MedVSR: Medical Video Super-Resolution with Cross State-Space Propagation

——— Supplementary Material ———

Xinyu Liu[1], Guolei Sun[2], Cheng Wang[1], Yixuan Yuan[1,*], Ender Konukoglu[2]
[1] The Chinese University of Hong Kong, [2] Computer Vision Laboratory, ETH Zurich

This supplementary material presents additional details of Section 1, 3.3, and 4.1. Besides, extra experiments and analysis further demonstrate the effectiveness and memory efficiency of MedVSR.

- **Analogy Between SSM and Linear Attention.** We present a proof of the characteristic of cross state-space propagation in Sec 3.3 based on the analogy between SSM and linear attention.
- **GPU Memory Consumption Analysis.** We provide the GPU memory consumption of different methods to showcase the memory efficiency of MedVSR.
- **More Implementation Details.** We present additional descriptions on the datasets used in the experiments in Sec. 4.1 and the optical flow error analysis in Sec. 1.
- **More Ablation Studies.** We present more comprehensive ablation studies of the design in the proposed MedVSR framework.
- **Results with Longer Training Schedule.** We compare the results with longer training schedule to explore the capability of our model.
- **Theoretical Analysis of CSSP.** We give a theoretical analysis of the proposed propagation scheme in an information bottleneck perspective.
- **More Qualitative Results.** We display more qualitative results for visual comparison of the super-resolution performance of our proposed MedVSR.
- **Limitations and Future Work.** We present the limitations of the proposed MedVSR and potential future directions.

1. Analogy Between SSM and Linear Attention

In Sec. 3.3 in the main paper, we have derived that by processing two state-space sequences from different frames with separate projection layers could produce compounded state-space. We now prove this characteristic through the analogy between SSM and the linear attention mechanism based on [3, 4]. Specifically, the SSM of each iteration during scanning is computed by:

$$h_i = \overline{\mathbf{A}}_i h_{i-1} + \overline{\mathbf{B}}_i x_i, \quad y_i = \mathbf{C}_i h_i, \tag{1}$$

where y_i is the output for the token x_i. After scanning over all tokens, the outputs are used to construct an updated feature $y_{ssm} = [y_1, ..., y_N]$, where N is the number of tokens. By rewriting it via decompositioning the discretized parameters, we have:

$$h_i = \tilde{\mathbf{A}}_i \odot h_{i-1} + \mathbf{B}(\Delta_i \odot x_i), \quad y_i = \mathbf{C}_i h_i, \tag{2}$$

where $\odot$ is the Hadammard product, and $\tilde{\mathbf{A}}_i = \mathrm{diag}(\overline{\mathbf{A}}_i)$ is the matrix composed of diagonal elements in $\overline{\mathbf{A}}_i$. Notably, for each input token, the linear attention mechanism [4] can be formulated by:

$$\mathbf{S}_i = \mathbf{1} \odot \mathbf{S}_{i-1} + \mathbf{K}_i^\top (\mathbf{1} \odot \mathbf{V}_i), \quad y_i = \mathbf{Q}_i \mathbf{S}_i \,/\, \mathbf{Q}_i \mathbf{Z}_i, \tag{3}$$

where $\mathbf{S}_i = \sum_{j=1}^{i} \mathbf{K}_j^\top \mathbf{V}_j, \mathbf{Z}_i = \sum_{j=1}^{i} \mathbf{K}_j^\top$. The outputs are also constructed to a feature $y_{att} = [y_1, ..., y_N]$. By connecting the outputs of the two mechanisms, e.g., y_{ssm} and y_{att}, the analogy from Eq. (2) and (3) shows that the data-dependent parameters correspond to the query, key, and values in the attention mechanism: $\mathbf{C}_i \sim \mathbf{Q}_i$, $\mathbf{B}_i \sim \mathbf{K}_i^\top$, $x_i \sim \mathbf{V}_i$. To this end, our proposed CSSB that applies separate input sequences and projection layers to produce the parameters is similar to the cross-attention mechanism [2, 5] but with significantly better efficiency.

