Discriminative Synergistic Temporal Diffusion for Ultrasound Video Thyroid Nodule Segmentation

Jialu Li[1], Hongqiu Wang[1], Junpu Hu[4], Ruiqiang Xiao[1], Ying Hu[3], Faqin Lv[4], Qiong Wang[3]✉, and Lei Zhu[1,2]

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] The Hong Kong University of Science and Technology, Hong Kong, China
[3] Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[4] Third Medical Centre of Chinese PLA General Hospital

Automatic ultrasound video object segmentation (VOS) is crucial for thyroid nodule diagnosis and treatment, which still remains challenging due to small nodules and blurred, noisy boundaries induced by tissue deformation and probe motion. Although recent diffusion VOS methods have shown superior performance by gradually reconstructing tumor structures from noise features, they typically treat heterogeneous lesion cues equally throughout the segmentation process, neglecting the fact that their contribution varies across different denoising stages. Motivated by the coarse-to-fine annotation logic of clinicians, we propose a Discriminative Synergistic Temporal Diffusion (DSTD) framework that can explicitly assess the step-wise contribution of motion and texture guidance during each denoising timestep. Specifically, a Retentive Motion Perception module is designed to capture temporally consistent motion patterns while suppressing irrelevant tissue noise, and a Discriminative Texture Perception module evaluates the confidence of different texture representations to enhance ambiguous boundaries. Furthermore, a Synergistic Controller is introduced to adaptively balance these complementary guidance cues in a timestep-dependent manner. Moreover, we construct a large-scale ultrasound thyroid nodule dataset with 374 pixel-level annotated videos. Extensive experiments on 3 medical video datasets demonstrate that DSTD consistently outperforms state-of-the-art methods, involving small nodules and blurred boundaries(link).

Keywords: Ultrasound video · Thyroid nodule segmentation · Synergistic temporal diffusion · Discriminative boundary learning

1 Introduction

Ultrasound imaging is widely adopted for clinical diagnosis and treatment due to its non-invasiveness, real-time capability, and cost-effectiveness [8–10]. However, as illustrated in Fig. 1, accurately delineating nodule regions could be a time-consuming task: Thyroid nodules are generally smaller (lower lesion pixel percentage), making them more difficult to detect. In addition, they exhibit blurred

✉ Q. Wang (wangqiong@siat.ac.cn) is the corresponding author of this work.

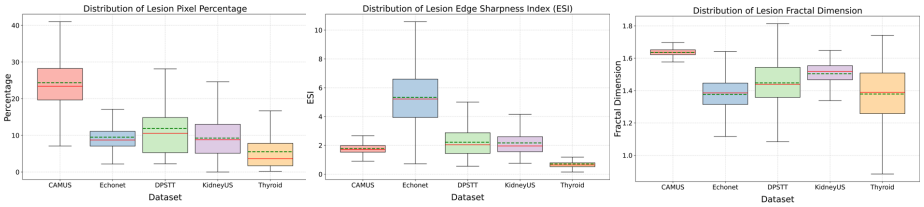

Fig. 1. Box plot of lesion characteristics across different ultrasound organ datasets [6, 8, 16, 20]. The red and green lines indicate the median and mean value, respectively.

boundaries with less pronounced grayscale boundary variations (lower ESI values) and demonstrate greater diversity in boundary shapes (longer whiskers of fractal distribution). These challenges are further exacerbated by operator-dependent probe motion and tissue deformation, leading to complex and inconsistent temporal variations across video frames.

Motivated by the observation [7, 18] that diffusion models tend to recover global coarse objects at early timesteps and then progressively refine details, which have shown stronger capability in reconstructing images from noisy and blurred scenarios. This may be particularly suitable for ultrasound videos, where clinicians tend to perform coarse lesion localization and then refine the boundary details. However, existing diffusion-based VOS methods equally and statically integrate heterogeneous lesion cues, overlooking the fact that their contribution to the prediction varies across different denoising stages.

This observation motivates our model to explicitly focus on the contributions of different guidance cues during the denoising step, rather than treating all lesion cues equally across timesteps, thereby accurately locating the tumor and optimizing boundary details. Specifically, we design a Retentive Motion Perception module to capture temporally consistent lesion patterns while suppressing irrelevant tissue motion, and a Discriminative Texture Perception module that assesses the contribution of multiple texture representations to enhance ambiguous and noisy boundaries. A Synergistic Controller is further introduced to adaptively integrate multiple guidance cues in a timestep-dependent manner, enabling robust and accurate segmentation.

