF-D2404009.

Disclosure of Interests. The authors have no competing interests

References

1. Alexopoulos, P., Madu, C., Wollstein, G., Schuman, J.S.: The development and clinical application of innovative optical ophthalmic imaging techniques. Frontiers in medicine **9**, 891369 (2022)
2. Bappy, D., Kang, D., Lee, J., Lee, Y., Koo, M., Baek, H.: Specular region detection and covariant feature extraction. In: International Conference on Pattern Recognition. pp. 170–186. Springer (2024)
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
4. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012)
5. El Meslouhi, O., Kardouchi, M., Allali, H., Gadi, T., Benkaddour, Y.A.: Automatic detection and inpainting of specular reflections for colposcopic images. Central European Journal of Computer Science **1**(3), 341–354 (2011)
6. Fu, G., Zhang, Q., Zhu, L., Li, P., Xiao, C.: A multi-task network for joint specular highlight detection and removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7752–7761 (2021)
7. Guo, X., Chen, X., Luo, S., Wang, S., Pun, C.M.: Dual-hybrid attention network for specular highlight removal. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10173–10181 (2024)
8. Kusters, C.H., Jaspers, T.J., Boers, T.G., Jong, M.R., Jukema, J.B., Fockens, K.N., De Groof, A.J., Bergman, J.J., Van Der Sommen, F., De With, P.H.: Will transformers change gastrointestinal endoscopic image analysis? a comparative analysis between cnns and transformers, in terms of performance, robustness and generalization. Medical Image Analysis **99**, 103348 (2025)
9. Li, R., Pan, J., Si, Y., Yan, B., Hu, Y., Qin, H.: Specular reflections removal for endoscopic image sequences with adaptive-rpca decomposition. IEEE transactions on medical imaging **39**(2), 328–340 (2019)
10. Li, W., Jia, F., Liu, W.: Endosrr: a comprehensive multi-stage approach for endoscopic specular reflection removal. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1203–1211 (2024)

11. Li, Z., Cui, M., Hu, D., Gong, J., Weng, J., Zhang, Z., Tian, L., Li, M., Huang, K.: A two-stage method for specular highlight detection and removal in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 23–33. Springer (2025)
12. Nie, C., Xu, C., Li, Z., Chu, L., Hu, Y.: Specular reflections detection and removal for endoscopic images based on brightness classification. Sensors **23**(2), 974 (2023)
13. Shafer, S.A.: Using color to separate reflection components. Color Research & Application **10**(4), 210–218 (1985)
14. Tang, F., Ding, J., Quan, Q., Wang, L., Ning, C., Zhou, S.K.: Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
15. Tang, F., Wang, L., Ning, C., Xian, M., Ding, J.: Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In: 2023 IEEE 20th international symposium on biomedical imaging (ISBI). pp. 1–5. IEEE (2023)
16. Wu, Z., Guo, J., Zhuang, C., Xiao, J., Yan, D.M., Zhang, X.: Joint specular highlight detection and removal in single images via unet-transformer. Computational visual media **9**(1), 141–154 (2023)
17. Xia, W.: Fast high quality highlight removal using a simplified matrix iteration method. Signal Processing **226**, 109641 (2025)
18. Xia, W., Chen, E.C., Pautler, S.E., Peters, T.M.: A global optimization method for specular highlight removal from a single image. IEEE Access **7**, 125976–125990 (2019)
19. Yang, Q., Wang, S., Ahuja, N.: Real-time specular highlight removal using bilateral filtering. In: European conference on computer vision. pp. 87–100. Springer (2010)
20. Yu, B., Chen, W., Zhong, Q., Zhang, H.: Specular highlight detection based on color distribution for endoscopic images. Frontiers in Physics **8**, 616930 (2021)
21. Zaidi, S.A., Chouvatut, V., Phongnarisorn, C., Praserttitipong, D.: Deep learning based detection of endometriosis lesions in laparoscopic images with 5-fold cross-validation. Intelligence-Based Medicine **11**, 100230 (2025)
22. Zhang, C., Liu, Y., Wang, K., Tian, J.: Specular highlight removal for endoscopic images using partial attention network. Physics in Medicine & Biology **68**(22), 225009 (2023)
23. Zhang, J., Ji, Z., Zhao, C., Huang, M., Li, M., Zhang, H.: Sdmffn: a novel specular detection median filtering fusion network for specular reflection removal in endoscopic images. Biomedical Physics & Engineering Express **12**(1), 015036 (2026)
24. Zhu, D., Huang, G., Yuan, X., Chen, X., Zhong, G., Pun, C.M., Deng, J.: Faqnet: Frequency-aware quaternion network for endoscopic highlight removal. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1408–1413. IEEE (2024)