LGRNet: Local-Global Reciprocal Network for Uterine Fibroid Segmentation in Ultrasound Videos

Huihui Xu[1*], Yijun Yang[1*], Angelica I Aviles-Rivero[3], Guang Yang[4], Jing Qin[5], and Lei Zhu[1,2](✉)

[1] The Hong Kong University of Science and Technology (Guangzhou), China
leizhu@ust.hk
[2] The Hong Kong University of Science and Technology, China
[3] University of Cambridge, UK
[4] Imperial College London, UK
[5] The Hong Kong Polytechnic University, China

Abstract. Regular screening and early discovery of uterine fibroid are crucial for preventing potential malignant transformations and ensuring timely, life-saving interventions. To this end, we collect and annotate the first ultrasound video dataset with 100 videos for uterine fibroid segmentation (UFUV). We also present Local-Global Reciprocal Network (LGRNet) to efficiently and effectively propagate the long-term temporal context which is crucial to help distinguish between uninformative noisy surrounding tissues and target lesion regions. Specifically, the Cyclic Neighborhood Propagation (CNP) is introduced to propagate the inter-frame local temporal context in a cyclic manner. Moreover, to aggregate global temporal context, we first condense each frame into a set of frame bottleneck queries and devise Hilbert Selective Scan (HilbertSS) to both efficiently path connect each frame and preserve the locality bias. A distribute layer is then utilized to disseminate back the global context for reciprocal refinement. Extensive experiments on UFUV and three public Video Polyp Segmentation (VPS) datasets demonstrate consistent improvements compared to state-of-the-art segmentation methods, indicating the effectiveness and versatility of LGRNet. Code, checkpoints, and dataset are available at `https://github.com/bio-mlhui/LGRNet`

Keywords: Uterine Fibroid Segmentation · Ultrasound Videos · State Space Model · Video Polyp Segmentation

1 Introduction

Uterine fibroids are the most common benign tumors in the female genital tract, with approximately 70% of women at risk of experiencing such diseases throughout their lifetime [17]. Consequently, regular screening and early detection of uterine fibroids are essential for initiating timely life-saving treatments. Since

* Equal contributions.

CT and MRI examinations are expensive and harmful to human bodies, ultrasound is becoming a more popular imaging modality for clinical diagnosis. Recently, automatic ultrasound detection and segmentation in videos have attracted much attention from the medical community [15,14,25,26]. For example, FLA-Net[15] presents a frequency and location feature aggregation network for ultrasound video breast lesion segmentation. UltraDet[26] proposes to aggregate the negative temporal context to facilitate filtering out false positive predictions in ultrasound video breast lesion detection. However, automatic uterine fibroid segmentation in ultrasound videos remains unexplored. Moreover, ultrasound video segmentation is challenged by several factors including noisy temporal motions, blurry boundaries, and changing lesion size over time.

In this paper, **1) we collect the first ultrasound video dataset for uterine fibroid segmentation (UFUV)**, which contains 100 videos with per-frame expert annotations. To handle aforementioned challenges in ultrasound segmentation, **2) we present Local-Global Reciprocal Net (LGRNet)** to efficiently and effectively aggregate the global temporal context using a set of frame bottleneck queries. In LGRNet, **3) we incorporate Cyclic Neighborhood Propagation (CNP) and Hilbert Selective Scan (HilbertSS)** which reciprocally propagate the crucial local-global temporal context through the bottleneck queries. **4)** We conduct extensive experiments to demonstrate that LGRNet can both quantitatively and qualitatively outperform state-of-the-art segmentation methods on UFUV and three publicly available Video Polyp Segmentation datasets.

2 Method

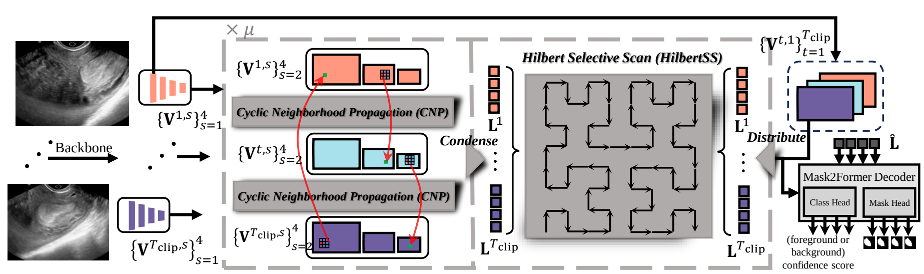

Fig. 1. LGRNet architecture. We introduce the notion of *frame bottleneck queries* which efficiently condense the global temporal context and distribute global context back for reciprocal local refinement.

