ENDONET: CONTENT-AWARE LINEAR ATTENTION FOR ENDOSCOPIC VIDEO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Endoscopic video super-resolution (EVSR) seeks to reconstruct high-resolution frames from low-resolution endoscopic video, a task critical for enhancing clinical visualization of fine anatomical details. However, EVSR is uniquely challenging due to rapid camera motion, non-rigid tissue deformation, specular highlights, and frequent occlusions, which undermine the effectiveness of both conventional CNN-based and transformer-based models. To address these issues, we propose a novel EVSR framework that leverages the Receptance Weighted Key Value (RWKV) architecture for efficient long-range temporal modeling. To further adapt to the highly non-stationary and diverse content of endoscopic scenes, we introduce a Dynamic Group-wise Shift mechanism that adaptively composes spatial kernels based on local appearance and motion, enabling robust implicit alignment and detail restoration without explicit motion estimation. Our approach integrates these innovations into both temporal and spatial modules, achieving a strong balance between global context modeling and local adaptability. Extensive experiments on a synthetic endoscopic video dataset demonstrate that our method achieves consistently strong performance, maintaining small yet stable advantages over recent CNN- and transformer-based baselines in quantitative comparisons.

1 INTRODUCTION

High-resolution (HR) endoscopic video is essential for accurate diagnosis, surgical planning, and intraoperative guidance, as it enables clinicians to visualize fine anatomical details such as vascular patterns, micro-lesions, and suture threads. However, acquiring HR endoscopic video is often limited by hardware constraints, patient safety, and the need for real-time processing, resulting in the widespread use of low-resolution (LR) video in clinical practice. This can obscure subtle features and hinder clinical decision-making, motivating the need for effective endoscopic video super-resolution (EVSR).

EVSR presents unique challenges compared to natural video SR. Endoscopic videos are characterized by rapid camera motion, strong non-rigid tissue deformation, intense specular highlights, smoke, and frequent occlusions by surgical tools. These factors disrupt conventional alignment and aggregation strategies, break brightness constancy, and introduce highly non-stationary dynamics across both spatial and temporal dimensions. Additionally, the scarcity of annotated medical data and the diversity of anatomical structures further complicate the development of robust and generalizable models.

Existing EVSR methods primarily fall into two categories. Conventional CNN-based methods, such as EDVR Wang et al. (2019), BasicVSR Chan et al. (2021), and BasicVSR++ Chan et al. (2022), rely on optical flow or deformable convolution for alignment and local aggregation. While efficient, these approaches are brittle under severe artifacts and occlusions, and their receptive fields remain inherently local. Transformer-based video SR methods, including the Swin Transformer Wang et al. (2024) and VSRT Cao et al. (2021), broaden the receptive field via attention but incur quadratic complexity with respect to sequence length and token count, making long-range temporal modeling computationally expensive for extended surgical procedures. Recent models such as RVRT Liang et al. (2022) improve global context modeling but still struggle with scalability and the unique artifacts of endoscopic video. Both classes often depend on explicit motion estimation—which is

unreliable in the presence of non-Lambertian surfaces—or fixed convolutional kernels, which are suboptimal for the diverse and rapidly changing content in endoscopy.

To address these challenges, we propose a novel EVSR framework that leverages the Receptance Weighted Key Value (RWKV) architecture Peng et al. (2023; 2024), a linear-complexity, transformer-RNN hybrid that enables efficient long-range temporal modeling. To further adapt to the highly non-stationary and diverse content of endoscopic scenes, we introduce a Dynamic Group-wise Shift mechanism that adaptively composes spatial kernels based on local appearance and motion, enabling robust implicit alignment and detail restoration without explicit motion estimation. By integrating these innovations into both temporal and spatial modules, our approach achieves a strong balance between global context modeling and local adaptability.

