G-021

Recurrent Route-Map Based Transformer for Endoscopic Video Super-Resolution

ヨ シンキ† 鎌田 清一郎†
Chenxi Yu Sei-ichiro Kamata

1. Abstract

Video super-resolution, as a fundamental computer vision task, is widely used in various fields. Particularly, in the field of endoscopic, Endoscopic video restoration is conducive for doctors to observe unclear areas or tiny lesions for making accurate diagnosis. However, existing endoscopic video super-resolution methods can only utilize limited frames, and it is difficult to utilize the information of frames that are far away, and the effective use of information cannot be guaranteed. Therefore, in this paper, we propose a route-map mechanism to guide the information flow within the transformer architecture, enabling the model to focus on relevant regions and features crucial for super-resolution. The route-map can record the relative position information of all frames. By dynamically updating the route-map based on the content of each frame, our model can capture more efficient features, thereby improving the quality of the results. Experimental results demonstrate the proposed method outperforms the state-of-the-art methods considered in this work.

2. Introduction

The objective of video super-resolution is to reconstruct a high-resolution video from a low-resolution video at a specified magnification scale [1]. Endoscopic video super-resolution, a fundamental task in medical imaging, has garnered significant attention from both industry and academia for applications such as tumor detection, colorectal polyp detection, and quality assessment of individual surgeon skills [4]. However, achieving high-quality endoscopic videos is challenging due to practical hardware limitations (e.g., capsule endoscopy cameras [7]) and uncontrolled endoscope movements, which can substantially degrade diagnostic performance. Therefore, investigating endoscopic video super-resolution is crucial in the current clinical process. However, some current VSR networks usually lack long-term modeling capability due to vanishing gradient [14], which inevitably leads to some unsatisfactory results.

Inspired by the recent advancements of Transformer in natural language processing [15], significant progress has been achieved in both visual recognition [16] and generation tasks [17]. A Transformer comprises a self-attention module and feed-forward networks [15]. In the vision Transformer method [18], a single image is segmented into multiple patches, which are then used as sequence data input to the Transformer. However, exploring proper ways of utilizing Transformers in endoscopic videos remains a big challenge.

To solve this problem and obtain high-quality endoscopic videos, we propose a Route-Map Based Transformer (RRMBT) to facilitate effective video representation learning for Endoscopic Video Super-Resolution. The main contributions of this paper are as follows:

- We propose a route-map to calculate the visual tokens in each endoscopic video frame and calculate Q, K, V in the same map. Once the route-map is learned, RRMBT only calculates self-attention for the most relevant visual tokens in the same map.
- We develop multi-scale temporal features to obtain different route-maps and select the best features to calculate the results.
- Extensive experiments demonstrate that our proposed method can achieve state-of-the-art performance in the endoscopic video super-resolution task.

3. Related Work

3.1 Video Super-Resolution

Existing video super resolution approaches can be classified into two types: sliding-window and recurrent.

Sliding-window structure. The methods based on the sliding window structure use adjacent frames within a sliding window as inputs to recover the high-resolution frame (e.g., 5 or 7 frames) [19]-[21]. EDVR [22] used deformable convolutions [3] to align adjacent frames. However, these methods cannot utilize textures from other moments, especially from relatively distant frames.

Recurrent structure: Instead of aggregating information from adjacent frames, methods based on recurrent structures use a hidden state to convey relevant information from previous frames [2]. OVSR [6], BasicVSR [12], IconVSR [12] and BasicVSR++ [13] fused the bidirectional hidden state from past and future frames for reconstruction, achieving significant improvements. However, due to the vanishing gradient problem [14], this mechanism causes the updated hidden state to lose some long-term modeling capabilities.

3.2 Endoscopic Video Super-Resolution

Recently, numerous deep learning-based approaches have been employed in endoscopic video super-resolution [23]. For example, MABPN [23] introduced a multilevel error feedback mechanism to recover details like texture and edges in frames. Additionally, endoscopic video super-resolution necessitates high-magnification reconstruction for identifying minute lesions, which demands precise details in images.

