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

MICCAI 2024

No Institute Given

Abstract. Keywords: Uterine Fibroid Segmentation · Ultrasound Videos · Selective 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 [20]. Consequently, regular screening and early detection of uterine fibroids are essential for initiating timely life-saving treatments. Since 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 segmentation in videos has attracted much attention from the medical community [17,16,32]. For example, FLA-Net[17] presents a frequency and location feature aggregation network, which incorporates frequency-based temporal features learning, for video breast lesion segmentation. UltraDet[32] proposes to aggregate the negative temporal context to facilitate filtering out false positive predictions in video breast lesion detection. However, uterine fibroid segmentation in ultrasound videos remains unexplored and significantly challenged by noisy temporal motions, blurry boundaries, and changing lesion size over time.

To this end, in this paper, **1) we collect the first ultrasound video benchmark dataset for uterine fibroids segmentation, dubbed UFUV**, which contains 100 videos with per-frame annotations by experts. The aforementioned ultrasound segmentation methods perform global temporal aggregation by processing dense, multi-scale features. Such an approach is time-consuming and inefficient, as it involves handling background tokens with weak semantic relevance. To address this problem, **2) we present Local-Global Reciprocal Net (LGRNet) which follows a local-global reciprocal learning strategy** to efficiently aggregate the temporal context utilizing a set of frame bottleneck tokens. In our LGRNet, **3) we incorporate the local Cyclic Neighborhood Propagation module and Hilbert Selective Scan module** to reciprocally learn from each other through a set of *frame bottleneck queries*. LGRNet can efficiently aggregate the temporal context information that is crucial to ultrasound segmentation. **4)** We conduct extensive experiments to demonstrate that

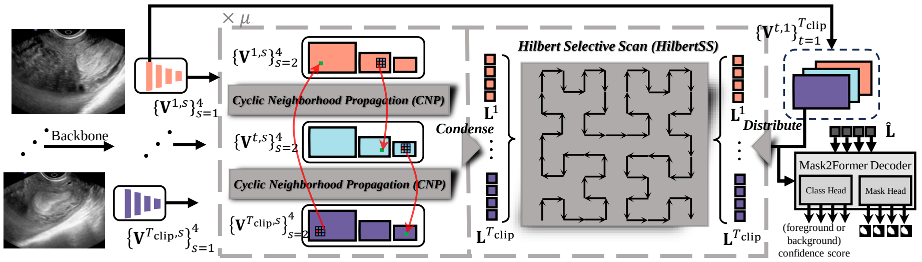

× μ
$\{\mathbf{V}^{1,s}\}_{s=1}^{4}$
Backbone
$\{\mathbf{V}^{T_{\text{clip}},s}\}_{s=1}^{4}$
$\{\mathbf{V}^{1,s}\}_{s=2}^{4}$
Cyclic Neighborhood Propagation (CNP)
$\{\mathbf{V}^{t,s}\}_{s=2}^{4}$
Cyclic Neighborhood Propagation (CNP)
$\{\mathbf{V}^{T_{\text{clip}},s}\}_{s=2}^{4}$
$\mathbf{L}^1$
Condense
$\mathbf{L}^{T_{\text{clip}}}$
Hilbert Selective Scan (HilbertSS)
$\mathbf{L}^1$
Distribute
$\mathbf{L}^{T_{\text{clip}}}$
$\{\mathbf{V}^{t,1}\}_{t=1}^{T_{\text{clip}}}$
$\hat{\mathbf{L}}$
Mask2Former Decoder
Class Head
Mask Head
(foreground or background) confidence score

Fig. 1. LGRNet architecture. We propose to use a set of *frame bottleneck queries* to efficiently aggregate the global temporal context and distribute global context back to local refinement, which forms a reciprocal net.

LGRNet both quantitatively and qualitatively achieves state-of-the-art performance on our UFUV dataset and three public video polyp segmentation datasets.