In addition, we construct a large-scale ultrasound video thyroid nodule dataset with 374 annotated videos to facilitate systematic evaluation. Extensive experiments on multiple medical video datasets demonstrate that our method consistently outperforms SOTA methods, especially in challenging scenarios involving small lesions and blurred boundaries. Our work can be summarized as follows:

- We construct a large-scale dataset consisting of 374 pixel-wise annotated US thyroid nodule videos to facilitate future research on this topic.
- We propose a novel Discriminative Synergistic Temporal Diffusion (DSTD) framework to dynamically learn the contribution of different guidance cues across denoising timesteps.

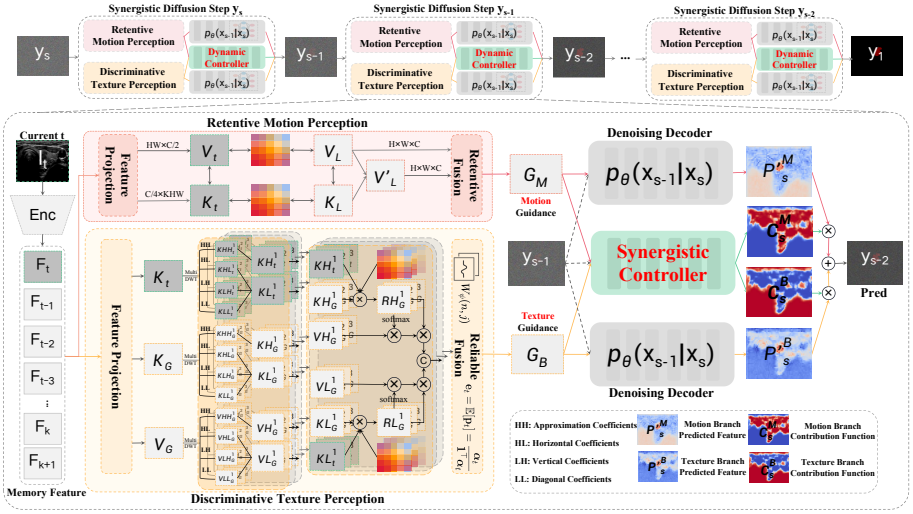

Fig. 2. Overview of our proposed DSTD framework that can dynamically evaluate the contributions of lesion boundary and motion guidance cues at each denoising timestep.

– We introduce a Discriminative Texture Perception module to explicitly assess the contribution of different wavelet lesion cues and enhance boundary robustness.
– Extensive experiments on three medical video datasets demonstrate that our method outperforms recent SOTA methods.

2 Method

As shown in Fig. 2, the current frame feature F_t together with the memory features are fed into the Retentive Motion Perception module to learn consistent tumor motion patterns (G_M), and the Discriminative Texture Perception module to selectively emphasize effective lesion texture cues and suppress unreliable feature interference (G_B). Finally, the Synergistic Controller generates refined predictions by adaptively evaluating the contributions of Retentive Motion Guidance G_M and Discriminative Texture Guidance G_B.

2.1 Retentive Motion Perception

In ultrasound videos, meaningful lesion motion could be spatially localized and temporally consistent. However, existing methods [1, 26] equally extract motion features from different memory frames, which may neglect their origin spatiotemporal relationships and disrupt relevant motion cues. Hence, we explicitly model pixel-level spatiotemporal proximity using a retentive distance matrix S_m, aiming to suppress noisy tissue interference and enhance the perception of reliable tumor motion cues among memory frames:

$$S_m = \gamma^{(|x_n - x_m| + |y_n - y_m|)}. \tag{1}$$

where (x_n, y_n) and (x_m, y_m) are the pixel-level coordinates of each frame feature, γ is a learnable attenuation parameter to control the retentive level.

The retentive distance matrix is used to reweight attended motion features, enforcing spatial consistency and suppressing irrelevant tissue motion using the key-value calculation [1, 8] of the memory and the current frame feature:

$$G_M = Proj[S \odot \big(Cat(Attn(K_t, K_M, \sum_{m=1}^{M} S_m) \odot V_M, V_t)\big)]. \tag{2}$$

where $Attn$ is the attention calculation [8], $Proj$ is the KAN layer [12], K_M-V_M are the key-value embedding of the memory feature F_M, and K_t-V_t are the key-value embedding of the current feature F_t, m is the memory frame index.

2.2 Discriminative Texture Perception

Discrete wavelet transform (DWT) could further provide multi-scale and multi-orientation texture representations [14], while different wavelet kernels exhibit distinct texture responses even within the same frequency subband [2]. However, existing DWT methods neglect to clearly distinguish the effectiveness of different texture characteristics, which may lead to performance degradation in thyroid nodules with small and inherently blurred nodule boundaries.