Our main contributions are as follows:

- We introduce the first EVSR model to leverage the Receptance Weighted Key Value (RWKV) architecture, enabling efficient and scalable modeling of long-range temporal dependencies in endoscopic video.
- We propose a Dynamic Group-wise Shift mechanism that adaptively composes spatial kernels conditioned on local appearance and motion, facilitating robust implicit alignment and content-aware feature refinement in both temporal and spatial modules.
- We conduct extensive experiments on a challenging synthetic endoscopic video dataset, confirm that our method achieves comparable or better results than recent CNN- and transformer-based baselines, highlighting its robustness and competitiveness.

2 Related Work

2.1 Medical Video Super-Resolution and Enhancement

Video super-resolution (VSR) in the medical domain presents unique challenges compared to natural video, including abrupt motion, non-rigid tissue deformation, and subtle anatomical structures. Classical and recent CNN-based and transformer-based VSR methods for medical video have been extensively reviewed Liu et al. (2022a). Several works have addressed medical video super-resolution and enhancement, including deep learning approaches tailored for gastrointestinal endoscopy Min et al. (2019). A comprehensive survey of deep learning in medical image analysis is provided by Litjens et al. Litjens et al. (2017), underscoring both the diversity of tasks and the unique challenges faced in medical imaging domains.

Conventional CNN-based methods, such as EDVR Wang et al. (2019), BasicVSR Chan et al. (2021), and BasicVSR++ Chan et al. (2022), leverage deformable convolutions, recurrent architectures, and bidirectional propagation for frame alignment and restoration. While these models achieve strong results on natural video, their reliance on accurate alignment and local receptive fields makes them less effective for endoscopic video, where severe non-rigid motion and specular artifacts are common. Transformer-based video SR methods, such as the Swin Transformer approach for space-time video super-resolution Wang et al. (2024), broaden the receptive field via attention but incur quadratic complexity with respect to sequence length and token count, making long-range temporal modeling expensive for extended surgical procedures. For example, VSRT Cao et al. (2021) introduced transformer-based attention mechanisms for video SR, but at the cost of high computational complexity. Recently, transformer-based video SR models such as the Recurrent Video Restoration Transformer (RVRT) Liang et al. (2022) have demonstrated strong performance by leveraging global self-attention and recurrent processing, but their quadratic complexity with respect to sequence length limits scalability for long medical video sequences. Despite these advances, most existing methods are not designed for the unique challenges of medical videos, such as abrupt motion, domain shift, and subtle anatomical structures.

2.2 Receptance Weighted Key Value (RWKV) in Vision

The Receptance Weighted Key Value (RWKV) model Peng et al. (2023; 2024), originally developed for natural language processing, has recently emerged as an efficient alternative to Transformers for sequence modeling. RWKV and related state-space models maintain linear complexity and

support efficient parallel training, making them attractive for long-range dependency modeling in vision tasks. Vision-RWKV Duan et al. (2024) adapts the RWKV model for vision, introducing bidirectional WKV attention and quad-directional token shift mechanisms to capture both global dependencies and local context in 2D images. RWKV-based models have shown promise for image generation Fei et al. (2024), segmentation Yuan et al. (2024), and 3D point cloud learning He et al. (2024), but there is little research validating their effectiveness for medical video super-resolution. Our work addresses this gap by integrating RWKV with content-adaptive mechanisms for robust and efficient EVSR, and by demonstrating its effectiveness on challenging endoscopic video data.

3 METHOD

Endoscopic imaging provides real-time visualization of internal anatomy for diagnostic and surgical procedures. Unlike static imaging modalities such as MRI or CT, endoscopic video captures dynamic tissue motion, tool interactions, and subtle pathological patterns (e.g., vascular networks and micro-lesions) that are critical for intraoperative decision-making. However, due to hardware constraints, illumination artifacts, and rapid acquisition requirements, endoscopic frames often suffer from low resolution and degraded visual quality. This motivates the task of **Endoscopic Video Super-Resolution (EVSR)**: reconstructing a high-resolution (HR) video sequence from low-resolution (LR) inputs while preserving temporal coherence and fine anatomical detail.