2. GPU Memory Consumption Analysis

We conducted an analysis to assess the GPU memory consumption of our proposed MedVSR model and compared it with two other state-of-the-art models, RVRT [6] and IART [9]. The comparative results are depicted in Fig. 1. For this analysis, we considered video clips consisting of 50 frames. It is evident from the figure that MedVSR demonstrates a gradual and moderate increase in GPU memory consumption with larger input sizes. Specifically, the peak memory usage ranges from 0.55 GB for inputs of 64×64 pixels to 19.93 GB for inputs of 512×512 pixels. In comparison, RVRT and IART exhibit a significantly higher demand for GPU memory, which becomes particularly evident as the input size grows. Both models encounter out-of-memory (OOM) errors when processing frames larger than

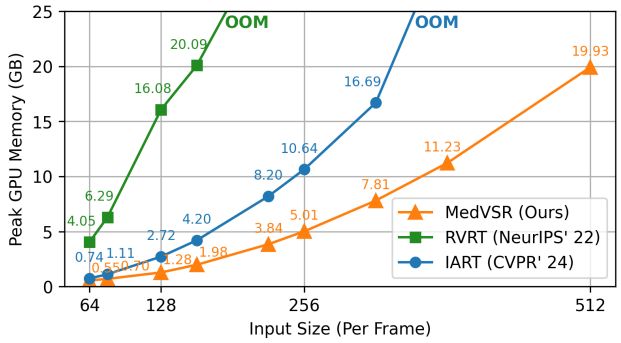

Figure 1. Comparison of peak GPU memory consumption for our MedVSR, RVRT [6], and IART [9] with varied input size.

384×384 pixels. Notably, even with substantial input sizes, MedVSR's memory consumption remains well below the 24 GB threshold, which is well within the capacity of a typical consumer-grade GPU. These findings validate the memory efficiency of our MedVSR model. It is also implied that MedVSR can be feasibly integrated into clinical environments, assisting doctors with diagnostic tasks and other image analysis needs without the requirement for expensive, high-performance computing equipment.

3. More Implementation Details

3.1. Datasets and Implementation Details

We construct our datasets by extracting video clips from the original medical video datasets. Specifically, we divide the video dataset into training, validation, and testing sets, and then sample clips with consecutive frames. This approach ensures that video clips from different sets do not originate from the same video, thereby maintaining set independence. We train and test all methods for exactly the same settings for fair comparison. The spatial resolution for the HyperKvasir is 576×720, with 50 frames per clip. The LDPolyp and EndoVis18 sets have spatial resolutions of 480×560 and 1024×1280, respectively. The spatial resolution of Cataract-101 is 540×720. The diverse spatial resolutions and frame rates across these datasets allow us to evaluate the robustness of our model to varying video qualities and medical scenes.

As introduced in Sec. 3.2 in main paper, our framework is constructed with multiple forward and backward Cross State-Space Propagation (CSSP) branches. Specifically, we employ a sequence of propagations: a backward branch, a forward branch, another backward branch, and a final forward branch. Hence, we obtain propagated and aligned features across a total of four branches. To facilitate the learning of the compound state-space between frame $t-2$ and $t-1$ while improving efficiency, we apply a shared CSSB layer across the four branches, ensuring that the extracted consistent feature will be processed and refined in the state-space of the following branch.

3.2. Optical Flow Error Analysis

In Fig. 1(b) of the main paper, we present an optical flow error analysis on natural domain and medical domain datasets. Here, we provide the details of the experimental setup. Specifically, for a dataset with N video clips $V_1, V_2, \ldots, V_N$, we utilize the averaged forward and reversed backward estimated flow to represent the optical flow error for each dataset. Specifically, for each video clip V_i, we estimate the forward optical flow $O_i^{\text{forward}} \in \mathbb{R}^{(N-1)\times H\times W\times 2}$ with a pretrained SpyNet [7]. Similarly, we compute the backward optical flow $O_i^{\text{backward}} \in \mathbb{R}^{(N-1)\times H\times W\times 2}$, then calculate its reverse flow, which is $\overline{O}_i^{\text{backward}} \in \mathbb{R}^{(N-1)\times H\times W\times 2}$. Lastly, the optical flow error Ω for the dataset with N video clips is given by:

$$\Omega = \frac{\sum_{i=1}^{N} \left\| O_i^{\text{backward}} - \overline{O}_i^{\text{backward}} \right\|_2}{N}. \tag{4}$$

