Multi-frame Restoration for 10 Hz Lissajous Confocal Laser Endomicroscopy

Minhee Lee[1], Sangyoon Lee[1], Jiwook Lee[2], Minki Hong[2], Kyuyoung Kim[2], Won Hwa Kim[1], and Jaeho Lee[1]

[1] POSTECH, Pohang, Republic of Korea
[2] VPIX Medical, Seoul, Republic of Korea
{mhlee02,sangyoon.lee,wonhwa,jaeho.lee}@postech.ac.kr

Abstract. Lissajous confocal laser endomicroscopy (CLE) is a promising solution for high-speed in vivo optical biopsy for handheld scenarios. However, Lissajous scanning traces a resonant trajectory and samples only the visited pixels per frame; at high frame rates, many pixels remain unvisited, creating structured holes. In this work, we introduce the first benchmark for 10 Hz Lissajous CLE, consisting of low-quality video clips paired with high-quality reference images. The reference images are wide-FOV mosaics obtained by stitching stabilized, slow-scan frames of the same tissue, enabling temporally aligned supervision. Using this dataset, we propose MIRA, a lightweight recurrent framework for Lissajous CLE restoration that iteratively aggregates temporal context through feature reuse and displacement alignment. Our experiments demonstrate that MIRA outperforms both lightweight and high-complexity baselines in restoration quality while maintaining a favorable computational efficiency suitable for clinical deployment.

Keywords: Confocal laser endomicroscopy · Lissajous CLE restoration

1 Introduction

Confocal laser endomicroscopy (CLE) enables *in vivo* imaging of tissue microstructures at cellular resolution, offering the potential for real-time "optical biopsy" during endoscopy [3,26]. However, since CLE forms images via sequential laser scanning over a microscopic field-of-view (FOV), it is inherently sensitive to probe-tissue motion. Even subtle motion can cause severe distortion (i.e., *motion artifacts*), necessitating robust, high-speed imaging for clinical use [1,13,16,23].

Increasing the frame rate is a natural way to reduce motion artifacts by shortening the acquisition window. While existing fiber-bundle CLE systems provide video-rate imaging, their spatial fidelity is fundamentally constrained by the finite number of fiber cores and inter-core honeycomb gaps. Single-fiber raster systems eliminate bundle artifacts, but their frame rate remains constrained by one-axis asymmetric scanning and sequential line sampling. In contrast, Lissajous scanning uses dual resonant motion along orthogonal axes, decoupling frame-rate scaling from per-axis line count and enabling smoother, inertia-efficient

acquisition. This work therefore focuses on 10 Hz Lissajous CLE, following its clinical relevance in prior handheld Lissajous CLE systems [6,7,8].

Despite its speed advantage, 10 Hz Lissajous CLE produces sparse, nonuniform pixel measurements, resulting in severe undersampling artifacts (called "holes") and degraded quality. Simply aggregating multiple frames can improve spatial coverage, but risks reintroducing motion artifacts and misalignments [13].

This setting differs from conventional video restoration methods [2,15,5] in its alignment requirement. First, dense correspondence methods, including optical flow [21,25], become unreliable because each Lissajous frame contains substantial missing pixels. Second, unlike natural videos with complex local motion, handheld CLE motion is dominated by global probe–tissue displacement, so most structures share a consistent inter-frame shift. These properties motivate MIRA to use lightweight flow-free global registration rather than dense local alignment.

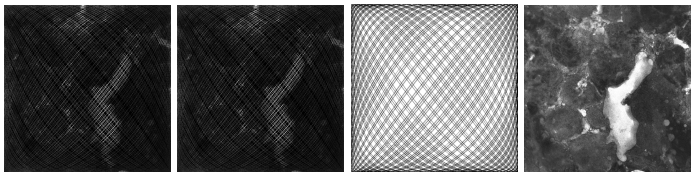
To address this challenge, we propose *learning* to restore high-quality frames from 10 Hz Lissajous CLE. We construct *the first open benchmark Lissajous CLE dataset*, **MaLissa** (Matched Lissajous CLE), comprising over 2,000 paired 10 Hz low-quality (LQ) and low-rate high-quality (HQ) frames collected from animal tissues. Since acquiring perfectly aligned LQ-HQ pairs is difficult, we propose a matching-based strategy: we first stitch HQ frames into a large mosaic, then extract for each LQ frame the best-matching HQ subimage as its target.

