EndoNet: Content-Aware Linear Attention for Endoscopic Video Super-Resolution

Anonymous Author(s)
Affiliation
Address
email

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 (22), BasicVSR (3), and BasicVSR++ (4), 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 (21) and VSRT (2), broaden the receptive field via attention but incur quadratic complexity with respect to sequence length and token count, making

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

long-range temporal modeling computationally expensive for extended surgical procedures. Recent models such as RVRT (10) 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 (18; 19), 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 (12). Several works have addressed medical video super-resolution and enhancement, including deep learning approaches tailored for gastrointestinal endoscopy (16). A comprehensive survey of deep learning in medical image analysis is provided by Litjens et al. (11), underscoring both the diversity of tasks and the unique challenges faced in medical imaging domains.

Conventional CNN-based methods, such as EDVR (22), BasicVSR (3), and BasicVSR++ (4), 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 (21), 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 (2) 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) (10) 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 (18; 19), 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 (5) 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 (6), segmentation (24), and 3D point cloud learning (8), 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 Background

Endoscopic imaging is a minimally invasive modality widely used in clinical diagnostics and surgery, providing real-time visualization of internal anatomical structures such as the gastrointestinal tract, airways, and abdominal cavity. Unlike static imaging modalities like MRI or CT, endoscopic video captures dynamic tissue motion, tool interactions, and subtle pathological features (e.g., micro-lesions, vascular patterns) that are critical for diagnosis and intraoperative decision-making. However, the quality of endoscopic video is often limited by hardware constraints, illumination artifacts, and the need for rapid acquisition, resulting in low-resolution (LR) frames that may obscure clinically relevant details.

Formally, the endoscopic video super-resolution (EVSR) problem can be defined as follows. Given a sequence of T consecutive LR frames $\{I_1, I_2, \ldots, I_T\}$, where $I_t \in \mathbb{R}^{H \times W \times C}$ denotes the t-th frame with spatial resolution $H \times W$ and C color channels, the goal is to reconstruct a sequence of high-resolution (HR) frames $\{I'_1, I'_2, \ldots, I'_T\}$, where $I'_t \in \mathbb{R}^{sH \times sW \times C}$ and s is the upscaling factor (typically $s = 4$). The mapping from LR to HR frames is learned by a function F_θ parameterized by θ, such that $I'_t = F_\theta(\{I_{t-k}, \ldots, I_t, \ldots, I_{t+k}\})$, where k controls the temporal window size. The objective is to maximize the fidelity of I'_t to the unknown ground truth HR frame, typically measured by metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM).

Several clinical and data-specific assumptions are made to facilitate EVSR modeling. First, the imaging protocol is assumed to be consistent within each video sequence, with fixed frame rates and illumination settings. Second, anatomical structures within the field of view may exhibit local homogeneity (e.g., mucosal surfaces) but are subject to rapid non-rigid deformation and occlusion by surgical tools. Third, LR frames may contain artifacts such as specular highlights, motion blur, and noise, which complicate both spatial and temporal alignment. Unlike natural video SR, explicit motion estimation is often unreliable due to these artifacts and the non-Lambertian nature of biological tissues (20). Therefore, EVSR models must be robust to domain-specific challenges and capable of leveraging both global temporal context and local content adaptivity.

Prior work in video super-resolution has explored CNN-based and transformer-based architectures (7; 20), but these approaches are limited by either local receptive fields or high computational complexity. Recent advances in state-space models and hybrid architectures such as RWKV (18; 19) offer promising solutions for efficient long-range dependency modeling. In the context of endoscopic video, our work builds on these foundations by integrating content-adaptive mechanisms and linear attention to address the unique challenges of medical video enhancement.