† Chenxi Yu and Sei-ichiro Kamata are with the Image Media Laboratory, Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 8080135, Japan.

3.3 Vision Transformer

Transformer [15] has become widely utilized in natural language processing due to its ability to model long-term sequences. In recent years, transformer-based architectures have also been applied to vision tasks and have shown remarkable performance, including image classification [25] and image restoration [11]. In the domain of video restoration tasks, VSR-T [10] presented a transformer-based model for VSR for the first time, achieving comparable results with CNN-based methods. However, it fails to exploit local information within each patch during the global self-attention process. To address this issue, the approach outlined in VRT [9] proposed a mutual attention mechanism to enhance temporal receptive fields and reduce computational complexity.

4. Proposed Method

4.1 Route-Map

In this paper, route-map represents the position of abstract object features in each frame as show in Figure. 1. For moving objects in endoscopy videos, the route-map records its position information in each frame. The route-map can locate the position information of objects in any frame so that we can use their features more effectively. It plays a huge role in alignment and feature utilization.

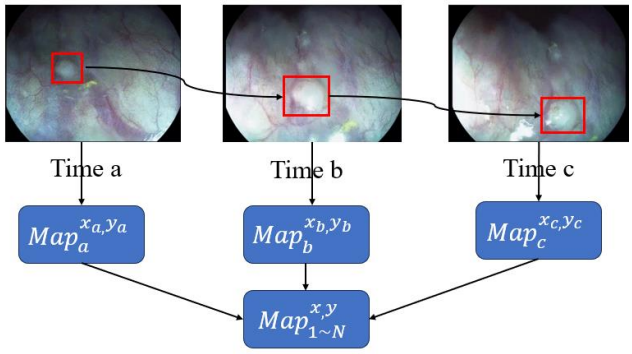

Figure 1. What is the route-map. The figure shows three frames of an endoscope video. The object in the red box changes as the camera moves. In different frames, its size and position are different, but it is still the same object. The route-map is a map that records the spatial information of the object in different frames.

So how to get the route-map? As shown in Figure 2, to address this problem, we propose a position map that is represented as a set of matrices over time. With this design, position generation can be represented as some matrix operations. Specifically, we try to use optical flow [8] to update the map. We can easily get the optical flow between adjacent frames. Then, we can convert the positional relationship between adjacent frames based on the optical flow. Then, based on the positional relationship between the current adjacent frames and the optical flow of the previous frame, we can convert the positional relationship between the previous frame and the current frame. When $M_t^{x,y}$ represents the position coordinates ending at (x, y) at time t, I_T^{LQ} represents current input frame, we can get the route transition graph:

$$M\left(I_T^{LQ} \cdot t\right) = M_t^{x,y} \quad t \in [1, N] \tag{1}$$

Where x ∈ [1, W] and y ∈ [1, H], N represents the total number of frames. This formula expresses the relative position of the information of the current frame at time t. And so on, we can get the positional relationship between the current frame and all frames, and thus generate a route-map containing the positional relationship.

4.2 Temporal Segmentation

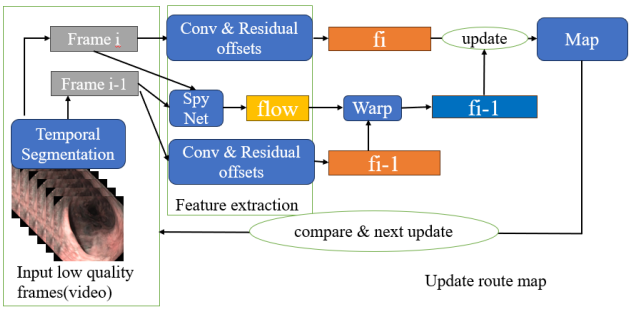

Figure 2. How to get the route-map. The input is an endoscopic video sequence. The output is a map that records relative position information. CONV and Residual offsets are used to extract features. SpyNet is used to get optical flow. Map refers to the route-map.