2 Method

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 $\{\{\mathbf{V}^{t,s} \in \mathcal{R}^{H_sW_s \times c}\}_{s=1}^{S}\}_{t=1}^{T_{\text{clip}}}$, 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[29] layer. Then, clip features of the last three scales are input to **Cyclic Neighborhood Propagation (CNP)** 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 tokens of all frames, i.e. $\{\mathbf{L}^t\}_{t=1}^{T_{\text{clip}}}$, are then input to **Hilbert Selective Scan (HilbertSS)** to efficiently path-connect each frame. Afterward, global-view queries of each frame are input to a *Distribute* layer to distribute the global temporal context back to multi-scale features for local temporal refinement. Finally, the local-global reciprocally encoded multi-scale features are input to the decoder for foreground/background classification and mask prediction. Compared with existing methods which only output single predicted mask, LGRNet can output a set of different possible mask predictions associated with a lesion confidence score, which also facilities more comprehensive diagnosis.

2.1 Local Cyclic Neighborhood Propagation (CNP)

Local CNP. Motion priories, such as optical flow [32], can be utilized as pixel-wise guidance to propagate inter-frame temporal information. However, they incorporate additional parameters of pretrained optical flow predictor [7] 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) [12] demonstrates that only

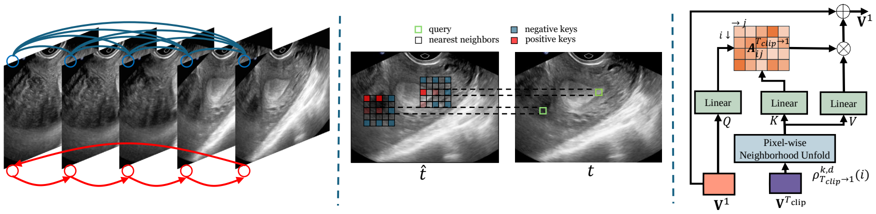


Fig. 2. Illustrations of *CNP*. (Left) We enforce local cyclic inter-frame dependencies (red), instead of estabaling fully connected connections (blue). (Middle) In *CNP*, each query point takes the corresponding nearest neighbors of the cyclic frame as attention keys. (Right) Detailed implementation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

involving nearest neighbors as attention keys can achieve comparable performance in 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 set of *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:

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

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

where M denotes the number of attention heads, $\rho^{k,d}_{\hat{t}\rightarrow t}(i)$ is the set of nearest neighbors at $\hat{t}$-th frame w.r.t i-th position at t-th frame, and the nearest neighbors are defined by a kernel with size k and dilation d. $\{A_{ij}^{\hat{t}\rightarrow t}\}_{j=1}^{|\rho^{k,d}_{\hat{t}\rightarrow t}(i)|}$ denotes the attention weights, which are the dot product of query with each neighbor key, i.e. softmax$(\frac{\mathbf{q}_i^{tT}\mathbf{W}_K\mathbf{V}_j^{\hat{t}}}{\sqrt{c}})$. 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 patch-level *fully connected* inter-frame dependencies, since when motion changes severely and is noisy, dense connections would connect a query patch to background patches with weak semantics at distant frames, which also leads to increased computation.

2.2 Global Hilbert Selective Scan (HilbertSS)

Frame Queries as Information Bottleneck. Radiologists often require visual features spanning long temporal context [23] to not only decide whether a possible region is lesion or not, but also refine their per-frame diagnosis. 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:

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

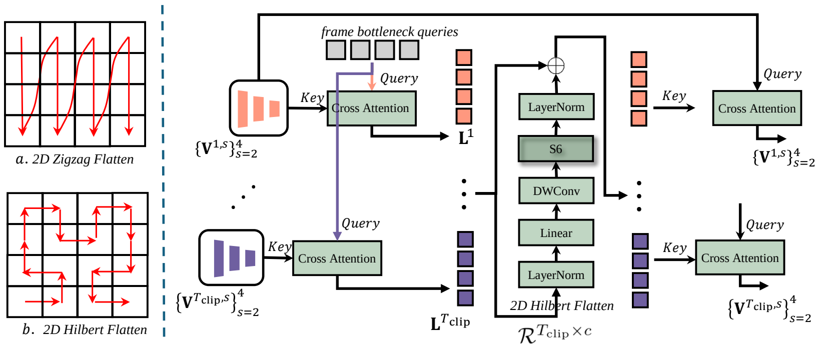


Fig. 3. Illustrations of *HilbertSS*. (Left) Comparison between (a) Zigzag Flatten (b) Hilbert Flatten. (Right) Detailed implementation of global part of LGRNet.