As shown in Fig.1, given a video clip $\{V^t\}_{t=1}^{T_{\text{clip}}}$, we first devise a backbone to extract its per-frame multi-scale features $\{\{\mathbf{V}^{t,s} \in \mathcal{R}^{H_s W_s \times c}\}_{s=1}^{S}\}_{t=1}^{T}$,

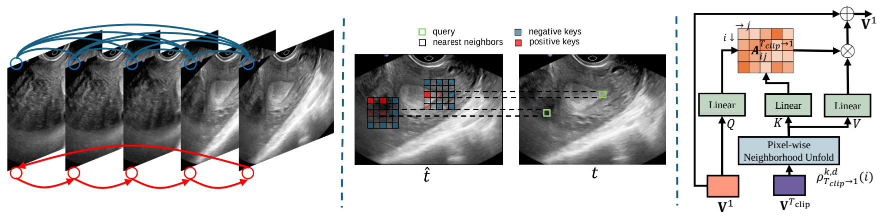

Fig. 2. Illustrations of CNP. (Left) We enforce local cyclic inter-frame dependencies (red), instead of estabaling fully connected connections (blue). (Middle) In CNP, each query point takes the corresponding nearest neighbors of the cyclic frame as attention keys. (Right) Detailed Implementation of CNP.

where c and H_sW_s are dimension and resolution of the s-th scale. Each scale is transformed into the common dimension c by a non-biased Conv2D with GroupNorm[24] layer. Then, clip features of the last three scales are input to **Cyclic Neighborhood Propagation** to propagate local inter-frame motion context in a cyclic manner. CNP is executed for each scale, with all scales sharing the same CNP parameters. Next, for each frame, the multi-scale features are input to a *Condense* layer and compressed into a short sequence of **frame bottleneck queries $\mathbf{L}^t$**. Bottleneck queries of all frames, i.e. $\{\mathbf{L}^t\}_{t=1}^{T_{\text{clip}}}$, are input to **Hilbert Selective Scan** to efficiently path connect all frames. Then, for each frame, global-view queries are input to a *Distribute* layer to disseminate the global temporal context back to multi-scale features. Finally, the reciprocally encoded multi-scale features are input to a Mask2former [4] decoder for foreground/background classification and mask prediction. LGRNet can output a set of different mask predictions each with a foreground lesion confidence score.

2.1 Local Cyclic Neighborhood Propagation (CNP)

Local CNP. Motion priories, such as optical flow [26], can be utilized as pixel-wise guidance to propagate inter-frame temporal information. However, they incorporate additional parameters of pretrained optical flow predictor [6] and may not generalize to ultrasound videos due to noisy and monochromatic color change. Recently, motivated by introducing locality inductive biases to vanilla attention mechanism, Neighborhood Attention (NA) [11] demonstrates that only involving nearest neighbors as attention keys can achieve comparable performance on image tasks. In this paper, we interpret inter-frame locality inductive biases as motion guidance and adapt NA to videos. We propose the Cyclic Neighborhood Propagation. CNP is executed for each scale and all scales share the same CNP parameters. We omit scale superscript s in this subsection for simplicity. As shown in Fig.2, for a query point $\mathbf{q}_i^t = \mathbf{W}_Q\mathbf{V}_i^t \in \mathcal{R}^c$ at i-th position of t-th frame, CNP aggregates local motion information from frame $\hat{t}$ to frame

t as:

$$CNP(\mathbf{q}_i^t, \mathbf{V}^{\hat{t}}) = \sum_{m=1}^{M} \mathbf{W}_m \sum_{j \in \rho_{\hat{t}\to t}^{k,d}(i)} A_{ij}^{\hat{t}\to t} \mathbf{W}_V \mathbf{V}_j^{\hat{t}}, \tag{1}$$