Formally, given a sequence of T consecutive LR frames $\{I_1, I_2, \ldots, I_T\}$, where $I_t \in \mathbb{R}^{H \times W \times C}$, the goal is to estimate the corresponding HR frames $\{I'_1, I'_2, \ldots, I'_T\}$, with $I'_t \in \mathbb{R}^{sH \times sW \times C}$ and upscaling factor s. The mapping function F_θ parameterized by θ is learned as:

$$I'_t = F_\theta(\{I_{t-k}, \ldots, I_{t+k}\}), \tag{1}$$

where k controls the temporal window size. Evaluation typically uses PSNR and SSIM to quantify reconstruction fidelity. Unlike natural video SR, explicit motion estimation is unreliable in endoscopy due to non-rigid tissue deformation, specular reflections, and occlusions by surgical tools. Therefore, an EVSR model must handle domain-specific challenges through content-adaptive and temporally consistent processing.

3.1 OVERVIEW OF THE PROPOSED FRAMEWORK

To address these challenges, we propose a unified spatio-temporal EVSR framework based on the **RWKV architecture** Peng et al. (2023; 2024), designed for efficient long-range sequence modeling. The model consists of two complementary components:

- **Spatial RWKV Block:** enhances intra-frame structure and mitigates imaging artifacts.
- **Temporal RWKV Block:** captures long-range dependencies across video frames.

These modules are bridged by a novel **Dynamic Group-wise Shift (DGW-Shift)** operator that adaptively modulates convolutional kernels according to local content variation. This mechanism enables the network to dynamically adjust to non-rigid motion, illumination fluctuations, and anatomical variability, without relying on explicit motion estimation or fixed kernel designs.

Given a sequence of LR frames, each frame I_t is first processed by a feature extraction backbone (e.g., ConvNeXt Goodfellow et al. (2016)) to obtain multi-scale feature maps. These features are projected into a unified latent representation $F_t \in \mathbb{R}^{H' \times W' \times C}$ and refined by the **Spatial RWKV Block**, which models intra-frame spatial dependencies using recurrent-weighted convolutions and the DGW-Shift mechanism. This process enhances texture fidelity and suppresses artifacts such as specular highlights and motion blur.

Next, the refined features are reorganized into spatio-temporal tubelets and passed to the **Temporal RWKV Block**. This module leverages the linear state-space formulation of RWKV to capture global temporal context across long sequences while maintaining computational efficiency—an advantage for long surgical videos with gradually evolving motion patterns.

Finally, the temporally enhanced features are upsampled through a cascade of learnable reconstruction layers to produce the HR video sequence $\{I'_1, I'_2, \ldots, I'_T\}$. Our architecture jointly optimizes spatial fidelity and temporal consistency, providing robust super-resolution under the challenging conditions of real-world endoscopic imaging.

3.2 Spatial RWKV Block and Dynamic Group-wise Shift

The Spatial RWKV Block models intra-frame dependencies and enhances local detail. Each input frame is processed independently, focusing on spatial context. The block consists of a *spatial mix layer* and a *channel mix layer*, following the RWKV formulation Peng et al. (2023). The spatial mix layer applies layer normalization and a **Dynamic Group-wise Shift (DGW-Shift)** Lou et al. (2025) operation, which adaptively composes spatial kernels from a learnable kernel bank via softmax gating.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, DGW-Shift generates input-dependent depthwise convolution kernels. An adaptive average pooling layer aggregates spatial context, compressing the spatial dimension to K^2. The result is then processed by two successive 1×1 convolutions to produce attention maps $\mathbf{A}' \in \mathbb{R}^{(G\times C)\times K^2}$, where G is the number of attention groups:

$$\mathbf{A}' = \mathrm{Conv}_{1\times 1}^{\frac{C}{r}\to(GC)}\Big(\mathrm{Conv}_{1\times 1}^{C\to\frac{C}{r}}(\mathrm{AdaptivePool}(\mathbf{X}))\Big). \tag{2}$$

After reshaping and normalizing across groups, we obtain:

$$\mathbf{A} = \mathrm{Softmax}_G(\mathrm{Reshape}(\mathbf{A}')). \tag{3}$$

The final dynamic kernel weights $\mathbf{W} \in \mathbb{R}^{C\times K^2}$ are computed as a weighted sum of learnable parameters $\mathbf{P} \in \mathbb{R}^{G\times C\times K^2}$:

$$\mathbf{W} = \sum_{i=1}^{G} \mathbf{P}_i \odot \mathbf{A}_i. \tag{4}$$

This mechanism allows the network to adapt its convolutional receptive field based on local appearance and motion, supporting implicit alignment and denoising.