4 Method

To address the dual challenges of spatial detail refinement and temporal coherence in endoscopic video super-resolution, we design a unified spatio-temporal processing pipeline based on the RWKV framework (18; 19). Our EVSR architecture consists of two complementary modules: (1) a **Spatial RWKV Block**, which enhances intra-frame structures and mitigates image artifacts, and (2) a **Temporal RWKV Block**, which captures long-range dependencies across video sequences. Both modules are tightly integrated with the proposed **Dynamic Group-wise Shift (DGW-Shift)** operator, enabling adaptive kernel composition to handle content-dependent variations in appearance and motion. This joint design is particularly suited for endoscopic scenarios, where non-rigid deformations, specular highlights, occlusions, and frequent topology changes make alignment-based methods unreliable.

Our framework incorporates several domain-specific adaptations for medical video. The Dynamic Group-wise Shift mechanism is designed to handle anatomical variability and imaging artifacts by enabling content-aware kernel selection. The RWKV-based temporal modeling supports long sequences, making the approach suitable for extended surgical procedures. Unlike prior methods that rely on explicit motion estimation or fixed kernels, our model adapts dynamically to the input, improving robustness to occlusions and non-rigid motion. These innovations collectively enable superior performance in endoscopic video super-resolution, as demonstrated in our experiments.

4.1 Overall Architecture and Processing Pipeline

Given a sequence of T consecutive LR frames $\{I_1, I_2, \ldots, I_T\}$, each frame is first processed by a feature extraction backbone (e.g., ConvNeXt (7)) to obtain multi-scale feature maps. These features are projected and fused to form a unified representation $F_t \in \mathbb{R}^{H' \times W' \times C}$ for each frame. The fused features are then passed through the Spatial RWKV Block, which models intra-frame spatial dependencies and refines local details. The output of the spatial module is subsequently downsampled and reorganized into spatio-temporal tubelets, which serve as input tokens for the Temporal RWKV Block. This block models long-range temporal dependencies across frames, leveraging the RWKV architecture (18; 19) for efficient sequence processing. The final output is upsampled through a cascade of learnable upsampling blocks to produce the HR video sequence $\{I'_1, I'_2, \ldots, I'_T\}$.

4.2 Spatial RWKV Block and Dynamic Group-wise Shift

The Spatial RWKV Block models intra-frame dependencies and enhances local details. 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 framework (18). The spatial mix layer uses layer normalization and a Dynamic Group-wise Shift (DGW-Shift) operation (15), which adaptively composes spatial kernels from a learnable bank via softmax gating. Specifically, to inject inductive bias and perform local feature aggregation in a dynamic input-dependent manner, the DGW-Shift module generates input-dependent depthwise convolution kernels. Given an input feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, an adaptive average pooling layer first aggregates spatial contexts, compressing the spatial dimension to K^2. This compressed representation is then processed by two successive 1×1 convolutional layers, producing attention maps $\mathbf{A'} \in \mathbb{R}^{(G \times C) \times K^2}$, where G denotes the number of attention groups. Subsequently, $\mathbf{A'}$ is reshaped into $\mathbb{R}^{G \times C \times K^2}$ and a softmax function is applied along the group dimension G to generate the normalized attention weights $\mathbf{A} \in \mathbb{R}^{G \times C \times K^2}$. These weights are then element-wise multiplied with a set of learnable parameters $\mathbf{P} \in \mathbb{R}^{G \times C \times K^2}$, and the product is summed over the group dimension, yielding the dynamic kernels $\mathbf{W} \in \mathbb{R}^{C \times K^2}$. This process is formally defined as:

$$\mathbf{A'} = \mathrm{Conv}_{1\times1}^{\frac{C}{r} \to (G \times C)} \left(\mathrm{Conv}_{1\times1}^{C \to \frac{C}{r}} \left(\mathrm{AdaptivePool}(\mathbf{X}) \right) \right) \tag{1}$$

$$\mathbf{A} = \mathrm{Softmax}\left(\mathrm{Reshape}(\mathbf{A'}) \right) \tag{2}$$

$$\mathbf{W} = \sum_{i=0}^{G} \mathbf{P}_i \mathbf{A}_i \tag{3}$$

This mechanism enables the model to adjust convolutional operators based on local appearance and motion, facilitating robust implicit alignment and denoising. The Bi-WKV attention mechanism (5) is then applied to capture long-range spatial dependencies with linear complexity. The channel mix layer performs feature fusion in the channel dimension, using squared ReLU activation for enhanced nonlinearity and a multi-layer perceptron for integration.