We try to segment the endoscopic video sequence to produce several video segments with different frame rates. For example, every third segment (1, 4, 7, 10...), every second segment (1, 3, 5, 7...), and no segment (1, 2, 3, 4...). As show in Figure 2, note that the reference frame appears in each segmentation. It is worth noting that our method can be easily generalized to arbitrary frames as input. Temporal segmentation enables the effective integration of neighboring frames with varying temporal distances. It not only can get contributions from these frames differ, especially with large deformations, occlusions, and motion blur, but also enhance information extraction and fusion. Temporal segmentation facilitates the extraction of features at different temporal frequencies, leading to higher quality restored endoscopic videos.

4.3 Overall Architecture

Given a low-quality (LQ) input video sequence, the goal of endoscopic video super-resolution is to recover a output high-quality (HQ) sequence. In this sequel, we denote $I^{LQ} \in \mathbb{R}^{T*C*H*W}$ as the input LQ sequence and $I^{HQ} \in \mathbb{R}^{T*C*H*W}$ as the output HQ sequence, where T, C, H, and W represents the frame number, channel dimensions, height, and width respectively. As illustrated in Figure. 3, we propose a route-map based transformer with the recurrent framework.

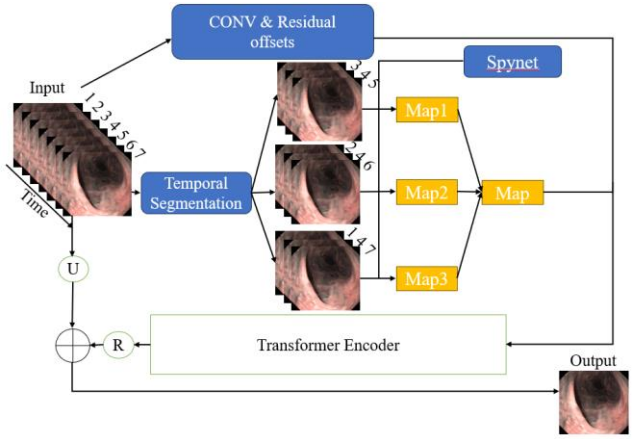

Figure 3. The overview of RRMBT. The input is a low-resolution endoscopic video sequence. The output is a generated high-resolution endoscopic video sequence. CONV and Residual offsets are used to extract features. Temporal Segmentation module segments frames in temporal range. Map refers to the route-map generated by the model. Transformer encoder means using vision transformer. R(·) represents the reconstruction network followed by a pixel-shuffle layer to resize feature maps to the desired size. U(·) represents the bicubic upsampling operation. ⊕ indicate element-wise addition.

For our task, when restoring the Tth frame I_T^{LQ}, we denote the current LQ frame as I_T^{LQ} and other LQ frames as $I^{LQ} = \{I_t^{LQ}\ t \in [1,N]\}$. The map M in our approach can be formulated as a set of maps, in which each map m^i is a sequence of coordinate over time and the end point of m^i is associated with the coordinate of token q^i:

$$M = \{AVG(m^i), i \in [1,N]\}$$
$$m^i = \{m_t^i = (x_t^i, y_t^i), t \in [1,N]\} \quad (2)$$

Among them, $x_t^i, \in [1,W]$, $y_t^i, \in [1,H]$, and (x_t^i, y_t^i) represents the coordinate of map m^i at time t. H and W represents the height and width of the feature maps, respectively. Since Temporal Segmentation is used, AVG operation is needed to fuse the maps at different time frequencies.

From the aspect of maps, the inputs of proposed route-map based transformer can be further represented as visual tokens which are aligned by map M:

$$Q = \left\{q_T^{m^i}, i \in [1,N]\right\}$$
$$K = \left\{k_t^{m^i}, i \in [1,N], t \in [1,N]\right\}$$
$$V = \left\{v_t^{m^i}, i \in [1,N], t \in [1,N]\right\} \quad (3)$$

By using map in the Transformer, it can avoid computations on the spatial dimensions and thus can significantly reduce the attention computations on Q, K and V compared to vision transformers.