Using MaLissa, we propose **MIRA** (Multi-frame Iterative RestorAtion), a lightweight framework to integrate the current LQ frame with past frames to estimate the HQ output. To handle sparse pixels while ensuring low latency, MIRA adopts two key strategies: (1) caching and reusing encoded features across timesteps to avoid redundant computations; (2) aligning spatially mismatched features using a lightweight flow-free registration module, which estimates inter-frame displacement from sparse pixels. MIRA achieves a latency–quality trade-off (22.88 dB PSNR at 236.66 ms end-to-end latency), outperforming baselines [2,15,20] and providing a practical basis for pipelined 10 Hz deployment.

Contribution. (i) We introduce the task of 10 Hz Lissajous CLE restoration and MaLissa, the first open dataset in this domain. (ii) We propose MIRA, a lightweight recurrent restoration framework for Lissajous CLE. (iii) MIRA achieves the best restoration quality among evaluated baselines with sub-250 ms end-to-end latency. The MaLissa dataset is available at: `https://huggingface.co/datasets/2mm/MaLissa`

2 Task: 10 Hz Lissajous CLE Restoration

The task of 10 Hz Lissajous CLE restoration can be formalized as follows: Let $\mathcal{I} = \{I_t\}_{t=1}^{T}$ denote the input frame sequence collected with 10 Hz Lissajous CLE, where each $I_t \in \mathbb{R}^{H \times W}$ is a grayscale LQ frame with sparse, motion-distorted pixels; noisy pixel measurements are acquired along a Lissajous trajectory, and other pixels have the value 0. Our goal is to reconstruct a HQ frame sequence $\mathcal{O} = \{O_t\}_{t=1}^{T}$ from $\mathcal{I}$, where each $O_t \in \mathbb{R}^{H \times W}$ should correspond to the same scene as I_t, but with dense measurements and less motion artifacts.



(a) (b) (c) (d)

Fig. 1: A visual description of the 10 Hz Lissajous CLE frames: (a) 10 Hz LQ frame with steady probe movement; (b) 10 Hz LQ frame of the same scene, with shaky probe movement; (c) Scanning trajectory of the probe while taking one frame, where the white "holes" correspond to the missing pixels; (d) The low-rate HQ frame with the same field-of-view.

At high frame rate, each LQ frame contains substantial missing pixels, making single-frame restoration ill-posed. Thus, we consider *multi-frame* restoration: Given the frame window $I_{t-\tau}, \ldots, I_t$, we estimate

$$\hat{O}_t = f(I_{t-\tau}, \ldots, I_{t-1}, I_t) \tag{1}$$

such that this estimate approximates the output frames well, *i.e.,* $O_t \approx \hat{O}_t$.

3 The MaLissa Dataset

To train and evaluate our model, we introduce an open Lissajous CLE restoration dataset, termed MaLissa (Matched Lissajous CLE), consisting of paired LQ and HQ frames (I_t, O_t) that share the same scene. We split the dataset at the clip level: each raw clip was acquired from a distinct, non-overlapping spatial location on the tissue, preventing spatial leakage between the training and validation sets. MaLissa contains 2,060 curated pairs (1,619 for training and 441 for validation), selected from 16 raw video clips (3,988 frames total). All videos were acquired at 1024×1024 resolution using the cCeLL Lissajous-scanning CLE system [6], with a $500 \times 500\,\mu$m FOV. The data were collected from porcine stomach tissue. Although non-pathological, these tissues closely resemble human gastrointestinal tract anatomically and physiologically, making them a common choice for open-to-public datasets [19,24].

To emulate clinical use, LQ frames were captured at 10 Hz, resulting in more than 70% missing pixels. During acquisition, the probe was moved irregularly in random directions to mimic a clinician's motion. HQ frames were captured at 2 Hz, leaving fewer than 10% pixels unmeasured. The missing pixels were filled via bilinear interpolation to obtain dense images. The probe followed an outward spiral trajectory with small overlaps to cover the entire sample.

