STDDNet: Harnessing Mamba for Video Polyp Segmentation via Spatial-aligned Temporal Modeling and Discriminative Dynamic Representation Learning

Supplementary Material

1. Ablation Studies on GDAU

Different to traditional GRUs [1, 2], our proposed GDAU utilizes an activation ratio to regulate the outputs. Through model training, it is able to adaptively focus on the regions of foreground targets and identify dynamic details. According to the components ablation studies in the main text, the activation ratio significantly influences the dynamic features learning of DDFE. By removing activation ratio, GDAU reverts to a standard GRU, and thus DDFE lose the ability to distinguish static and dynamic regions. Furthermore, in this section, we extend the ablation studies to analyze the GDAU's calculations involving the static, varying values and the dynamic map. As outlined in the main text, four widely-used metrics are employed for evaluations, including structure measure ($\mathcal{S}_\alpha$) [3], enhanced-alignment measure (E_ϕ) [4], weighted F-measure (F_β^w) [6] and mean dice coefficient (Dice). Dice measures the similarity between predictions and their corresponding ground truth annotations. $\mathcal{S}_\alpha$ servers as a metric for structural similarity, with α set to 0.5. In addition, E_ϕ is used to evaluate the heterogeneous location and shape of polyps, where ϕ denotes the enhanced-alignment matrix. F_β^w is the weighted harmonic mean of precision and recall for more comprehensive evaluation. The quantitative results are presented in Tab. 1. According to Eq. 11 in the main text, when the static value is omitted, the equation simplifies to:

$$U^t = S \odot (1 - V^t_{vary}) + V^t_{vary} \odot M^t_{dynamic} \quad (1)$$

Similarly, when the varying value is not computed, we use the following formula to calculate the dynamic regions:

$$U^t = V^t_{static} \odot S + (1 - V^t_{static}) \odot M^t_{dynamic} \quad (2)$$

In addition, we also perform an ablation study where the calculations of static, varying values and dynamic map are replaced with a cross-attention mechanism. The results show that varying value contributes more in enabling DDFE to distinguish dynamic regions. Meanwhile, although the combination of cross-attention and the activation ratio provides a similar focus on the dynamic regions, it introduces additional computational cost.

2. Quantitative Results on Different Scenarios

In this section, we present more statistical comparisons on both the seen and unseen cases of the SUN-SEG test sets, as shown in Tab. 2 and Tab. 3. Note that STDDNet^1 is constructed on the Res2Net-50 backbone, while STDDNet^2 employs PVTv2-B2 as the image encoder backbone. On both the 'easy' and 'hard' test sets, most image-based segmentation networks demonstrate a more pronounced performance degradation in unseen scenarios, primarily due to their inability to model spatiotemporal consistency from video sequences. In addition, the segmentation performance of PNS+ [5] also suffers significant degradation in the unseen scenarios, as its normalized self-attention mechanism lacks sufficient training on similar scenes. Compared to other competitors, our proposed STDDNet achieves superior performance on both SUN-SEG 'easy' and 'hard' test sets, owing to its effective spatiotemporal dependency modeling via STDP and dynamic features extraction through

Module	Setting	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice
GDAU	**All components**	**88.71**	**92.11**	**82.92**	**83.79**
	No static value	88.19	91.78	82.08	83.47
	No varying value	88.12	91.80	81.94	83.25
	Cross-attention	87.66	91.45	81.51	82.93

Table 1. Ablation studies on the SUN-SEG-Hard test set, evaluating the effects of different calculations in GDAU.

Method	SUN-SEG-Easy-Seen (%)				SUN-SEG-Easy-Unseen (%)			
	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice
Vim	91.74	95.12	88.94	88.81	76.91	80.15	67.09	66.28
VMamba	92.74	96.00	89.07	90.02	78.98	83.14	68.11	68.72
Polyp-SAM	91.77	94.05	87.13	87.71	79.71	84.36	68.65	69.71
G-CASCADE	93.03	96.26	90.71	90.92	82.59	88.30	75.96	76.19
VideoMamba	93.46	96.20	88.99	90.07	83.21	88.09	76.17	77.26
PNS+	93.21	95.44	89.54	89.96	80.27	79.87	68.80	68.90
SALI	93.58	96.62	89.85	90.91	84.66	88.40	76.66	77.80
STDDNet^1	94.06	**97.16**	90.76	**91.87**	85.41	**90.91**	77.56	79.52
STDDNet^2	**94.12**	96.93	**91.22**	91.80	**85.56**	89.92	**78.93**	**79.91**

Table 2. Statistical comparison with different state-of-the-art methods on the SUN-SEG-Easy test set.