The process of recovering the Tth HQ frame I_T^{HQ} can be further expressed as:

$$I_T^{HQ} = R\big(M(Q,K,V)\big) + U(I_T^{LQ}) \quad (4)$$

Among them, M(·) denotes route-map based transformer. R(·) denotes the reconstruction network followed by a pixel shuffle layer to resize the feature map to the desired size. U(·) denotes the bicubic upsampling operation.

5. Experiments

5.1 Datasets and Metrics

In the field of medical imaging [23], it is challenging to obtain corresponding datasets due to the difficulty of data acquisition. Therefore, we constructed an endoscopic video super-resolution dataset based on OBRDataset [5], which contains in vivo sequences and ground truth data from endoscopy. These videos are collected during standard gastrointestinal examinations, involving challenges such as tissue deformation and rapid endoscope motion. For fair comparison, we divide the dataset into 27 video sequences, 23 for training and 4 for testing. Each sequence contains 300 frames with a resolution of 640×480. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are selected as evaluation metrics.

5.2 Implementation Details

For fair comparison, we use the pre-trained SpyNet [8] as the motion estimation module. The number of channels in our model is set to 64. During training, we use the Adam optimizer with β1 = 0.9, β2 = 0.99, and a cosine annealing scheme. The initial learning rate of the motion estimation module is 2.5 × 10−5, and the rest of the model is set to 2 × 10−4. We set the batch size to 8 and the input LR frame to 64×64. The total number of iterations is 200K, and we fix the weights in the motion estimation module in the first 5K iterations to stabilize the training process. All experiments are conducted on a server with PyTorch 1.8.0 and a single NVIDIA GeForce RTX 3090 GPU.

5.3 Comparisons with State-of-the-art Methods

Method	PSNR	SSIM
Bicubic	31.37	0.8463
EDVR [22]	36.74	0.9332
BasicVSR [12]	36.62	0.9313
IconVSR [12]	36.71	0.9328
VRT-T [10]	36.77	0.9324
VRT [9]	36.96	0.9348
BaiscVSR++ [13]	37.18	0.9358
RRMBT	37.23	0.9368

Table 1. Quantitative comparison (PSNR (dB) and SSIM) on OBRDataset. Red text indicates the best and blue text indicates the second best performance.

We compare our model with 6 state-of-the-art methods listed in Table 1. Our proposed RRMBT outperforms VRT-T by 0.46 dB in PSNR and 0.0044 in SSIM. RRMBT outperforms BasicVSR++, another state-of-the-art loop-based method, by 0.05 dB in PSNR and 0.0010 in SSIM.

The visual qualitative results are shown in Figure 4. By comparing the magnified results, the HR outputs of EDVR and BasicVSR++ are blurred, and the blood vessel edges are not clear enough. Our proposed method can generate

finer edge contour information, and the restored texture details are more consistent with the ground truth data.

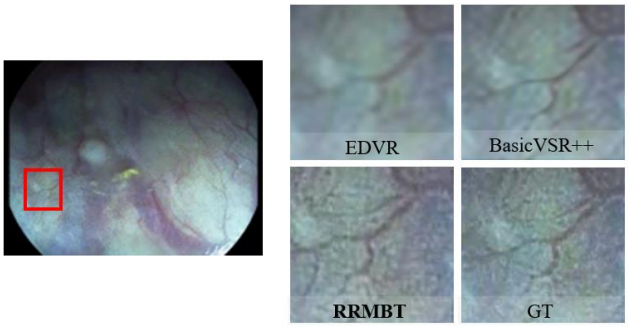

Figure 4. Visual comparison on OBRDataset. Zoom in for best view.

5.4 Ablation Study

We conducted an ablation study on the proposed RRMBT model for endoscopic video super-resolution. By removing the route-map and temporal segmentation, we directly use a stack of normal self-attention layers in the "base" model. As shown in Table 2, by adding the Route-Map (Baseline + RM), the PSNR can be improved from 36.72 dB to 37.13 dB. Adding the Temporal Segmentation on Baseline + RM basis, the PSNR can be improved from 37.13 dB to 37.23 dB.