3.1 Matching LQ and HQ frames

As the raw LQ and HQ videos are acquired along different probe trajectories, their frames do not share the same FOV, making direct frame-wise matching

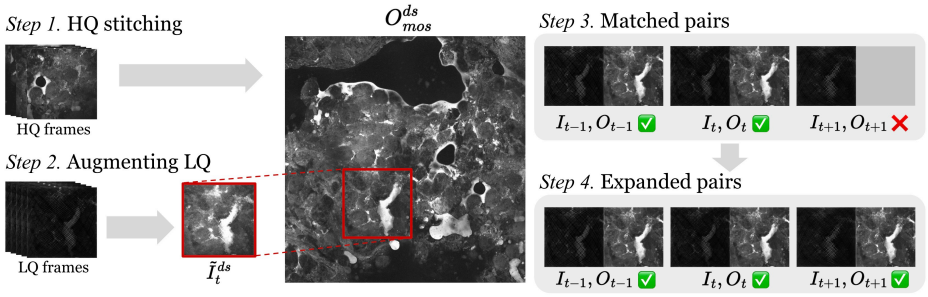


Fig. 2: Overview of dataset construction procedure for the MaLissa dataset.

infeasible. Moreover, LQ frames contain substantial missing pixels, and asynchronous acquisition introduces local intensity inconsistencies. To obtain paired frames (I_t, O_t) with identical FOV, we construct an HQ mosaic by stitching HQ frames and match each LQ frame to a corresponding subimage of this mosaic. Our method consists of four steps:

Step 1: HQ stitching. We select 25 HQ frames with approximately 20% spatial overlap and stitch them to generate the HQ mosaic O_{mos}. To maintain intensity continuity while suppressing edge artifacts, all frames are mapped to a unified coordinate system and blended using a Hann window.

Step 2: Augmenting LQ frames. As LQ frames contain many missing pixels, direct comparison with HQ subimages is unreliable. We therefore construct *augmented* LQ frames $\tilde{\mathcal{I}} = \{\tilde{I}_t\}_{t=1}^{T}$, by aggregating information from neighboring frames. For each I_t, we estimate the translations to the past $(I_{t-4}, \ldots, I_{t-1})$ and future $(I_{t+1}, \ldots, I_{t+4})$ frames using FFT-based phase correlation [27]:

$$\mathbf{r}(I, I') = \mathcal{F}^{-1}\left(\texttt{norm}\left(\mathcal{F}(I) \odot \overline{\mathcal{F}(I')}\right)\right), \tag{2}$$

where $\mathcal{F}(\cdot)$ denotes the 2D FFT, $\overline{(\cdot)}$ the complex conjugate, $\odot$ the Hadamard product, and `norm` the elementwise normalization for complex values. The translation maximizing this correlation is selected to align neighboring frames with I_t. The aligned frames are then averaged with I_t, using only measured pixels while ignoring unobserved locations, yielding the augmented frame $\tilde{I}_t$.

Step 3: Matching via phase correlation. We match augmented LQ frames $\tilde{I}_t$ to cropped subimages of the HQ mosaic O_{mos} using FFT-based phase correlation. To mitigate the effect of missing pixels, we apply max-pooling downsampling to both $\tilde{I}_t$ and O_{mos}, and compute the correlation $r(\tilde{I}^{\text{ds}}_t, O^{\text{ds}}_{\text{mos}})$. Pairs with correlation above a threshold (0.05 for the MaLissa dataset) are retained. If a match is found, the corresponding subimage is assigned to the HQ label O_t for I_t.

Step 4: Temporal expansion. The pairs (I_t, O_t) identified in Step 3 guide further matching, as temporally adjacent LQ frames are likely to correspond to nearby regions in O_{mos}. For each (I_t, O_t), we search for subimages that align with I_{t-1} and I_{t+1}. Since unmatched frames often exhibit motion artifacts that degrade structural cues and hinder phase correlation, we instead employ tem-