Method	SUN-SEG-Hard-Seen (%)				SUN-SEG-Hard-Unseen (%)			
	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice
Vim	85.17	89.61	79.54	79.17	77.36	82.36	66.58	66.31
VMamba	88.69	92.57	83.59	84.28	77.67	83.73	65.77	66.96
Polyp-SAM	87.89	91.48	81.88	82.71	80.95	85.31	70.08	71.61
G-CASCADE	89.61	94.14	85.79	86.43	83.03	88.83	75.03	75.80
VideoMamba	90.43	93.45	84.23	85.68	83.76	89.59	77.71	76.76
PNS+	90.77	94.16	86.41	86.98	81.51	82.16	70.48	71.00
SALI	90.31	94.44	85.74	86.93	85.44	89.92	76.82	78.85
STDDNet^1	**91.59**	**95.39**	87.19	**88.38**	85.39	90.32	76.02	78.51
STDDNet^2	91.22	94.81	**87.31**	88.35	**86.27**	**90.51**	**79.33**	**80.82**

Table 3. Statistical comparison with different state-of-the-art methods on the SUN-SEG-Hard test set.

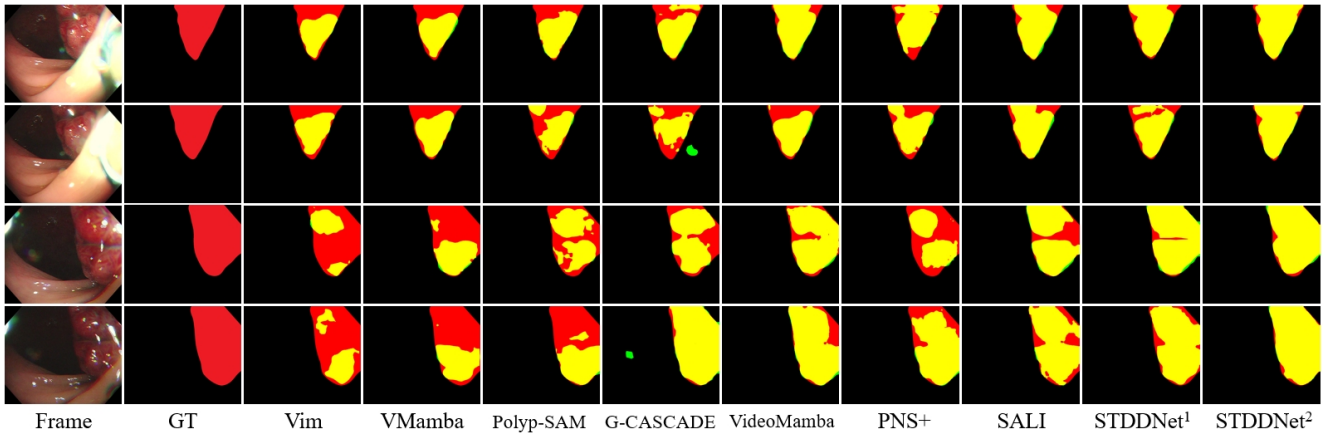

Figure 1. Visual comparisons with different state-of-the-art methods on the SUN-SEG test sets. Red, green and yellow areas represent the ground truth, prediction and their overlapping regions, respectively.