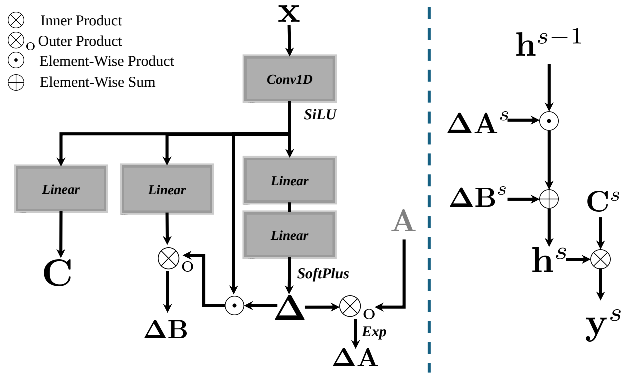


Fig. 4. Implementation of S6 block. (Left) Precomputation of input-dependent weights. (Right) Scanning for iterative hidden state transformation.

where the *Condense* layer is implemented as a Cross Attention Layer [25] with Residual connection [13]. In the cross attention layer, the 2D multi-scale features are flattened and concatenated. Specifically, the query length is $\bar{N}$ and the key length equals to the sum of multi-scale shapes, i.e. $\sum_{s=1}^{S} H_s \times W_s$. Frame queries can be seen as bottlenecks extracting semantically rich lesion information and filtering out irrelevant and redundant features of each frame, which facilitates later effective global temporal information exchange.

Preliminary of Selective State Space Model. Recently, Selective State Space Model (Mamba) [11] was proposed as a new sequence model with linear-time complexity, which even outperforms Transformers[25] on some language tasks.

Mamba is built based on the Selective Scan Block (S6). As shown in Fig.4, each S6 block transforms the input sequence $\{\mathbf{x}^s \in \mathcal{R}^c\}_{s=1}^{S}$ to $\{\mathbf{y}^s \in \mathcal{R}^c\}_{s=1}^{S}$ through the precomputation stage and the scanning stage. During precomputation, a set of linear layers are used to precompute the input-dependent dynamic weights $\mathbf{\Delta B}, \mathbf{\Delta A} \in \mathcal{R}^{S \times c_{\text{state}} \times c}$, $\mathbf{C} \in \mathcal{R}^{S \times c_{\text{state}}}$, where c_{state} denotes the dimension of the hidden state. The precomputation stage can be formalized as:

$$\begin{aligned} \mathbf{x} &= \text{SiLU}(\text{Conv1D}(\mathbf{x})) \\ \mathbf{C} &= \text{Linear}_{c_{\text{state}}}(\mathbf{x}) \in \mathcal{R}^{S \times c_{\text{state}}} \\ \mathbf{\Delta} &= \text{Softplus}(\text{Linear}_c(\text{Linear}_{c_{\text{rank}}}(\mathbf{x})) \in \mathcal{R}^{S \times c} \\ \mathbf{\Delta A} &= \exp\{\otimes_{\text{o}}(\mathbf{\Delta}, \mathbf{A})\} \in \mathcal{R}^{S \times c_{\text{state}} \times c} \\ \mathbf{\Delta B} &= \otimes_{\text{o}}((\mathbf{x} \odot \mathbf{\Delta}), \text{Linear}_{c_{\text{state}}}(\mathbf{x})) \in \mathcal{R}^{S \times c_{\text{state}} \times c}, \end{aligned} \tag{4}$$

where $\mathbf{A}$ denotes a learnable parameter of shape $c_{\text{state}} \times c$, c_{rank} denotes the dimension of low-rank projection [11], Linear_* denotes a linear layer transforming input to dimension $*$, SiLU denotes the Sigmoid Linear Unit activation, Conv1D denotes the depth-wise 1D convolution, $\otimes_{\text{o}}$ denotes the outer product, and Softplus denotes the softplus activation. To some extent, the precomputation stage in S6 corresponds to weights computation (normalized inner product of input queries and input keys) in attention [25]. Both two procedures in S6 and attention generate input-dependent, i.e. dynamic, weights for subsequent information aggregation. This characteristic enables S6 to attend to the in-context information.