$$\hat{t} = \begin{cases} t-1, & t > 1 \\ T, & t = 1 \end{cases} \tag{2}$$

where M denotes the number of attention heads, $\rho_{\hat{t}\to t}^{k,d}(i)$ is the set of nearest neighbors w.r.t $\mathbf{V}_i^t$ at $\hat{t}$-th frame, and the nearest neighbors are defined by a kernel with size k and dilation d. $\{A_{ij}^{\hat{t}\to t}\}_{j=1}^{|\rho_{\hat{t}\to t}^{k,d}(i)|}$ denotes the attention weights, which are the normalized dot product of query with each neighbor key, i.e. $\text{softmax}_j(\frac{\mathbf{V}_i^{t\,T}\mathbf{W}_Q^T\mathbf{W}_K\mathbf{V}_j^{\hat{t}}}{\sqrt{c}})$. Since CNP is applied in each encoder layer, it enables the local inter-frame temporal information to circulate within the clip. As shown in Fig.2, we do not build fully connected inter-frame dependencies, since when motion changes severely and is noisy, dense connections would connect a query point to uninformative background tokens at distant frames with weak semantics, which also leads to increased computation.

2.2 Global Hilbert Selective Scan (HilbertSS)

Frame Queries as Information Bottleneck. Radiologists often need longer temporal context [19] to not only decide whether a possible region is lesion or not, but also refine their local per-frame predictions. Inspired by this behavior, we devise a set of learnable **frame bottleneck queries** $\mathbf{L} \in \mathcal{R}^{\bar{N}\times c}$ to first summarize each frame into a query sequence:

$$\mathbf{L}^t = Condense(\text{query} = \mathbf{L}, \text{key} = \{\mathbf{V}^{t,s}\}_{s=1}^{S}), t = 1, ..., T_{\text{clip}}. \tag{3}$$

Frame queries can be seen as bottlenecks extracting semantically rich lesion information from each frame and facilitating later efficient global information exchange.

Global HilbertSS. Recently, Selective State Space Model (S6)[10] proposes a new sequence transform model with linear complexity. Formally, each S6 block transforms input sequence $x \in R^{L\times c}$ to $y \in R^{L\times c}$ as, where $A \in R^{c\times c_i}$, $B = \text{Linear}_{c_i}(x)$, $C = \text{Linear}_{c_i}(x)$, $\Delta = \text{softplus}(\text{Bias}_c + \text{broadcast}_c(\text{Linear}_1(x)))$:

$$\begin{gathered} h_k = \bar{A}h_{k-1} + \bar{B}x_k, y_k = Ch_k \\ \bar{A} = \exp(\Delta A), \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I)\cdot \Delta B. \end{gathered} \tag{4}$$

Linear_d is a linear projection to dimension d. Bias_d is a bias vector of dimension d. Similar to the attention mechanism, weights of S6 are *input-dependent*, which facilitates context modelling.

To apply S6 to 2D input, a direct approach is flattening 2D input into a sequence using a Zigzag curve just like attention-based models. However, Eq.4

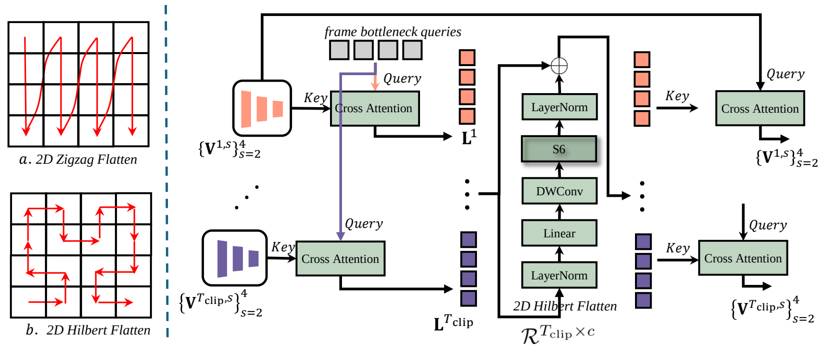

Fig. 3. Illustrations of *HilbertSS*. (Left) Comparison between (a) Zigzag Flatten (b) Hilbert Flatten. (Right) Detailed Implementation of global part of LGRNet.

implies S6 is position-aware. As an intuitive 2D example shown in Fig.3, Zigzag flattening may corrupt the locality structure. More formally, any flattening curve can be generalized to a *Space Filling Curve* (SFC) σ, which maps point x in [0, 1] to $\sigma(x)$ on [0,1]× [0,1]. *Dilation Factor* of a space-filling curve is defined as the upper bound of $\frac{|\sigma(x)-\sigma(y)|^2}{|x-y|}$. As proved in [1][3], the dilation factor of Hilbert curve is 6, while the Zigzag curve is $4^n - 2^{n+1} + 2$, which diverges to ∞ when curve order n is bigger. This shows that Hilbert curve preserves the 2D locality structure, which accords with the intuition that lesions of close frames should be scanned in groups and tracked together.