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

$$M_s = DGW - Shift(LayerNorm(M)). \tag{4}$$

$$R_s = M_s W_{R_s}, \quad K_s = M_s W_{K_s}, \quad V_s = M_s W_{V_s}, \tag{5}$$

where W_{R_s}, W_{K_s}, and W_{V_s} are learnable projections. The Bi-WKV attention output for token l is:

$$wkv_l = Bi - WKV(K_s, V_s)_l = \frac{\sum_{i=1, i\neq l}^{L} e^{-(|l-i|-1)/L\cdot w + k_i} v_i + e^{u+k_l} v_l}{\sum_{i=1, i\neq l}^{L} e^{-(|l-i|-1)/L\cdot w + k_i} + e^{u+k_l}}, \tag{6}$$

with relative position bias and gating as described in Eq. 6. The final output is modulated by the receptance gate:

$$M' = (\sigma(R_s) \odot wkv) W_{LN}^s + M, \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function.

The channel mix layer processes M' as:

$$M_c = LayerNorm(M'), \tag{8}$$

$$R_c = M_c W_{R_c}, \quad K_c = M_c W_{K_c}, \quad V_c = \gamma(K_c) W_{V_c}, \tag{9}$$

where $\gamma(\cdot)$ is squared ReLU, and W_{R_c}, W_{K_c}, W_{V_c} are learnable projections. The output is:

$$M_o = (\sigma(R_c) \odot V_c) W_{LN}^o + M', \tag{10}$$

with W_{LN}^o as the output projection.

4.3 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 is designed to capture long-range inter-frame dependencies without relying on explicit motion estimation. Unlike recurrent units that accumulate errors over time or transformers that incur quadratic complexity, RWKV provides a linear-time sequence model with strong memory retention and efficient parallelization.

Concretely, we first reorganize refined spatial features into spatio-temporal tubelets, which serve as tokens encoding both local spatial context and short-term dynamics. These tokens are then fed into the RWKV-based temporal mix layer, which leverages recurrent gating and attention-inspired weighting to integrate information across frames. Importantly, RWKV maintains a memory state that scales with sequence length, enabling the model to process long endoscopic videos without truncation. This property is crucial for surgical workflows, where sequences often span tens of thousands of frames.

The temporal module is also augmented with DGW-Shift, extending its adaptive kernel selection into the temporal domain. By dynamically adjusting temporal filters according to motion cues, DGW-Shift allows the block to suppress inconsistent artifacts caused by rapid camera movement, occlusions, or tissue deformation. This design yields robustness to highly non-rigid dynamics and ensures that subtle temporal correlations, such as gradual appearance changes or periodic motions, are faithfully reconstructed.

The outputs of the spatial and temporal RWKV modules are fused to form a joint representation that balances high-frequency spatial detail with temporally consistent context. Residual connections preserve low-level fidelity, while RWKV attention captures long-range spatio-temporal correlations. The fused features are progressively upsampled using learnable reconstruction blocks to generate the high-resolution 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 not only reduces reliance on optical flow or deformable convolutions, which are prone to failure in endoscopic settings, but also scales efficiently to long surgical procedures, making it well-suited for real-world medical deployment.

4.4 Loss Functions and Training Protocol

We train the model using a combination of pixel-wise reconstruction loss and perceptual loss. The primary objective is the Charbonnier loss, $\mathcal{L} = \sqrt{\|I'_t - I_t^{HR}\|^2 + \epsilon^2}$, where I'_t is the reconstructed HR frame and I_t^{HR} is the ground truth. Perceptual loss is optionally added to encourage preservation of clinically relevant textures. Training is performed on synthetic endoscopic video datasets with realistic degradations, using data augmentation to simulate domain variability.