This error value provides a quantitative measure of the consistency between the forward and backward optical flow estimations. A lower value of Ω indicates higher accuracy in flow estimation in the corresponding dataset, suggesting that the optical flow is more reliable across the dataset and and that the frames are easier to align.

4. More Ablation Studies

4.1. Ablation on the Propagation Schemes

Our framework uses two forward and two backward CSSP branches to propagate the features from both previous and future frames. To assess the effectiveness of this design, we compare our method with using only forward or backward propagation in Tab. 1. The results suggest that using both directions yields the best performance with comparable speed, owing to our efficient module design.

Table 1. Ablation study on different propagation schemes.

Method	FLOPs (T)	Latency (s)	HyperKvasir PSNR	HyperKvasir SSIM	LDPolyp PSNR	LDPolyp SSIM
Forward	9.33	1.0936	31.8925	0.9055	31.7188	0.8656
Backward	9.33	1.0877	31.9183	0.9056	31.7525	0.8651
Both (Ours)	9.46	1.1486	**32.0958**	**0.9069**	**31.8333**	**0.8673**

4.2. Ablation on the Number of Propagation Frames

In CSSB, we propagate the $t-2$-th frame feature as the distant support frame. We further ablate the number of propagation frames from $t-2$ to $t-7$ in Tab. 2. Using more distant frames ($\geq$4) negatively impacts both efficiency and performance. This degradation is likely due to the sharp transitions in medical videos, which result in a lack of relevant supporting features in these longer-distance frames. Consequently, we select $t-2$ as the optimal choice.

Table 2. Ablation study on different number of frames propagated.

Metric	$t-2$	$t-3$	$t-4$	$t-5$	$t-6$	$t-7$
PSNR	32.0958	**32.0962**	31.9858	31.9103	31.8903	31.8061
SSIM	**0.9069**	0.9065	0.9054	0.9050	0.9052	0.8907
FLOPs	**9.46**	9.71	9.96	10.20	10.44	10.67
Latency	**1.1486**	1.4109	1.9085	2.2201	2.7136	3.2977

Table 3. Ablation study on the cross state-space with Mamba2 and cross-attention.

Method	FLOPs (T)	Latency (s)	HyperKvasir PSNR	HyperKvasir SSIM	LDPolyp PSNR	LDPolyp SSIM
Attention	9.48	1.5591	31.9106	0.9066	31.8142	**0.8676**
Linear Attention	9.47	1.1892	31.8628	0.9068	31.7989	0.8660
Mamba2 (Ours)	**9.46**	**1.1486**	**32.0958**	**0.9069**	**31.8333**	0.8673

Table 4. Ablation study on more scaling factors.

Factor	Metric	BVSR	BVSR++	VSRT	RVRT	IART	**MedVSR**
2x	PSNR	36.210	36.891	37.107	35.619	37.308	**37.756**
	SSIM	0.9326	0.9381	0.9380	0.9271	0.9389	**0.9401**
6x	PSNR	27.354	27.801	28.052	27.096	28.332	**28.4262**
	SSIM	0.8606	0.8630	0.8635	0.8051	0.8688	**0.8734**

4.3. Ablation on the Cross State-Space and Cross-Attention

In CSSB, we propose a cross state-space mechanism with Mamba2 for propagating relevant features from distant frames. To verify the effectiveness of this design, we present an ablation of replacing Mamba with cross-attention in Tab. 3, and observe that attention mechanisms need larger computational overhead but have comparable performance, due to its quadratic computation complexity. Even with linear attention, using Mamba in the proposed CSSB achieves superiority in both performance and latency.