As shown in Fig.3, the scan order is a Hilbert curve on the 2D $\bar{N} \times T_{\text{clip}}$ grid. We use the implementation of [27]. In all, we have:

$$\{\mathbf{L}^t\}_{t=1}^{T_{\text{clip}}} = S6(Hilbert - Flatten(\{\mathbf{L}^t\}_{t=1}^{T_{\text{clip}}})) \tag{5}$$

Reciprocal Local-Global Refinement Radiologists may use global view to refine their per-frame predictions. We use a *Distribute* layer to distribute the global temporal context back to the multi-scale features for each frame:

$$\{\mathbf{V}^{t,s}\}_{s=1}^{S} = Distribute(\text{query} = \{\mathbf{V}^{t,s}\}_{s=1}^{S}, \text{key} = \mathbf{L}^t), t = 1, ..., T_{\text{clip}}. \tag{6}$$

Both *Condense* and *Distribute* are implemented as a cross attention layer.

2.3 Decoder

Our decoder uses the same architecture with Mask2Former[4]. A set of learnable *temporal queries* $\hat{\mathbf{L}} \in \mathcal{R}^{\hat{N}\times c}$ are used to cross-attend different scale features at each cross attention layer, where the query length is $\hat{N}$ and the key length is $T_{\text{clip}} \times H_s \times W_s$. The final masks are the dynamic convolution between the stride 4 scale (s=1) and the temporal queries. The bipartite matching loss is composed of classification (foreground/background) cross-entropy loss, mask dice loss and binary cross entropy loss: $\lambda_{class}L_{class}+\lambda_{dice}L_{dice}+\lambda_{ce}L_{ce}$. Our model can generate multiple, i.e. $\hat{N}$, different predictions, each with a lesion confidence score. We choose the mask with highest foreground score to compare with other methods.

3 Experiments

Dataset. We collect and annotate the first ultrasound video uterine fibroid segmentation dataset (UFUV). Our UFUV dataset contains 100 videos and each video has 50 frames. The ultrasound videos were collected using Mindray Resona 8 and Supersonic Alxplorer. The dataset encompasses a cohort of female subjects aged between 20 to 45 years. We chose video that showcases at least one clearly delineated hypoechoic region (indicative of a fibroid) within the uterine wall, with a diameter exceeding 1 cm. The annotation process was rigorously conducted by two experienced gynecological ultrasound diagnosticians with over five years of professional experience. To ensure the accuracy and reliability of the annotations, the collected data underwent a cross-annotation procedure between the two diagnosticians. We randomly select 83 videos for training, and the remaining 17 videos are utilized for testing.
Compared Methods. We compare LGRNet against 9 recent state-of-the-art segmentation methods, including five image-based methods and four video-based methods. These image-based methods are UNet++[29], PraNet[8], LDNet[28], WeakPolyp[21], and BUSSeg[23], while video-based methods are PNS-Net[12], DPSTT[14], FLA-Net[15], and MS-TFAL[5]. For each compared method, we utilize the hyperparameters settings from the original paper or their official codes for fair comparisons.
Evaluation Metrics. For quantitative comparison, we utilize five common metrics, including Dice Coefficient (Dice), Intersect of Union (IoU), Sensitivity, Mean Absolute Error (MAE), and S-Measure (structural similarity [7]). We also compute the inference Multiply-Accumulate Counts (MACs, GFLOPS) and the number of parameters (Params) for efficiency comparisons.
Implementation Details. We set $T_{clip}=6$ in our experiments. Each frame is resize to 352×352. The training augmentation contains the horizontal flip, the vertical flip, and the perspective transform with magnitude 0.12. We use AdamW [16] and set initial learning rate as 1e-3 with a backbone multiplier of 0.1. The multistep scheduler the learning rate by 0.5 every 3 epochs. Gradient clipping with square norm value 1e-2 is used. We use a point sampling [4] with an oversampling ratio of 3.0 and importance of 0.76 for the bipartite matching mask loss computation. Both the number of encoder and decoder layers are set to 3. We use Res2Net-50[9] as backbone, and empirically set $\lambda_{class}=2$, $\lambda_{dice}=5$, $\lambda_{ce}=2$, $c=64$, $k=5$, $d=2$, $\bar{N}=20$, $\mu=3$, and $\hat{N}=10$. More ablation studies on hyperparameters are demonstrated in Sec. 3.2.