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 (3)	31.46	0.899
BasicVSR++ (4)	31.73	**0.904**
RVRT (10)	29.26	0.894
TCNet (13)	31.11	0.889
IART (23)	31.30	0.903
EndoNet (Ours)	**31.89**	0.899

5 Experiments

To evaluate the effectiveness of linear attention mechanisms for endoscopic video super-resolution, we conduct controlled experiments on the HyperKvasir dataset (1). Our evaluation quantifies improvements in reconstruction quality, training stability, and generalization, following established protocols in the medical video super-resolution literature.

5.1 Implementation Details

Models are implemented in PyTorch (17) and trained from scratch using the AdamW optimizer (14) 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 (9) 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.

5.2 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.

5.3 Quantitative Comparison

Table 1 summarizes the quantitative results of EndoNet and several state-of-the-art baselines, including BasicVSR (3), BasicVSR++ (4), RVRT (10), TCNet (13), and IART (23). 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.

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

5.4 Ablation Studies

5.4.1 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. 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.

5.4.2 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.

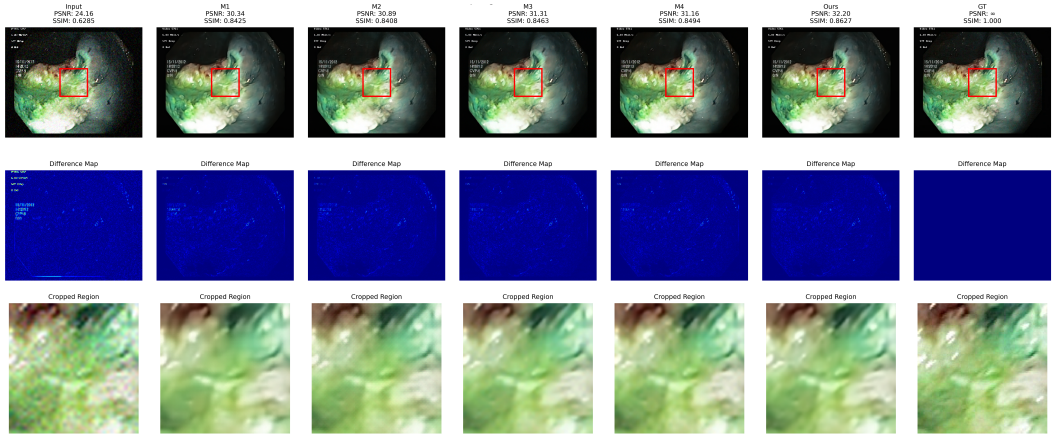

Figure 1: Visual comparisons of results produced by our ablation study on video frames from the HyperKvasir dataset. (Zoom in for more details)

6 Discussion

6.1 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.

6.2 Discussion on Societal Impacts

On the positive side, endoscopic video super-resolution can enhance the visual quality of surgical recordings, providing surgeons with clearer views of fine anatomical structures and potentially improving diagnostic accuracy, surgical safety, and training quality. Such advancements may contribute to better patient outcomes and facilitate knowledge transfer in minimally invasive procedures. On the other hand, we acknowledge possible negative impacts. Improved visual clarity may lead to over-reliance on AI-enhanced images, which could obscure the limitations of the original acquisition hardware. There is also a risk that misuse of enhanced medical videos outside proper clinical or regulatory contexts could cause misinterpretation of findings. Furthermore, privacy concerns must be carefully managed when handling surgical video data. To mitigate these risks, we emphasize that our framework is intended as a decision-support tool rather than a replacement for medical expertise, and we advocate for integration with clinical validation and ethical guidelines.