During scanning, the hidden state is zero-initialized, i.e. $\mathbf{h}^0 = \mathbf{0} \in \mathcal{R}^{c_{\text{state}} \times c}$. Formally, each step of hidden state transformation can be formalized as:

$$\begin{aligned} \mathbf{h}^s &= \mathbf{h}^{s-1} \odot \mathbf{\Delta A}^s \oplus \mathbf{\Delta B}^s \in \mathcal{R}^{c_{\text{state}} \times c} \\ \mathbf{y}^s &= \otimes(\mathbf{C}^s, \mathbf{h}^s) \in \mathcal{R}^c, \end{aligned} \tag{5}$$

where $\otimes$ denotes the inner product, $\oplus$ denotes the element-wise sum, $\odot$ denotes the element-wise product. Although the scanning stage is recurrent, the transition from $\mathbf{h}^{s-1}$ to $\mathbf{h}^s$ is associative. As shown in [3][22], the scanning process can thus be parallelized to logarithmic complexity w.r.t the sequence length.

Global HilbertSS. In LGRNet, we aim to devise S6 on the aggregated frame queries $\{\mathbf{L}^t\}_{t=1}^{T_{\text{clip}}} \in \mathcal{R}^{\bar{N} \times T_{\text{clip}} \times c}$ to efficiently propagate the global clip-level temporal information. To apply S6 to 2D input, a direct approach is flattening input into a sequence using Zigzag curve[1] as in attention-based models. However, Eq.5 implies the scanning stage is position-aware, in the sense that different scanning order would generate different set of hidden states. Moreover, the contextual information is propagated along the sequence only through the hidden state $\mathbf{h}^s$. Although the selection mechanism may enable boundary resetting [11] and context filtering [11] on some language tasks, we believe locality is more important to vision tasks, in the sense that for a local structured visual region, the hidden state should aggregate them in a group, instead of corrupting their neighboring structure in the flattened 1D sequence.

[1] https://pytorch.org/docs/2.0/generated/torch.flatten.html?highlight=flatten

As an intuitive example shown in Fig.3, zigzag scanning may corrupt the original neighborhood structure, in the sense that two tokens which are neighboring to each other on original 2D layout would be distant to each other on the flattened 1D sequence. Formally, any flattening curve can be generalized to a continuous *Space Filling Curve* (SFC) σ, which maps each point $x \in$ [0, 1] to $\sigma(x) \in$ [0,1]× [0,1]. We also denote n as the curve order of the Hilbert curve, which, in our discrete case, approximates to height and width of the filled grid. For any two points x, y in [0,1], their *Space to Linear Ratio* (SLR) is defined as:

$$\frac{|\sigma(x) - \sigma(y)|^2}{|x - y|}. \tag{6}$$

The *Dilation Factor* (DF) of a SFC is defined as the upper bound of the SLR. For same two points in [0, 1], if a SFC has lower DF, their mappings will also be closer in the 2D grid, which accords with locality preserving requirement in the scanning stage. As proved in [1][4], the DF of Hilbert curve is 6, while the Zigzag curve is $4^n - 2^{n+1} + 2$, which diverges to ∞ as the curve order n increases. This shows that Hilbert curve could preserve the 2D locality structure, which accords with our intuition that lesions spatial-temporally close to each other should be scanned in groups and tracked together.