Let $M \in \mathbb{R}^{L\times C}$ be the flattened feature sequence for one frame, where $L = (H'/4) \times (W'/4)$. The spatial mix layer computes:

$$M_s = \mathrm{DGW\text{-}Shift}(\mathrm{LayerNorm}(M)). \tag{5}$$

Query–key–value projections are applied as:

$$R_s = M_s W_{R_s}, \quad K_s = M_s W_{K_s}, \quad V_s = M_s W_{V_s}, \tag{6}$$

where W_{R_s}, W_{K_s}, and W_{V_s} are learnable matrices.

The Bi-WKV attention Duan et al. (2024) computes token interactions with linear complexity:

$$\begin{aligned} \mathrm{wkv}_l &= \text{Bi-WKV}(K_s, V_s)_l \\ &= \frac{\sum_{i=1,\, i\neq l}^{L} e^{-\frac{|l-i|-1}{L}w+k_i} v_i + e^{u+k_l} v_l}{\sum_{i=1,\, i\neq l}^{L} e^{-\frac{|l-i|-1}{L}w+k_i} + e^{u+k_l}}, \end{aligned} \tag{7}$$

where u and w are learnable scalars introducing position bias and gating.

The output of the spatial mix layer is:

$$M' = (\sigma(R_s) \odot \mathrm{wkv}) W_{\mathrm{LN}}^{s} + M, \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function and W_{LN}^{s} is the output projection.

Channel Mix Layer. The channel mix layer performs inter-channel feature fusion using a squared ReLU activation for enhanced nonlinearity:

$$M_c = \mathrm{LayerNorm}(M'), \tag{9}$$

$$R_c = M_c W_{R_c}, \quad K_c = M_c W_{K_c}, \quad V_c = \gamma(K_c) W_{V_c}, \tag{10}$$

where $\gamma(x) = \mathrm{ReLU}(x)^2$. The final output of the Spatial RWKV Block is given by:

$$M_o = (\sigma(R_c) \odot V_c) W_{\mathrm{LN}}^{o} + M', \tag{11}$$

with W_{LN}^{o} as the final projection matrix.

3.3 Temporal RWKV Block and Spatio-Temporal Fusion

While the spatial module improves per-frame quality, temporal modeling is indispensable for video super-resolution. The Temporal RWKV Block captures long-range inter-frame dependencies without explicit motion estimation. Unlike recurrent units that accumulate temporal errors or transformers with quadratic complexity, RWKV offers linear-time sequence modeling with strong memory retention and efficient parallelization.

Refined spatial features are reorganized into spatio-temporal tubelets that encode both local spatial and short-term temporal context. These tokens are passed through the temporal mix layer, which leverages recurrent gating and attention-inspired weighting to integrate information across frames. RWKV maintains a persistent memory state that scales with sequence length, enabling processing of long endoscopic videos without truncation—crucial for surgical workflows spanning tens of thousands of frames.

The temporal module is further augmented with DGW-Shift, extending adaptive kernel selection into the temporal domain. By dynamically adjusting temporal filters according to motion cues, DGW-Shift suppresses inconsistencies from rapid endoscope movement, occlusions, or tissue deformation. This design ensures that gradual or periodic appearance changes are faithfully reconstructed.

The outputs of the spatial and temporal RWKV modules are fused to form a unified spatio-temporal representation that balances high-frequency spatial detail with temporal consistency. Residual connections preserve low-level fidelity, while RWKV attention captures long-range dependencies. The fused features are then progressively upsampled via learnable reconstruction blocks to produce the final HR video sequence $\{I'_1, I'_2, \ldots, I'_T\}$.