7 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

[1] Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 283 (2020)

[2] Cao, J., Li, Y., Zhang, K., Gool, L.V.: Video super-resolution transformer. arXiv (Cornell University) (2021)

[3] Chan, K.C.K., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

[4] Chan, K.C.K., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[5] Duan, Y., Wang, W., Chen, Z., Zhu, X., Lu, L., Lu, T., Qiao, Y., Li, H., Dai, J., Wang, W.: Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. arXiv preprint arXiv:2403.02308 (2024)

[6] Fei, Z., Fan, M., Yu, C., Li, D., Huang, J.: Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. arXiv preprint arXiv:2404.04478 (2024)

[7] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT Press (2016)

[8] He, Q., Zhang, J., Peng, J., He, H., Wang, Y., Wang, C.: Pointrwkv: Efficient rwkv-like model for hierarchical point cloud learning. arXiv preprint arXiv:2405.15214 (2024)

[9] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[10] 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. arXiv (Cornell University) (2022)

[11] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Laak, J.v.d., Ginneken, B.v., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis (2017)

[12] Liu, H., Ruan, Z., Zhao, P., Dong, C., Shang, F., Liu, Y., Yang, L., Timofte, R.: Video super-resolution based on deep learning: a comprehensive survey. Artificial Intelligence Review (2022)

[13] Liu, M., Jin, S., Yao, C., Lin, C., Zhao, Y.: Temporal consistency learning of inter-frames for video super-resolution. IEEE Transactions on Circuits and Systems for Video Technology **33**(4), 1507–1520 (2022)

[14] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[15] Lou, M., Zhang, S., Zhou, H.Y., Yang, S., Wu, C., Yu, Y.: 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)

[16] Min, J.K., Kwak, M.S., Myung, J.: Overview of deep learning in gastrointestinal endoscopy. Gut and Liver (2019)

[17] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[18] Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K.K., et al.: Rwkv: Reinventing rnns for the transformer era. arXiv preprint arXiv:2305.13048 (2023)

[19] Peng, B., Goldstein, D., Anthony, Q., Albalak, A., Alcaide, E., Biderman, S., Cheah, E., Ferdinan, T., Hou, H., Kazienko, P., et al.: Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. arXiv preprint arXiv:2404.05892 (2024)

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[21] Wang, X., Wang, H., Zhang, M., Zhang, F.: Combining optical flow and swin transformer for space-time video super-resolution. Engineering Applications of Artificial Intelligence (2024)

[22] Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C.: Edvr: Video restoration with enhanced deformable convolutional networks. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)

[23] Xu, K., Yu, Z., Wang, X., Mi, M.B., Yao, A.: 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)

[24] Yuan, H., Li, X., Qi, L., Zhang, T., Yang, M.H., Yan, S., Loy, C.C.: Mamba or rwkv: Exploring high-quality and high-efficiency segment anything model. arXiv preprint arXiv:2406.19369 (2024)

Agents4Science AI Involvement Checklist

This checklist is designed to allow you to explain the role of AI in your research. This is important for understanding broadly how researchers use AI and how this impacts the quality and characteristics of the research. **Do not remove the checklist! Papers not including the checklist will be desk rejected.** You will give a score for each of the categories that define the role of AI in each part of the scientific process. The scores are as follows:

- **[A] Human-generated**: Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI**: The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human**: The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated**: AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

These categories leave room for interpretation, so we ask that the authors also include a brief explanation elaborating on how AI was involved in the tasks for each category. Please keep your explanation to less than 150 words.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "Agents4Science AI Involvement Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[C]**

   Explanation: Researchers have assigned AI to complete an endoscopic video super-resolution task. AI generated the idea for this research follow-up.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: The researchers provided a code template for the experiment. AI has further improved and optimized the structure of deep learning networks based on the proposed ideas. Researchers need to participate to some extent in the experimental operation.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[C]**

   Explanation: The researchers provided the dataset for this study. AI automatically processed and analyzed the experimental results.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[D]**

   Explanation: The paper is generated by AI.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: It is difficult for AI to truly generate innovative ideas for complex algorithm design. AI cannot provide accurate network diagrams.

Agents4Science Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **Papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers and area chairs. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given. In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "Agents4Science Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately present our contribution.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes discussion of both limitations in the analysis section.

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The article describes the implementation of our method in the methodology section.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental details, hyperparameters, and methodology are disclosed in 547 sufficient detail for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and data will be made available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental settings are detailed in the methodology section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results include error bars and statistical significance testing across multiple experimental runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. They will be in the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: The research adheres to all ethical guidelines specified by the Agents4Science conference.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discussed the potential positive societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.