4.4. More Scaling Factors

Beyond 4x, we examined MedVSR on 2x and 6x in Tab. 4. MedVSR also performs better than existing models for both smaller and larger scaling factors, suggesting robustness and adaptability in varied situations.

5. Results with Longer Training Schedule

To further explore the capabilities of our model, we extended the training schedule to 300,000 training iterations, and test the results on the HyperKvasir testing dataset. [1]. The results are recorded in Tab. 5. It is observed that our MedVSR remains effective and outperforms compared methods, which demonstrate that the strong capability and generalization ability of MedVSR can be further explored with longer training schedules.

Table 5. Results on the HyperKvasir testing dataset [1] of different methods with longer training schedule.

Metric	BasicVSR	BasicVSR++	MedVSR
PSNR	32.8695	32.6034	33.0981
SSIM	0.9046	0.9020	0.9089

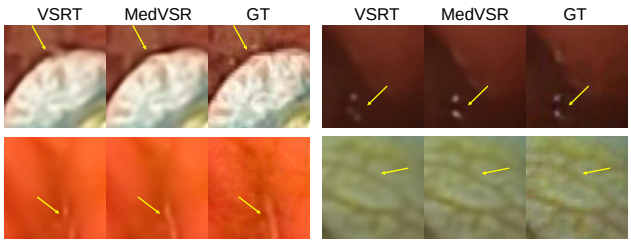

Figure 2. Qualitative comparison with VSRT.

6. Theoretical Analysis of CSSP

Our CSSP mechanism is inherently connected to the *Information Bottleneck (IB) principle [8]*, which aims to efficiently compress the information from the input while preserving its relevance to the target variable. In our sequential frame modeling, the propagation can be represented as a Markov chain, $F_{t-2} \rightarrow F_{t-1} \rightarrow F_t$, where each frame only depends on its immediate predecessor. Under the IB framework, at each propagation step, we consider F_{t-1} as a "bottleneck" variable that mediates the information flow from past frame F_{t-2} to current frame F_t. The objective can thus be viewed as maximizing the mutual information $I(F_{t-1}; F_t)$ to retain features critical for accurate super-resolution. Meanwhile, by implicitly reducing $I(F_{t-2}; F_{t-1})$, CSSP encourages the framework to suppress noise or redundancy accumulated from earlier frames, allowing only salient and informative features to be propagated. This behavior is reflected empirically in our results (as visualized in Fig. 7 of the main paper), where we observe reduced noise and sharper structural details in the propagated feature representations.

7. More Qualitative Results

In Fig. 2, we present more results with significant gains over VSRT on HyperKvasir dataset. Moreover, in Fig. 3 and 4, more visual comparisons between existing VSR methods and the proposed MedVSR are provided. It can be observed that, the compared methods tend to suffer from losing textures and distorting shapes, while our proposed method can continuously generate high quality super-resolved frames with clear edges and detailed structures, such as subtle wrinkles and vessels, and textures on the instruments.

8. Limitations and Future Work

One limitation is that, some medical imaging processes involve 4D data (3D spatial dimensions plus time), such as functional MRI or cardiac imaging. Our model is primarily designed for videos with 2D frames and might require additional adaptation to handle temporal dynamics in 4D scenes. Besides, Mamba operations may not be well-supported on mobile or wearable surgical devices at present. Future work will focus on extending support for 4D data and optimizing the model for compatibility with edge devices, ensuring robustness and efficiency in real-time applications.