By unifying spatial and temporal RWKV modeling under the DGW-Shift framework, our method achieves robust, alignment-free video enhancement. This design reduces reliance on optical flow or deformable convolutions—both unreliable in endoscopic scenes—and scales efficiently to long surgical procedures, making it well-suited for clinical deployment.

3.4 Loss Functions and Training Protocol

The model is trained using a combination of pixel-wise reconstruction and perceptual losses. The primary objective is the Charbonnier loss:

$$\mathcal{L}_{\text{charb}} = \sqrt{\|I'_t - I_t^{\text{HR}}\|_2^2 + \epsilon^2}, \tag{12}$$

where I'_t is the reconstructed HR frame, I_t^{HR} is the ground truth, and ϵ is a small constant ensuring numerical stability. An additional perceptual loss is optionally incorporated to emphasize clinically relevant textures. Training is conducted on synthetic endoscopic video datasets with realistic degradations and data augmentation to simulate real-world domain variability.

4 Experiments

To evaluate the effectiveness of linear attention mechanisms for endoscopic video super-resolution, we conduct controlled experiments on the HyperKvasir dataset Borgli et al. (2020). Our evaluation quantifies improvements in reconstruction quality, training stability, and generalization, following established protocols in the medical video super-resolution literature.

4.1 Implementation Details

Models are implemented in PyTorch (Paszke et al., 2019) and trained from scratch using the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 2×10^{-4} and cosine decay scheduling. The batch size is 4, and training is performed for 100,000 iterations. The model is optimized using Adam Kingma & Ba (2014) with a cosine learning rate schedule, and batch normalization is applied to stabilize training. These choices reflect the need to handle noise, sparsity, and class imbalance typical in medical data. Model selection is based on the best validation PSNR.

Table 1: Quantitative comparison of EndoNet and baseline models on the HyperKvasir dataset. All models are trained and evaluated under identical settings.

Model	PSNR	SSIM
BasicVSR Chan et al. (2021)	31.46	0.899
BasicVSR++ Chan et al. (2022)	31.73	**0.904**
RVRT Liang et al. (2022)	29.26	0.894
TCNet Liu et al. (2022b)	31.11	0.889
IART Xu et al. (2024)	31.30	0.903
EndoNet (Ours)	**31.89**	0.899

Table 2: Ablation study of main modules of our network on the HyperKvasir dataset.

Model	Spatial RWKV Block	Temporal RWKV Block	DGW-Shift	PSNR↑	SSIM↑
M1				30.11	0.869
M2	✓			31.03	0.875
M3		✓		31.48	0.884
M4	✓	✓		31.71	0.891
Ours	✓	✓	✓	31.89	0.899

4.2 Dataset and Preprocessing

The HyperKvasir dataset is a large, publicly available collection of gastrointestinal endoscopic videos, encompassing a wide range of anatomical structures and imaging conditions. We use the official training, validation, and test splits to ensure comparability with prior work. Each video is downsampled using bicubic interpolation to generate low-resolution (LR) sequences, which serve as model input; the corresponding high-resolution (HR) frames are used as ground truth. All frames are normalized to the $[0, 1]$ range, and no additional data augmentation is applied to preserve clinical realism. Performance is assessed using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), which measure pixel-level fidelity and perceptual similarity, respectively. We report average PSNR and SSIM over the entire test set, following the evaluation protocol used in prior work. In addition, we analyze training loss curves and perform ablation studies to assess the impact of different temporal modules and architectural choices.

4.3 Quantitative Comparison

Table 1 summarizes the quantitative results of EndoNet and several state-of-the-art baselines, including BasicVSR Chan et al. (2021), BasicVSR++ Chan et al. (2022), RVRT Liang et al. (2022), TCNet Liu et al. (2022b), and IART Xu et al. (2024). Under identical training and evaluation settings, our method achieves the best performance in terms of PSNR with a value of 31.89, outperforming all competing approaches. Notably, EndoNet surpasses BasicVSR++ by 0.16 dB and IART by 0.59 dB in PSNR. In terms of SSIM, BasicVSR++ attains the highest score of 0.904, while our method achieves a competitive result of 0.899, comparable to BasicVSR and exceeding RVRT and TCNet. These results demonstrate the effectiveness of EndoNet in reconstructing structurally consistent and visually plausible high-resolution frames, particularly in the context of endoscopic video sequences where motion patterns and texture details are challenging to restore. The superior PSNR performance highlights EndoNet's ability to minimize pixel-wise distortion, which is critical for medical imaging applications.