To execute Hilbert scan on a 2D $\bar{N} \times T_{\text{clip}}$ grid, we use the implementation of [33][2] to generate the pseudo Hilbert curve. In all, the HilbertSS module can be formalized as:

$$\{\mathbf{L}^t\}_{t=1}^{T_{\text{clip}}} = HilbertUnFlatten(S6(HilbertFlatten(\{\mathbf{L}^t\}_{t=1}^{T_{\text{clip}}}))), \tag{7}$$

where $HilbertFlatten$ means we use the `torch.index_select`[3] function to flatten 2D input into a 1D sequence using the generated Hilbert indices. $HilbertUnFlatten$ means we use the `torch.Tensor.scatter_add_`[4] to add the transformed features back to the original input according to the same Hilbert indices.

It should also be noted that existing S6-based vision models, including Vim[36], Vmamba[18], Segmamba[30], and Vivim[31], scan the multi-scale feature maps, which include background tokens with weak semantics and thus may affect the recurrent hidden state transformation in Eq.5. LGRNet devises a set of queries to filter out noisy information and extract the semantically rich lesion information, which would help the scanning process to efficiently attend to more informative visual regions.

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

$$\{\mathbf{V}^{t,s}\}_{s=1}^{S} = Distribute(\text{query} = \{\mathbf{V}^{t,s}\}_{s=1}^{S}, \text{key} = \mathbf{L}^t), t = 1, ..., T_{\text{clip}}. \tag{8}$$

[2] https://github.com/jakubcerveny/gilbert

[3] https://pytorch.org/docs/stable/generated/torch.index_select.html

[4] https://pytorch.org/docs/stable/generated/torch.Tensor.scatter_add_.html

Same to the *Condense* layer, the *Distribute* layer is also implemented as a Cross Attention Layer [25] with Residual connection [13], where the query length is $\sum_{s=1}^{S} H_s \times W_s$ and the key length is $\bar{N}$.

2.3 Decoder

Our decoder uses the same architecture with Mask2Former[5]. A set of learnable *temporal queries* $\hat{\mathbf{L}} \in \mathcal{R}^{\hat{N} \times c}$ are used to cross-attend to 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 mask is 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: $\lambda_{class} L_{class} + \lambda_{dice} L_{dice} + \lambda_{ce} L_{ce}$. During inference, the foreground classification probability can be interpreted as the lesion confidence of the corresponding predicted mask region. LGRNet can generate multiple, i.e. $\hat{N}$, mask predictions, each with a lesion confidence score. We choose the mask with the highest foreground probability as final output to compare with other methods in our experiments.

3 Experiments

Dataset. We collected and annotated the first ultrasound video uterine fibroid segmentation dataset (UFUV dataset). 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 videos that showcase 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 is utilized for testing.

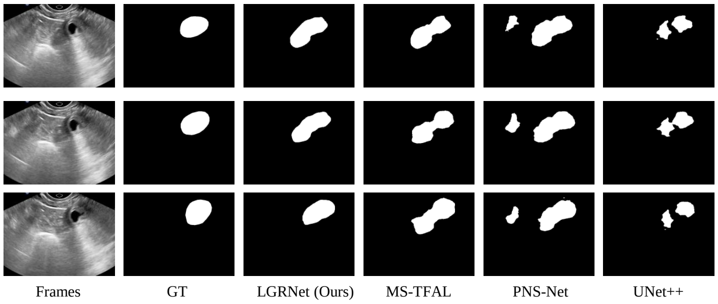
Compared Methods. We compare our network against 9 recent state-of-the-art (SOTA) segmentation methods, including five image-based methods and four video-based methods. These image-based methods are UNet++[35], PraNet[9], LDNet[34], WeakPolyp[26], and BUSSeg[28], while video-based methods are PNS-Net[14], DPSTT[16], FLA-Net[17], and MS-TFAL[6]. For each compared method, we utilize the hyperparameters settings of the original paper or their official codes for fair comparisons.

Evaluation Metrics. For quantitative comparison, we devise five common metrics, including Dice Coefficient (Dice), Intersect of Union (IoU, Jaccard), Sensitivity, Mean Absolute Error (MAE), and S-Measure (structural similarity [8]). We also compute the inference Multiply-Accumulate Counts (MACs, GFLOPS) and the number of parameters (Params) for comparisons.