References

[1] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1): 283, 2020. 3

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021. 1

[3] Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *arXiv preprint arXiv:2405.16605*, 2024. 1

[4] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 1

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[6] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 1, 2

[7] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[8] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 3

[9] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2546–2555, 2024. 1, 2

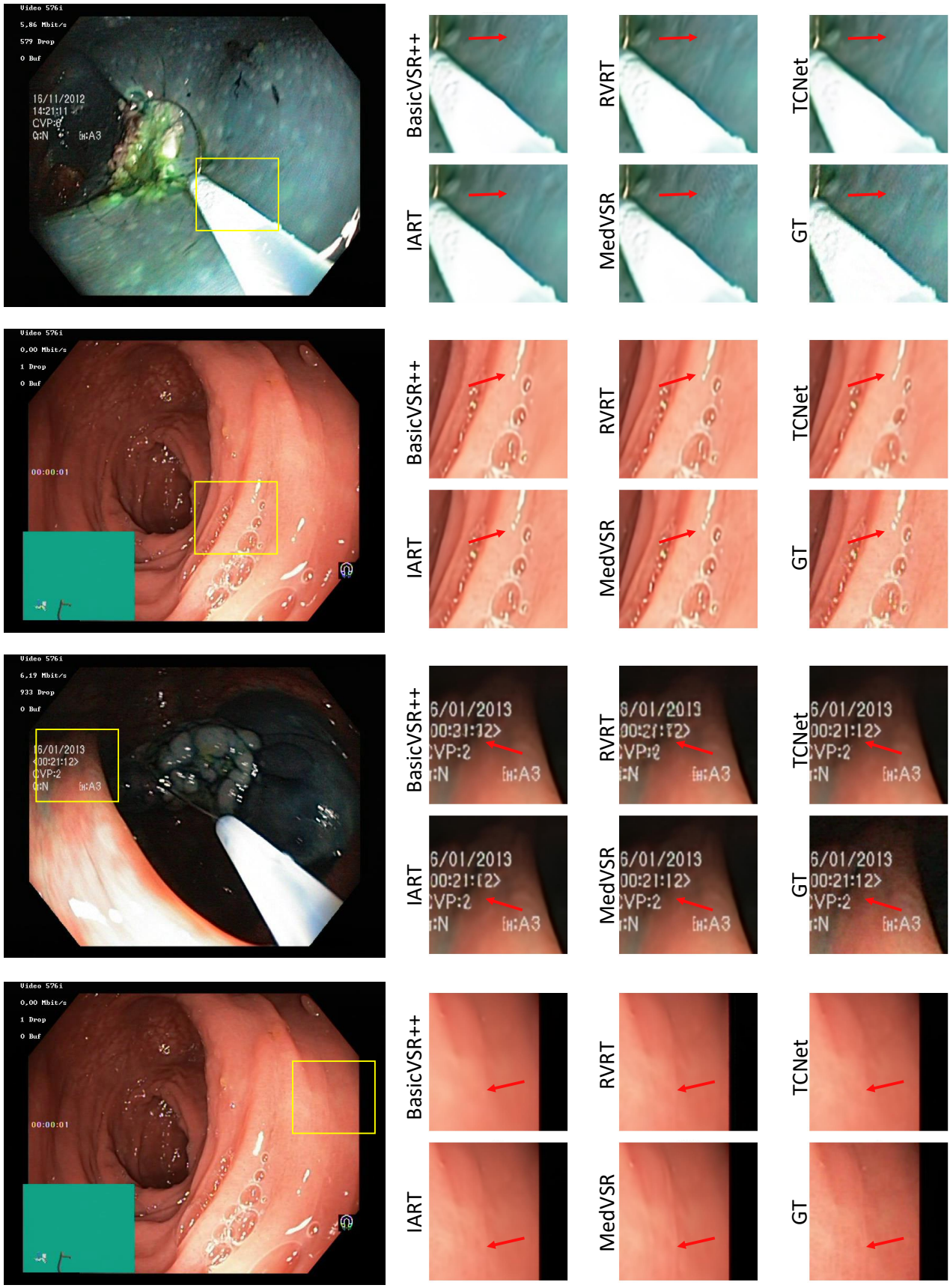

Figure 3. Qualitative comparison on HyperKvasir dataset.