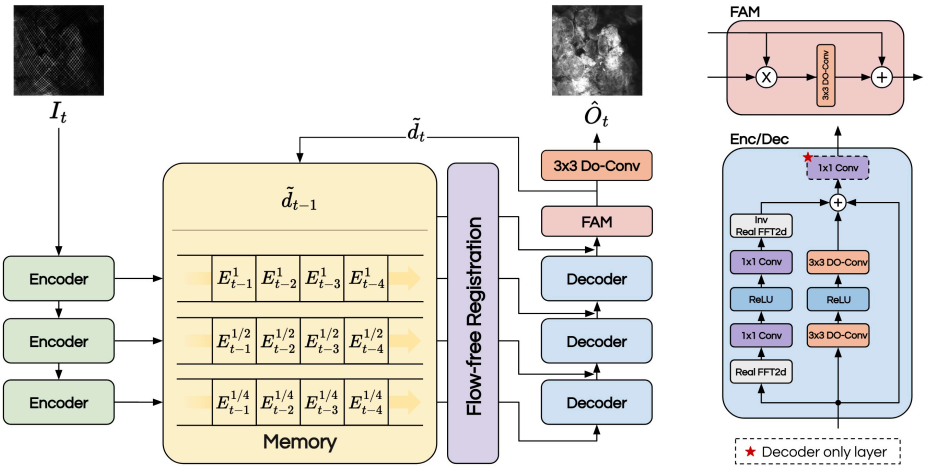


Fig. 3: Overview of the MIRA architecture. Encoder features from previous frames $E_{t-4}, \ldots, E_{t-1}$ and the previous FAM output $\tilde{d}_{t-1}$ are cached and aligned to the current frame via global registration.

plate matching in this step. The process is propagated to neighboring frames until no additional matches are found.

4 Method: Multi-frame Iterative Restoration (MIRA)

MIRA adopts a lightweight U-Net architecture [22] for low-latency processing (Fig. 3). Both encoder and decoder blocks use ResFFT modules [17], which combine frequency and spatial branches to suppress motion artifacts that are often separable in the frequency domain [9]. At timestep t, the encoder extracts multiscale features E_t^l from I_t and stores them in a memory bank with a window size of 4. Each decoder block aggregates current features with cached features of $\{E_{t-i}^l\}_{i=0}^4$ to compensate for pixel sparsity. Concretely, each block conducts

$$E_t^{l/2} = \mathrm{enc}\big(E_t^l\big), \qquad D_t^l = \mathrm{dec}\big(D_t^{l/2}, E_t^l, \{\mathrm{shift}(E_{t-i}^l, \Delta\mathbf{s}_{t,t-i})\}_{i=1}^4\big), \tag{3}$$

where D_t^l denote the decoder features at step t and scale l, and $\mathrm{shift}(\cdot)$ denotes the registration (described below). A 1×1 convolution is added at each decoder block to enhance the representational capacity. The decoded features are further refined by a lightweight feature attention module (FAM) [4,17], which recursively incorporates the previous FAM output $\hat{d}_{t-1}$ to stabilize high-frequency details.

Flow-free registration of cached features. Before aggregation, the cached features $\{E_{t-i}^l\}_{i=1}^4, \tilde{d}_{t-1}$ are aligned to the current frame I_t. This is done by estimating the global inter-frame displacement via phase correlation:

$$\Delta\mathbf{s}_{t,t'} = \mathrm{argmax}_{(x,y)}[\mathbf{r}(\tilde{I}_t^{ds}, \tilde{I}_{t'}^{ds})]_{x,y} \tag{4}$$

Table 1: Quality and efficiency of various video restoration models, with the best in **bold** and the runner-up in underlined. We use the patch size 256, 5 frame window. The latency is measured in ms, on an NVIDIA RTX 6000 Ada.

Model	Backbone	PSNR	SSIM	MS-SSIM	Latency	GFLOPs	Params
MedVSR [12]	SSM	9.95	0.0597	0.2747	290.03	1428.92	7.12M
Turtle [5]	Transformer	21.81	0.5957	0.7037	1668.69	9280.94	59.08M
RVRT [11]	Transformer	21.45	0.5885	0.6948	491.77	1363.74	13.57M
BasicVSR++ [2]	CNN	20.89	0.5813	0.6876	333.19	1968.80	7.39M
EMVD [15]	CNN	18.71	0.5441	0.6349	**165.63**	**10.41**	**0.02M**
MIRA (Ours)	CNN	**22.88**	**0.6008**	**0.7197**	236.66	16.57	7.29M

This strategy avoids registration methods such as optical flow which are sensitive to missing pixels and computationally expensive, and exploits the contact-constrained nature of CLE, where motion is predominantly lateral [10,18].