3.1 Comparisons with SOTA methods

Our UFUV dataset. As shown in Table1, although our network does not have the smallest inference time and the smallest number of parameters, our method

Table 1. Quantitative comparisons on our UFUV dataset.

Method	Publication		Dice↑	IoU↑	Sensitivity ↑	S-Measure↑	MAE ↓	GFLOPs↓	Params↓
UNet++[29]	TMI'19	image	0.681	0.540	0.611	0.728	0.081	22.6 G	29.8 M
PraNet[8]	MICCAI'20	image	0.724	0.583	0.709	0.751	0.076	20.3 G	16.1 M
LDNet[28]	MICCAI'22	image	0.738	0.588	0.707	0.753	0.068	12.6 G	15.8 M
WeakPolyp[21]	MICCAI'23	image	0.725	0.579	0.682	0.747	0.075	5.2 G	25.8 M
BUSSeg[23]	TMI'23	image	0.740	0.612	0.711	0.770	0.066	23.8 G	28.6 M
PNS-Net[12]	MICCAI'21	video	0.735	0.601	0.685	0.750	0.065	19.5 G	15.7 M
DPSTT[14]	MICCAI'22	video	0.738	0.609	0.707	0.769	0.065	24.8 G	30.2 M
FLA-Net[15]	MICCAI'23	video	0.741	0.615	0.710	0.773	0.066	18.4 G	87.6 M
MS-TFAL[5]	MICCAI'23	video	0.748	0.625	0.714	0.781	0.063	12.2 G	24.6 M
Ours	– –	video	**0.775**	**0.658**	**0.776**	**0.793**	**0.060**	13.2 G	26.6 M

Fig. 4. Visual comparisons on UFUV of our network and compared SOTA methods.

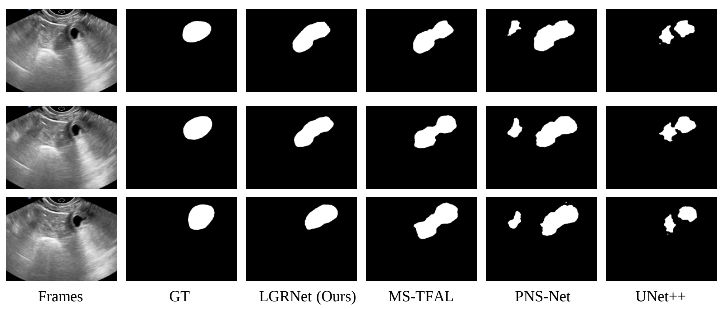

Table 2. Quantitative comparisons CVC-612[2] and CVC-300[20] for video polyp segmentation.

	Metrics	UNet++[29] TMI'19	PraNet[8] MICCAI'20	PNS-Net [12] MICCAI'21	LDNet[28] MICCAI'22	FLA-Net[15] MICCAI'23	MS-TFAL[5] MICCAI'23	Ours –
CVC-612-V	maxDice↑	0.684	0.869	0.873	0.870	0.885	0.911	**0.933**
	maxIoU↑	0.570	0.799	0.800	0.799	0.814	0.846	**0.877**
	S_α ↑	0.805	0.915	0.923	0.918	0.920	0.961	**0.947**
	maxSpe↑	0.952	0.983	0.991	0.987	0.992	0.994	**0.995**
	E_ϕ ↑	0.830	0.936	0.944	0.941	0.963	0.971	**0.977**
	MAE ↓	0.025	0.013	0.012	0.013	0.012	0.010	**0.007**
CVC-300-TV	maxDice↑	0.649	0.739	0.840	0.835	0.874	0.891	**0.916**
	maxIoU↑	0.539	0.645	0.745	0.741	0.789	0.810	**0.852**
	S_α ↑	0.796	0.833	0.909	0.898	0.907	0.912	**0.937**
	maxSpe↑	0.944	0.993	0.996	0.994	0.996	0.997	**0.997**
	E_ϕ ↑	0.831	0.852	0.921	0.910	0.969	0.974	**0.986**
	MAE ↓	0.024	0.016	0.013	0.015	0.010	0.007	**0.005**