Table 1. Quantitative comparisons on our UFUV dataset.

Method	Publication		Dice↑	IoU↑	Sensitivity ↑	S-Measure↑	MAE ↓	GFLOPs↓	Params↓
UNet++[35]	TMI'19	image	0.681	0.540	0.611	0.728	0.081	22.6 G	29.8 M
PraNet[9]	MICCAI'20	image	0.724	0.583	0.709	0.751	0.076	20.3 G	16.1 M
LDNet[34]	MICCAI'22	image	0.738	0.588	0.707	0.753	0.068	12.6 G	15.8 M
WeakPolyp[26]	MICCAI'23	image	0.725	0.579	0.682	0.747	0.075	5.2 G	25.8 M
BUSSeg[28]	TMI'23	image	0.740	0.612	0.711	0.770	0.066	23.8 G	28.6 M
PNS-Net[14]	MICCAI'21	video	0.735	0.601	0.685	0.750	0.065	19.5 G	15.7 M
DPSTT[16]	MICCAI'22	video	0.738	0.609	0.707	0.769	0.065	24.8 G	30.2 M
FLA-Net[17]	MICCAI'23	video	0.741	0.615	0.710	0.773	0.066	18.4 G	87.6 M
MS-TFAL[6]	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. 5. Visual comparisons on UFUV of our network and compared SOTA methods.



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 [19] and set initial learning rate as 1e-3 with a backbone multiplier of 0.1. The multistep scheduler decays the learning rate by 0.5 under each 3 epochs. Gradient clipping with square norm value 1e-2 is used. We use point sampling [5] with oversampling ratio of 3.0 and importance of 0.76 for the bipartite matching mask loss computation. The number of decoder layers is set to 3. We use Res2Net-50[10] 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, our method outperforms existing state-of-the-art medical image/video segmentation methods under all metrics. Specifically, LGRNet achieves 0.658 IoU score and 0.793 structural similarity[8], which is notably higher than state-of-the-art ultrasound video segmentation methods FLA-Net[17] and MS-TFAL[6]. Moreover, compared with DPSTT[16] and FLA-Net[17] whose flops and number of parameters are 24.8G/30.2M and 18.4G/87.6M respectively, LRGNet achieves significantly better performance with less complexity and parameters, i.e. 13.2G/26.6M. Therefore, our method also achieves a favorable balance between performance and speed.

Besides, Fig.5 compares the visual results produced by different methods. Notably, our network can more accurately segment uterine fibroid than state-of-

Table 2. Quantitative comparisons CVC-612[2] and CVC-300[24] for video polyp segmentation.

		UNet++[35]	PraNet[9]	PNS-Net [14]	LDNet[34]	FLA-Net[17]	MS-TFAL[6]	Ours
	Metrics	TMI'19	MICCAI'20	MICCAI'21	MICCAI'22	MICCAI'23	MICCAI'23	–
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**
CVC-612-T	maxDice↑	0.740	0.852	0.860	0.857	0.861	0.864	**0.875**
	maxIoU↑	0.635	0.786	0.795	0.791	0.795	0.796	**0.814**
	S_α ↑	0.800	0.886	0.903	0.892	0.904	0.906	**0.907**
	maxSpe↑	0.975	0.986	0.992	0.988	0.993	0.995	**0.998**
	E_ϕ ↑	0.817	0.904	0.903	0.903	0.904	0.910	**0.915**
	MAE ↓	0.059	0.038	0.038	0.037	0.036	0.038	**0.035**

Table 3. Quantitative comparisons on SUN-SEG[15] for video polyp segmentation.