Method	RM	TS	PSNR	SSIM
Baseline			36.72	0.9331
Base + RM	✓		37.13	0.9355
Base +RM + TS	✓	✓	37.23	0.9368

Table 2. Ablation study results on OBRDataset.

6. Conclusion

In this paper, we study endoscopic video super-resolution by exploiting long-range frame dependencies. Specifically, we propose a route-map based Transformer architecture. We record the positions of abstract objects in the video and record it in a map. And use this map to do some alignment operations and to exploit more features. We also tried to use route-map at different frame rates and obtained high-quality endoscopic video sequences. Experimental results show that this method is effective for improving the quality of endoscopic videos. In the future, we will focus on 1) trying to evaluate our method on more endoscopic video datasets, and 2) optimizing the performance of the model.

Acknowledgement

The authors declare no conflict of interest. Chenxi Yu: Conceptualization, Methodology, Research, Experiment, Validation. Sei-ichiro Kamata: Conceptualization, Research, Validation. I would like to express my great appreciation to my supervisor Sei-ichiro Kamata for his consecutive suggestions.

References

[1] Y. Luo, L. Zhou, S. Wang, “Video satellite imagery super resolution via convolutional neural networks,” IEEE Geosci. Remote. Sens Lett., Vol.n 14, No.n 12 (2017).
[2] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, “An image is worth 16×16 words: Transformers for image recognition at scale,” ICLR (2021).
[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, “Deformable convolutional networks.” ICCV (2017).
[4] S. A. Karkanis, D. K. Iakovidis, “Computer-aided tumor detection in endoscopic video using color wavelet features,” IEEE Trans. Inf. Technol. Biomed., Vol.n 7, No.n 3 (2003).
[5] M. Ye, S. Giannarou, A. Meining, “Online Tracking and Retargeting with Applications to Optical Biopsy in Gastrointestinal Endoscopic Examinations,” Med. Image Anal., Vol.n 30 (2016).
[6] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo, “Learning texture transformer network for image super-resolution.” CVPR (2020).
[7] J. Park, Y. Hwang, J. H. Yoon, “Recent development of computer vision technology to improve capsule endoscopy,” Clin. Endosc., Vol.n 52, No.n. 4 (2019).
[8] A. Ranjan, M. J. Black, "Optical Flow Estimation Using a Spatial Pyramid Network," CVPR (2017).
[9] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, and Y. Li, “VRT: A video restoration transformer,” arXiv (2022).
[10] J. Cao, Y. Li, K. Zhang, and Van Gool, “Video super-resolution transformer,” arXiv (2021).
[11] J. Liang, J. Cao, G. Sun, and K. Zhang, “Swinir: Image restoration using swin transformer,” CVPR (2021).
[12] K. C. K. Chan, X. Wang, K. Yu, “BasicVSR: The search for essential components in video super-resolution and beyond,” CVPR (2021).
[13] K. C. K. Chan, S. Zhou, X. Xu, C. C. Loy, "BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment," CVPR (2022).
[14] Sepp Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, Vol.n 6, No.n 2 (1998).
[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, “Attention is all you need.” NeurIPS. Vol.n 30 (2017).
[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, “Endto-end object detection with transformers.” ECCV (2020).
[17] Fuzhi Yang, Huan Yang, “Learning texture transformer network for image super-resolution.” CVPR (2020).
[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv (2020).
[19] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Shengjin Wang, and Qi Tian, “Video super-resolution with temporal group attention.” CVPR (2020).
[20] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and MingHsuan Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution.” CVPR (2017).
[21] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf, “Spatio-temporal transformer network for video restoration.” ECCV (2018).
[22] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy, “EDVR: Video restoration with enhanced deformable convolutional networks.” CVPRW (2019).
[23] S. Ren, J. Li, K. Guo, and F. Li, “Medical video super-resolution based on asymmetric back-projection network with multilevel error feedback,” IEEE Access, Vol.n 9 (2021).