achieves superior performance over all compared SOTA image/video segmentation methods on all segmentation metrics. Moreover, our method also achieves significant efficiency improvement compared with most methods. Besides, Fig.4 compares the visual results produced by different methods. Our network can more accurately segment uterine fibroid than SOTA methods, and our segmentation results are most consistent with the ground truth. More visual comparison results are presented in the supplementary material.

Video Polyp Segmentation (VPS) datasets To further demonstrate the effectiveness of our method, we compare our network against SOTA methods on three public Video Polyp Segmentation (VPS) benchmark datasets, which are CVC-612[2], CVC-300[20], and SUN-SEG[13]. For CVC-612[2] and CVC-300[20],

Table 3. Quantitative comparisons on SUN-SEG[13] for video polyp segmentation.

Model	Publication	Backbone	Easy Testing		Hard Testing	
			Dice	IoU	Dice	IoU
PraNet [8]	MICCAI'20	Res2Net-50	0.689	0.608	0.660	0.569
2/3D [18]	MICCAI'20	ResNet-101	0.755	0.668	0.737	0.643
SANet [22]	MICCAI'21	Res2Net-50	0.693	0.595	0.640	0.543
PNS+ [13]	MIR'22	Res2Net-50	0.787	0.704	0.770	0.679
DPSTT[14]	MICCAI'22	Res2Net-50	0.804	0.725	0.794	0.709
WeakPolyp[21]	MICCAI'23	Res2Net-50	0.792	0.715	0.807	0.727
		PVTv2-B2	0.853	0.781	0.854	0.777
FLA-Net[15]	MICCAI'23	Res2Net-50	0.805	0.723	0.811	0.730
		PVTv2-B2	0.856	0.784	0.858	0.781
MS-TFAL[5]	MICCAI'23	Res2Net-50	0.822	0.742	0.826	0.751
		PVTv2-B2	0.859	0.792	0.862	0.788
Ours	– –	Res2Net-50	**0.843**	**0.765**	**0.843**	**0.774**
Ours	– –	PVTv2-B2	**0.875**	**0.810**	**0.876**	**0.805**

we follow PNS-Net[12] to use the same training setting and test datasets. For SUN-SEG[13], we follow WeakPolyp [21] to combine the "Hard (Easy) Seen" and "Hard (Easy) Unseen" split into "Hard (Easy) Testing". As shown in Table2 & 3, our network also achieves better metric performance than all compared methods on all three benchmark datasets, which indicates that our network has the best video polyp segmentation performance.

Table 4. Component Analysis.

CNP	*HilbertSS*	Dice↑	IoU↑	S-Measure↑	MAE ↓
✗	✗	0.722	0.581	0.750	0.074
✗	✓	0.753	0.633	0.784	0.062
✓	✗	0.747	0.626	0.776	0.067
✓	✓	**0.775**	**0.658**	**0.793**	**0.060**

Table 5. Hyperparameter Ablations

Component	Version	Dice↑	IoU↑	S-Measure↑	MAE ↓
CNP	k=3, d=1	0.768	0.652	0.786	0.062
	k=3, d=2	0.771	0.656	0.789	0.061
	k=5, d=2	**0.775**	**0.658**	**0.793**	**0.060**
	k=7, d=2	0.766	0.647	0.784	0.064
HilbertSS	Zigzag Scan	0.761	0.639	0.788	0.060
	Hilbert Scan	**0.775**	**0.658**	**0.793**	**0.060**
	$\bar{N}=10$	0.764	0.643	0.786	0.062
	$\bar{N}=20$	**0.775**	**0.658**	**0.793**	**0.060**
	$\bar{N}=30$	0.771	0.657	0.792	0.060

3.2 Ablation Study

We conduct ablation analysis on *CNP* and *HilbertSS* by removing them or ablating their hyperparameters. As shown in Tab.4, removing either component leads to a performance drop. Removing both components causes the model to ignore the vital temporal context information. Moreover, the global HilbertSS (0.722→0.753) achieves more improvement than local *CNP* (0.722→0.747), which validates the design of frame bottleneck queries and reciprocal local-global learning. For the hyperparameter ablation, we set different kernel size k and dilation d for *CNP*, different selective scan strategy, and number of frame bottleneck queries $\bar{N}$ for *HilbertSS*. As shown in Fig.5, for *CNP*, both bigger kernel size and larger dilation may lead to improvement, but when much bigger kernel size is used (k=7 compared with k=5), the performance saturates. For *HilbertSS*, using Zigzag scan leads to lower performance. Moreover, using more bottleneck queries increases performance, but the performance also saturates after some threshold.