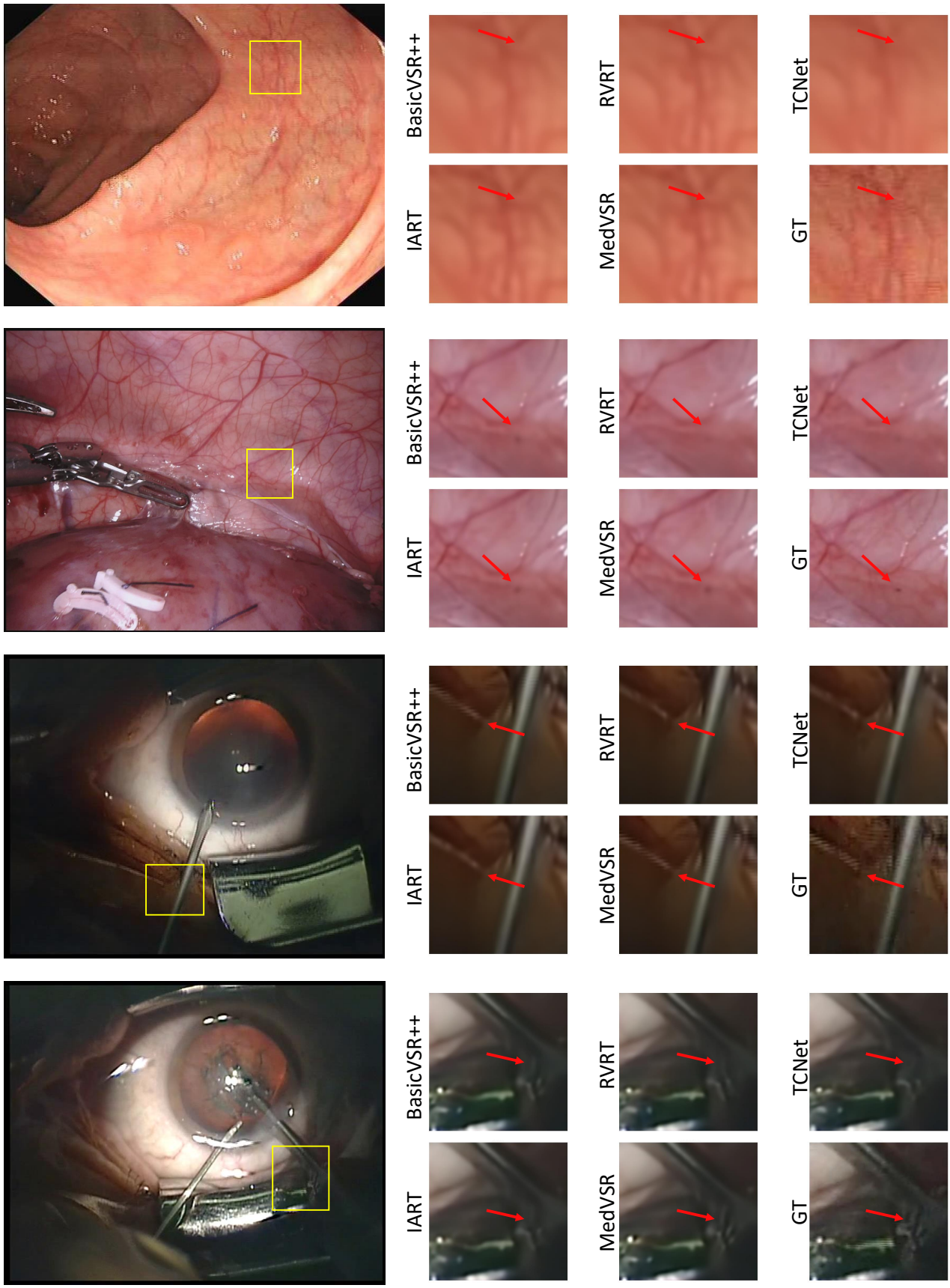

Figure 4. Qualitative comparison on LDPolyp, EndoVis18, and Cataract-101 datasets.