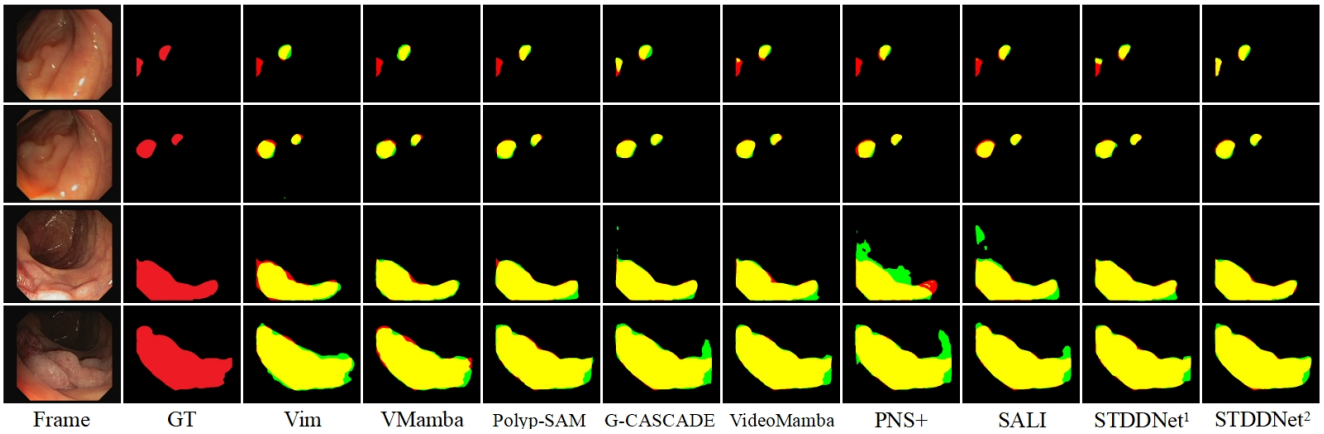

Figure 2. Visual comparisons with different state-of-the-art methods on the test set of CVC-Clinic DB. Red, green and yellow areas represent the ground truth, prediction and their overlapping regions, respectively.

DDFE. The enhanced spatiotemporal and dynamic features enable our STDDNet to address the challenges of VPS more effectively and efficiently.

3. Qualitative Results

To qualitatively evaluate the effectiveness of our proposed STDP and DDFE modules, we compare the prediction results of different components, as shown in Fig. 3. By adding STDP to the baseline network, the model is able to model temporal consistency from the video clips, thereby providing more positional information of foreground targets to improve polyp localization. In addition, by adding DDFE to the baseline network, more dynamic areas across frames are identified, particularly including the boundary details. Furthermore, the network equipped with both the STDP and DDFE modules leverages their complementary strengths for addressing VPS task, where STDP learns temporal consistency across adjacent frames and DDFE explores frame-wise dynamic differences, collectively yielding more accurate segmentation results.

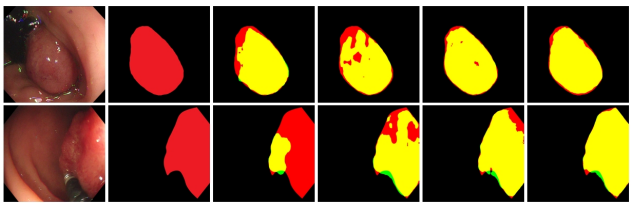

Figure 3. Segmentation results of ablation studies on the SUN-SEG test sets. (a) Input images. (b) Ground truth. (c) Baseline. (d) Model with STDP. (e) Model with DDFE. and (f) ours. Red, green and yellow areas represent the ground truth, prediction and their overlapping regions, respectively.

More visualization comparisons between our STDDNet and different state-of-the-art methods on SUN-SEG dataset

and CVC-Clinic DB are as shown in Fig. 1 and Fig. 2, respectively. Fig. 1 presents continuous segmentation results on consecutive frames with a distinct perspective alteration. The results show that VPS methods predominantly outperforms IPS methods, and our STDDNet generates refined predictions owing to the dynamic details captured by DDFE. Furthermore, even in cases with weak inter-frame continuity, STDDNet maintains prediction quality comparable to IPS methods, as illustrated in Fig. 2.

References

[1] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 1

[2] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, page 103. Association for Computational Linguistics, 2014. 1

[3] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017. 1

[4] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 698–704. International Joint Conferences on Artificial Intelligence Organization, 2018. 1

[5] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022. 1

[6] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014. 1