4 Conclusion

This paper collects and annotates the first ultrasound video uterine fibroid segmentation (UFUV) dataset, which contains 100 videos with 5,000 annotated video frames. We further propose the Local-Global Reciprocal Net (LGRNet) to efficiently aggregate global temporal context information for ultrasound video segmentation. The *Condense* and *Distribute* layers with our proposed frame bottleneck queries bridge the local *CNP* and global *HilbertSS*, and facilitate reciprocally propagating the crucial local-global temporal context information. Experimental results on UFUV dataset and other three public Video Polyp Segmentation (VPS) datasets show that LGRNet quantitatively and qualitatively outperforms existing state-of-the-art image and video segmentation methods.

Acknowledgments This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0671), the Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0671), and the InnoHK funding launched by Innovation and Technology Commission, Hong Kong SAR.

Disclosure of Interests The authors declare that they have no competing interests.

References

1. The dilation factor of the peano-hilbert curve. Mathematical Notes **80**, 609–620 (2006)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
3. Chen, W., Zhu, X., Chen, G., Yu, B.: Efficient point cloud analysis using hilbert curve. In: European Conference on Computer Vision. pp. 730–747. Springer (2022)
4. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
5. Cui, B., Zhang, M., Xu, M., Wang, A., Yuan, W., Ren, H.: Rectifying noisy labels with sequential prior: Multi-scale temporal feature affinity learning for robust video segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 90–100. Springer (2023)
6. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015)
7. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)
8. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
9. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence **43**(2), 652–662 (2019)
10. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
11. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6185–6194 (2023)
12. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2021)
13. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)
14. Li, J., Zheng, Q., Li, M., Liu, P., Wang, Q., Sun, L., Zhu, L.: Rethinking breast lesion segmentation in ultrasound: A new video dataset and a baseline network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–400. Springer (2022)
15. Lin, J., Dai, Q., Zhu, L., Fu, H., Wang, Q., Li, W., Rao, W., Huang, X., Wang, L.: Shifting more attention to breast lesion segmentation in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 497–507. Springer (2023)

16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Okolo, S.: Incidence, aetiology and epidemiology of uterine fibroids. Best practice & research Clinical obstetrics & gynaecology **22**(4), 571–588 (2008)
18. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 295–305. Springer (2020)
19. Spak, D.A., Plaxco, J., Santiago, L., Dryden, M., Dogan, B.: Bi-rads® fifth edition: A summary of changes. Diagnostic and interventional imaging **98**(3), 179–190 (2017)
20. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015)
21. Wei, J., Hu, Y., Cui, S., Zhou, S.K., Li, Z.: Weakpolyp: You only look bounding box for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 757–766. Springer (2023)
22. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 699–708. Springer (2021)
23. Wu, H., Huang, X., Guo, X., Wen, Z., Qin, J.: Cross-image dependency modelling for breast ultrasound segmentation. IEEE Transactions on Medical Imaging (2023)
24. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
25. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)
26. Yu, H., Li, Y., Wu, Q., Zhao, Z., Chen, D., Wang, D., Wang, L.: Mining negative temporal contexts for false positive suppression in real-time ultrasound lesion detection. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 3–13. Springer Nature Switzerland, Cham (2023)
27. Zhang, J., Kamata, S.i., Ueshige, Y.: A pseudo-hilbert scan algorithm for arbitrarily-sized rectangle region. In: International Workshop on Intelligent Computing in Pattern Analysis and Synthesis. pp. 290–299. Springer (2006)
28. Zhang, R., Lai, P., Wan, X., Fan, D.J., Gao, F., Wu, X.J., Li, G.: Lesion-aware dynamic kernel for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 99–109. Springer (2022)
29. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)