4.1 Training with patch-wise rejection sampling

MIRA is trained with the joint loss

$$\ell(O_t, \hat{O}_t) = \ell_{\text{char}}(O_t, \hat{O}_t) + \lambda \cdot \|\mathcal{F}(O_t) - \mathcal{F}(\hat{O}_t)\|_1, \tag{5}$$

where the Charbonnier loss ℓ_{char} ensures spatial robustness and the latter loss enhances structural details [4,17], which are balanced by $\lambda = 0.01$. During training, we randomly crop 256×256 patches from paired LQ-HQ frames.

As phase-correlation-based alignment can introduce brightness mismatches that destabilize the training, we apply *patch-wise rejection sampling* to filter out problematic crops: For each crop, we compute a binary inconsistency mask between the HQ patch and the (downsampled) augmented LQ frame $\tilde{I}_t^{ds}$, which aggregates aligned neighboring frames to reduce missing pixels. The patch is divided into 8×8 blocks, and block-wise MSE is evaluated between LQ and HQ blocks. Blocks above a threshold are marked inconsistent. The patch is rejected and resampled if inconsistent regions exceed 12.5% of the area.

5 Experiments & Results

5.1 Experimental setup

MIRA was trained with Adam using a step size of 2×10^{-4} and a batch size of 32. We trained for 600 epochs with cosine annealing. Baselines were trained using the hyperparameters in their original papers. We also applied random cropping, patch-wise rejection, and temporal reversal augmentation to all methods.

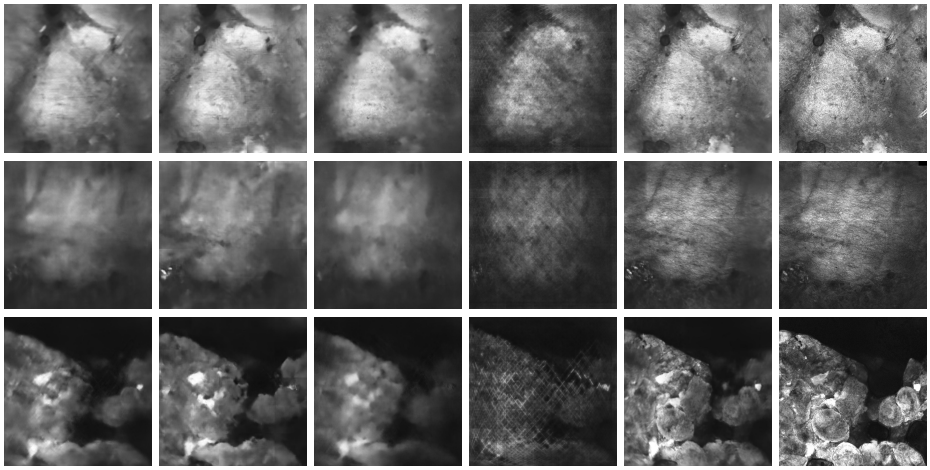


Fig. 4: Qualitative comparison of LQ frames restored by MIRA and the baselines.

Table 2: Ablation of key technical components of MIRA. "Registration" denotes the flow-free registration, and "Rej. sampling" denotes the patch-wise rejection sampling.

Variant	PSNR↑	SSIM↑	MS-SSIM↑
w/o Registration	22.87	0.6006	0.7093
w/o FAM	**22.90**	0.6000	0.7083
w/o rej. sampling	22.70	0.5966	0.7050
MIRA	22.88	**0.6008**	**0.7197**

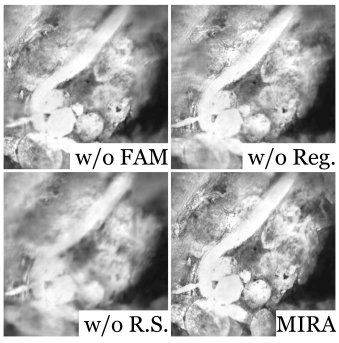


Fig. 5: Qualitative comparison of ablations.

5.2 Main results on image restoration

We compare MIRA with state-of-the-art video restoration models on the MaLissa validation split (Table 1). MIRA achieves the highest restoration quality while maintaining substantially lower latency than transformer-based methods. Compared with convolutional approaches, MIRA delivers markedly better reconstruction with only a modest increase in computation. Overall, MIRA offers a favorable accuracy-efficiency tradeoff, bridging the gap between heavy and lightweight models. Qualitatively, MIRA recovers sharper structures and preserves edge better than BasicVSR++ and Turtle, which tend to over-smooth details (Fig. 4).