Model	Publication	Backbone	Easy Testing		Hard Testing	
			Dice	IoU	Dice	IoU
PraNet [9]	MICCAI'20	Res2Net-50	0.689	0.608	0.660	0.569
2/3D [21]	MICCAI'20	ResNet-101	0.755	0.668	0.737	0.643
SANet [27]	MICCAI'21	Res2Net-50	0.693	0.595	0.640	0.543
PNS+ [15]	MIR'22	Res2Net-50	0.787	0.704	0.770	0.679
DPSTT[16]	MICCAI'22	Res2Net-50	0.804	0.725	0.794	0.709
WeakPolyp[26]	MICCAI'23	Res2Net-50	0.792	0.715	0.807	0.727
		PVTv2-B2	0.853	0.781	0.854	0.777
FLA-Net[17]	MICCAI'23	Res2Net-50	0.805	0.723	0.811	0.730
		PVTv2-B2	0.856	0.784	0.858	0.781
MS-TFAL[6]	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**

the-art methods. Moreover, due to the local-global reciprocally learning strategy, LGRNet can utilize global temporal information to refine the local predictions. Compared with MS-TFAL[6] and PNS-Net[14] which have more False Positive (FP) predictions, LRGNet can filter out these predictions, thus generating segmentations that are most consistent with the ground truth. More visual comparison results are presented in the supplementary material.

Video Polyp Segmentation (VPS) benchmark datasets. To further demonstrate the effectiveness of LGRNet, we compare against state-of-the-art VPS methods on three public VPS benchmark datasets, which are CVC-612[2], CVC-300[24], and SUN-SEG[15]. For CVC-612[2] and CVC-300[24], we follow PNS-Net[14] to use the same training setting and test dataset splits. For SUN-SEG[15], we follow WeakPolyp [26] to combine the "Hard (Easy) Seen" and "Hard (Easy) Unseen" split into "Hard (Easy) Testing". As shown in Table2 & 3, LGRNet also achieves better performance than other compared methods for all three VPS datasets, which validates the universality and effectiveness of LGRNet.

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 the whole component or devising different combinations of their hyperparameters. As shown in Tab.4, removing either component leads to a segmentation 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 effectiveness of information bottleneck token design and reciprocal local-global learning design. For the component hyperparameter ablation, we set different kernel size k and dilation d for *CNP* and different selective scan strategy, number of frame bottleneck queries $\bar{N}$ for *HilbertSS*. As shown in Fig.5, for *CNP*, both larger kernel and dilation size may lead to performance improvement, but the improvement statures after some threshold, specifically when much larger kernel size is used (k=7 compared with k=5). For *HilbertSS*, using Zigzag scan leads to lower performance, which also validates our locality preserving design principle for *HilbertSS*. Moreover, using more bottleneck queries increases performance, but the performance saturates after some threshold around $\bar{N} = 20$.

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. Armed with **Local Cyclic Neighborhood Propagation (CNP)** and **Global Hilbert Selective Scan (HilbertSS)**, the proposed **Local-Global Reciprocal Net (LGRNet)** effectively aggregate local and global temporal context and enable the two parts to reciprocally learn from each other through a set of Frame Bottleneck Queries. Experimental results on our UFUV dataset show that LGRNet quantitatively and qualitatively outperforms state-of-the-art medical image and video segmentation methods. We also carry out thorough ablations on the component and hyperparameters of *CNP* and *HilbertSS* to support our design motivations. Moreover, LGRNet also achieves state-of-the-art performance on three public Video Polyp Segmentation (VPS) datasets, validating the universality and effectiveness of LGRNet.

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. Blelloch, G.E.: Prefix sums and their applications (1990)
4. 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)
5. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
12. 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)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. 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)
15. 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)
16. 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)

17. 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)
18. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Okolo, S.: Incidence, aetiology and epidemiology of uterine fibroids. Best practice & research Clinical obstetrics & gynaecology **22**(4), 571–588 (2008)
21. 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)
22. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933 (2022)
23. 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)
24. 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)
25. 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)
26. 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)
27. 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)
28. Wu, H., Huang, X., Guo, X., Wen, Z., Qin, J.: Cross-image dependency modelling for breast ultrasound segmentation. IEEE Transactions on Medical Imaging (2023)
29. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
30. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
31. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)
32. 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)
33. 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)

34. 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)
35. 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)
36. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)