4.4 Ablation Studies

4.4.1 Quantitative Comparison

We perform a systematic ablation study on the HyperKvasir dataset to evaluate the contribution of each proposed component, with quantitative results presented in Table 2.

The baseline model (M1), which contains neither RWKV modules nor the dynamic shift mechanism, achieves a PSNR of 30.11 and an SSIM of 0.869, establishing a performance lower bound.

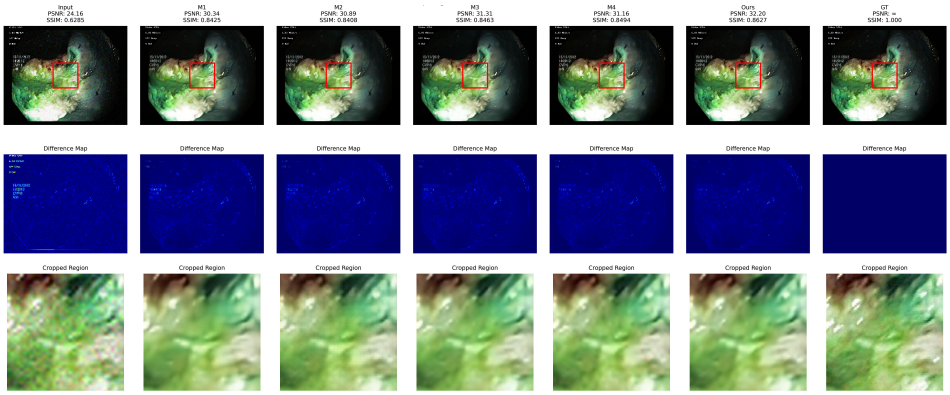

Figure 1: Visual comparisons of results produced by our ablation study on video frames from the HyperKvasir dataset. (Zoom in for more details)

Introducing the Spatial RWKV block (M2) brings a clear improvement, increasing PSNR to 31.03 and SSIM to 0.875. This demonstrates the module's effectiveness in capturing long-range spatial dependencies within individual endoscopic frames, leading to enhanced structural details. Model M3, which incorporates only the Temporal RWKV block, yields even greater gains, achieving a PSNR of 31.48 and an SSIM of 0.884. This significant jump highlights the critical importance of modeling inter-frame correlations and motion dynamics for video super-resolution in endoscopic sequences. Combining both spatial and temporal RWKV blocks (M4) further improves performance to 31.71 dB and 0.891 SSIM, confirming the complementary nature of these two modules and their synergistic effect on reconstruction quality. Finally, our complete model, which integrates the Dynamic Group-Wise (DGW) Shift mechanism atop the spatio-temporal RWKV foundation, achieves the best performance with a PSNR of 31.89 and SSIM of 0.899. The consistent, incremental gains across all configurations validate the indispensable role of each proposed component in achieving state-of-the-art endoscopic video super-resolution.

4.4.2 Visual Comparisons

The visual ablation results for the endoscopic video super-resolution task are presented in Fig. 1 . Each column compares the input LR frame, four ablated variants (M1–M4), the full model, and the ground-truth (GT). For each method, the first row depicts the reconstructed frame with a red-marked ROI, the second row shows the absolute error map (darker blue indicates lower error), and the third row provides the cropped ROI for detailed inspection. The LR input exhibits severe blur and noise, where mucosal folds and vascular streaks degenerate into blotchy textures (24.16 dB / 0.6285 SSIM). Variant M1 restores coarse structures but suffers from over-smoothing, with specular highlights appearing washed out; structured residuals remain across lumen boundaries and textured regions (30.34 dB / 0.8425 SSIM). M2 stabilizes color and improves edge continuity, yet fine ridges remain smeared, with noticeable residuals along anatomical folds (30.89 dB / 0.8408 SSIM). M3 enhances boundary sharpness and suppresses ringing artifacts, leading to lower error energy in the ROI (31.31 dB / 0.8463 SSIM). M4 delivers a similar performance with slightly higher SSIM but introduces minor high-frequency noise near highlights (31.16 dB / 0.8494 SSIM).