Table 3: Downstream task performances. Acc@1 (Full) and segmentation metrics (95% upper-trimmed mean) are reported. Bold indicates the best performance.

	Retrieval	Segmentation		
Model	Acc@1 ↑	Dice ↑	HD95 ↓	BFScore ↑
BasicVSR++	59.4	0.8670 ± 0.0953	16.3700 ± 7.9237	0.2849 ± 0.1614
Turtle	60.6	0.8683 ± 0.0958	15.9190 ± 7.0681	0.2752 ± 0.1452
MIRA (Ours)	**70.7**	**0.8688 ± 0.1004**	**15.2473 ± 7.2288**	**0.2919 ± 0.1711**

5.3 Ablation studies

We conduct an ablation study for each technical component of MIRA in Table 2. Removing the rejection sampling leads to a substantial quality decrease in all metrics. Removing either registration or FAM does not noticeably decrease the PSNR or SSIM, but has visible impacts on MS-SSIM; indeed, the removal does degrade the perceptual quality of the images, as shown in Fig. 5.

5.4 Downstream applications

To evaluate predictive quality from restored frames, we consider two downstream tasks: 1) sample region retrieval and 2) semantic segmentation.

Retrieval. We evaluate the task of retrieving the correct HQ mosaic given a restored LQ frame as the query; see Table 3. MIRA achieves 70.7% accuracy, outperforming BasicVSR++ and Turtle by 11.3%p and 10.1%p, respectively.

Segmentation. We assess structural consistency via segmentation, measuring the overlap between segmentation maps from restored LQ and their corresponding HQ frames. Since MaLissa lacks clinician-annotated masks, we construct pseudo-GT masks on HQ images by prompting MedSAM [14] with human-annotated bounding boxes on visible objects. The same pipeline is applied to restored LQ frames. We report Dice, HD95, and BFScore, using the 95% upper-trimmed mean to mitigate the influence of a small number of outlier frames with degenerate predictions. MIRA outperforms the baselines across all metrics.

Discussion. The results show that MIRA's advantage goes beyond pixel-level fidelity to improved reconstruction of structural properties. Such improvements may facilitate more reliable clinical assessment during endoscopic procedures.

6 Conclusion

In this work, we introduce the task of 10 Hz Lissajous CLE restoration and present MaLissa, the first publicly available dataset constructed via a matching-based strategy to address field-of-view inconsistencies. We also propose MIRA, a lightweight U-Net-based recurrent framework that effectively aligns and aggregates sparse low-quality frames. Trained on MaLissa, MIRA outperforms existing video restoration methods while maintaining low latency.

Future work. Despite these advances, latency remains a bottleneck for strict real-time deployment. We clarify that our reported latency measures non-pipelined end-to-end processing rather than streaming throughput. In a fully integrated system, pipelining the encoding and decoding of consecutive frames can reduce the inter-frame processing gap toward the 100 ms requirement. In addition, hardware-efficient optimizations, such as model compression and lower-precision quantization (e.g., FP8/FP16), provide promising directions for further acceleration while preserving restoration performance. Finally, physical resolution validation using standard targets such as USAF charts would further complement the current pixel-level metrics and downstream functional evaluations.

Acknowledgements. This work was supported by the Technology Development Program funded by the Ministry of SMEs and Startups (MSS, Korea) (No. RS-2024-00487593), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00457882, No. RS-2019-II191906), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00453301, No. RS-2026-25494004), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2025-25422953), and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health and Welfare, Republic of Korea (No. RS-2025-23872969).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Carbone, F., Fochi, N.P., Di Perna, G., Wagner, A., Schlegel, J., Ranieri, E., Spetzger, U., Armocida, D., Cofano, F., Garbossa, D., et al.: Confocal laser endomicroscopy: enhancing intraoperative decision making in neurosurgery. Diagnostics (2025)
2. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
3. Chauhan, S.S., Dayyeh, B.K.A., Bhat, Y.M., Gottlieb, K.T., Hwang, J.H., Komanduri, S., Konda, V., Lo, S.K., Manfredi, M.A., Maple, J.T., et al.: Confocal laser endomicroscopy. Gastrointestinal Endoscopy (2014)
4. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
5. Ghasemabadi, A., Janjua, M.K., Salameh, M., Niu, D.: Learning truncated causal history model for video restoration. In: Advances in Neural Information Processing Systems (2024)
6. Hwang, K., Seo, Y.H., Ahn, J., Kim, P., Jeong, K.H.: Frequency selection rule for high definition and high frame rate Lissajous scanning. Scientific Reports (2017)