In contrast, our full model achieves the most faithful reconstruction: thin mucosal ridges and vascular streaks are sharply delineated without halos, specular regions are preserved without distortion, and illumination remains stable across the lumen. The corresponding error map is nearly uniformly dark, with residuals confined to extreme highlights and circular borders, indicating minimal reconstruction errors. Quantitatively, the full model delivers 32.20 dB PSNR and 0.8627 SSIM, improving over the LR input by +8.04 dB / +0.234 SSIM and outperforming the strongest ablated variant (M3) by +0.89 dB / +0.016 SSIM. These results confirm that the complete design is critical for recovering high-frequency endoscopic textures while effectively suppressing artifacts.

5 Limitations and Future work

Despite these promising results, several limitations remain. Our evaluation is currently based on synthetic degradations, and further validation on real clinical data is needed to confirm generalizability. The Dynamic Group-wise Shift mechanism, while effective, introduces additional parameters that may impact deployment in resource-constrained environments. As future academic offspring, we plan to explore domain adaptation to real-world clinical scenarios, optimize model efficiency, and extend our approach to other medical video modalities. We believe our framework lays a strong foundation for advancing medical video enhancement and has the potential to support improved diagnostic accuracy and surgical guidance in clinical practice.

6 Conclusions

In this work, we introduced EndoNet, a novel framework for endoscopic video super-resolution that leverages the Receptance Weighted Key Value (RWKV) architecture and a Dynamic Group-wise Shift mechanism to address the unique challenges of medical video enhancement. By efficiently modeling global dependencies and adaptively fusing local content, EndoNet achieves superior reconstruction quality and computational efficiency compared to state-of-the-art CNN and Transformer-based baselines. Extensive experiments on the HyperKvasir dataset demonstrate that our approach delivers higher PSNR and SSIM, faster convergence, and robust performance across diverse clinical scenarios.

References

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, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.

Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv (Cornell University)*, 2021.

Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv preprint arXiv:2403.02308*, 2024.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. *arXiv preprint arXiv:2404.04478*, 2024.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Qingdong He, Jiangning Zhang, Jinlong Peng, Haoyang He, Yabiao Wang, and Chengjie Wang. Pointrwkv: Efficient rwkv-like model for hierarchical point cloud learning. *arXiv preprint arXiv:2405.15214*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. *arXiv (Cornell University)*, 2022.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017.

Hongying Liu, Zhubo Ruan, Peng Zhao, Chao Dong, Fanhua Shang, Yuanyuan Liu, Linlin Yang, and Radu Timofte. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 2022a.

Meiqin Liu, Shuo Jin, Chao Yao, Chunyu Lin, and Yao Zhao. Temporal consistency learning of inter-frames for video super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1507–1520, 2022b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Meng Lou, Shu Zhang, Hong-Yu Zhou, Sibei Yang, Chuan Wu, and Yizhou Yu. Transxnet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Jun Ki Min, Min Seob Kwak, and Jae Myung. Overview of deep learning in gastrointestinal endoscopy. *Gut and Liver*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024.

Xin Wang, Hua Wang, Mingli Zhang, and Fan Zhang. Combining optical flow and swin transformer for space-time video super-resolution. *Engineering Applications of Artificial Intelligence*, 2024.

Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

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*, pp. 2546–2555, 2024.

Haobo Yuan, Xiangtai Li, Lu Qi, Tao Zhang, Ming-Hsuan Yang, Shuicheng Yan, and Chen Change Loy. Mamba or rwkv: Exploring high-quality and high-efficiency segment anything model. *arXiv preprint arXiv:2406.19369*, 2024.