7. Hwang, K., Seo, Y.H., Kim, D.Y., Ahn, J., Lee, S., Han, K.H., Lee, K.H., Jon, S., Kim, P., Yu, K.E., et al.: Handheld endomicroscope using a fiber-optic harmonograph enables real-time and in vivo confocal imaging of living cell morphology and capillary perfusion. Microsystems & Nanoengineering (2020)
8. Jeon, J., Kim, H., Jang, H., Hwang, K., Kim, K., Park, Y.G., Jeong, K.H.: Handheld laser scanning microscope catheter for real-time and in vivo confocal microscopy using a high definition high frame rate Lissajous mems mirror. Biomedical Optics Express (2022)
9. Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
10. Latt, W.T., Newton, R.C., Visentini-Scarzanella, M., Payne, C.J., Noonan, D.P., Shang, J., Yang, G.Z.: A hand-held instrument to maintain steady tissue contact during probe-based confocal laser endomicroscopy. IEEE Transactions on Biomedical Engineering (2011)
11. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. In: Advances in Neural Information Processing Systems (2022)
12. Liu, X., Sun, G., Wang, C., Yuan, Y., Konukoglu, E.: MedVSR: Medical video super-resolution with cross state-space propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
13. Loewke, N.O., Qiu, Z., Mandella, M.J., Ertsey, R., Loewke, A., Gunaydin, L.A., Rosenthal, E.L., Contag, C.H., Solgaard, O.: Software-based phase control, video-rate imaging, and real-time mosaicing with a Lissajous-scanned confocal microscope. IEEE Transactions on Medical Imaging (2019)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications (2024)
15. Maggioni, M., Huang, Y., Li, C., Xiao, S., Fu, Z., Song, F.: Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
16. Mahé, J., Linard, N., Tafreshi, M.K., Vercauteren, T., Ayache, N., Lacombe, F., Cuingnet, R.: Motion-aware mosaicing for confocal laser endomicroscopy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2015)
17. Mao, X., Liu, Y., Liu, F., Li, Q., Shen, W., Wang, Y.: Intriguing findings of frequency selection for image deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
18. Marques, M.J., Hughes, M.R., Vyas, K., Thrapp, A., Zhang, H., Bradu, A., Gelikonov, G., Giataganas, P., Payne, C.J., Yang, G.Z., et al.: En-face optical coherence tomography/fluorescence endomicroscopy for minimally invasive imaging using a robotic scanner. Journal of Biomedical Optics (2019)
19. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., Sahin, H., Araujo, H., Alexandrino, H., Durr, N.J., Gilbert, H.B., Turan, M.: EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical Image Analysis (2021)
20. Qi, C., Chen, J., Yang, X., Chen, Q.: Real-time streaming video denoising with bidirectional buffers. In: Proceedings of the ACM International Conference on Multimedia (2022)

21. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2015)
23. Stoeve, M., Aubreville, M., Oetter, N., Knipfer, C., Neumann, H., Stelzle, F., Maier, A.: Motion artifact detection in confocal laser endomicroscopy images. Bildverarbeitung für die Medizin (2018)
24. Studier-Fischer, A., Seidlitz, S., Sellner, J., Bressan, M., Özdemir, B., Ayala, L., Odenthal, J., Knoedler, S., Kowalewski, K.F., Haney, C.M., Salg, G., Dietrich, M., Kenngott, H., Gockel, I., Hackert, T., Müller-Stich, B.P., Maier-Hein, L., Nickel, F.: HeiPorSPECTRAL - the Heidelberg porcine hyperspectral imaging dataset of 20 physiological organs. Scientific Data (2023)
25. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
26. Wallace, M.B., Fockens, P.: Probe-based confocal laser endomicroscopy. Gastroenterology (2009)
27. Zitova, B., Flusser, J.: Image registration methods: A survey. Image and Vision Computing (2003)