Hence, we propose to adaptively aggregate different texture features of wavelet kernels according to their texture response effectiveness, selectively emphasizing discriminative lesion cues while suppressing irrelevant texture components.

$$\pi = \text{argsort}_{i \in [1, t-1]} \big(\text{Cos}(F_t, F_i)\big), \quad K_M, V_M = Conv[\text{Cat}\left(F_{\pi_1}, \ldots, F_{\pi_K}\right)]. \tag{3}$$

where π denotes the index permutation that ranks the memory frames by their cosine similarity.

Then, as shown in Fig. 2, multiple DWT filters, Haar, DB3, and Dmey, are performed to explore the low-frequency structural cues and high-frequency texture features that are associated with the current frame F_t:

$${V'}_t^w = Cat\big(\hat{S}\big(\exp(KH_t^w \otimes KH_t^w)\big) \odot VH_t^w, \hat{S}\big(\exp(KL_t^w \otimes KL_t^w)\big) \odot VL_t^w\big). \tag{4}$$

where w denotes the DWT filter, KH and VH denote the high-frequency subbands H of key and value embedding, and KL and VL denote the low-frequency subbands L of key and value embedding, $\hat{S}$ denotes the Softmax function, $\otimes$ denotes the dot product and the $\odot$ denotes the element-wise multiplication.

Then, we explicitly evaluate these texture features and suppress noise-band interference using a learnable Dirichlet-based KAN layer. As a result, their effectiveness is naturally aggregated into the [0,1] interval, enabling the model to explicitly integrate diverse texture features:

$$\mathbf{e}_t = \mathbb{E}[\mathbf{p}_t] = \frac{\boldsymbol{\alpha}_t}{\mathbf{1}^\top \boldsymbol{\alpha}_t}, \quad \boldsymbol{\alpha}_t = \text{softplus}[Proj\Big({V'}_t^{Haar}; {V'}_t^{DB3}; {V'}_t^{Dmey}\Big)] + \mathbf{1}, \tag{5}$$

$$\tilde{G}_t = \sum_{w=1}^{W} e_{t,w} \odot G_t^{(w)}, \quad G_B = \mathrm{Proj}\left(S_m \odot \tilde{G}_t\right). \tag{6}$$

where $e_{t,w}$ denotes the reliability weight of the w-th wavelet kernel, and $G_t^{(w)}$ represents the corresponding texture guidance, and $Proj$ denotes a KAN Layer [12].

2.3 Synergistic Controller

Following the observations in prior works [7, 18], different guidance cues tend to contribute unequally across diffusion timesteps. Clinicians tend to perform lesion localization based on motion cues and then refine lesion structures based on reliable texture cues. Hence, previous methods that equally leverage different types of guidance features without distinguishing their effectiveness may cause interference between coarse localization and fine-grained boundary refinement.

Hence, we introduce the Synergistic Controller to estimate the contribution functions C_s^M and C_s^B at each timestep s for the coarse predictions (P_s^M and P_s^B) and generating the refined segmentation result $\hat{O}_s$:

$$P_s^M = f_{\mathrm{dec}}(s, G_M, F_t), \quad P_s^B = f_{\mathrm{dec}}(s, G_B, F_t), \tag{7}$$

$$C_s^M, C_s^B = f_{\mathrm{SC}}\left(s, P_s^M, P_s^B\right), \tag{8}$$

where f_{dec} denotes the denoising decoder, and f_{SC} is the proposed Synergistic Controller, which shares the same architecture as the denoising decoder but different learnable parameters.

The segmentation results $\hat{O}_s$ can be controlled with:

$$\hat{C}_s^M, \hat{C}_s^B = \mathrm{Softmax}\left(C_s^M, C_s^B\right), \quad \hat{O}_s = \hat{C}_s^M \odot P_s^M + \hat{C}_s^B \odot P_s^B. \tag{9}$$

where s denotes the diffusion timestep, P_s^M and P_s^B represent the segmentation predictions at timestep s, respectively. f_{SC} denotes the proposed Synergistic Controller that adaptively estimates the relative contributions of heterogeneous guidance cues, and $\hat{O}_s$ is the refined segmentation output at timestep s.

3 Experiments

3.1 Implementation

Our proposed ultrasound video thyroid nodule dataset comprises 374 video sequences with pixel-level annotations provided by experienced clinicians. The spatial resolution of the acquired ultrasound video includes axial (along the sound beam direction) and lateral (perpendicular to the sound beam direction) components, with axial resolution of 0.1–0.3 mm and lateral resolution of 0.2–0.4 mm. The temporal resolution ranges from 20 to 30 fps. The contrast resolution is set to 50–60 dB. The average size of thyroid nodules was 10.7 mm × 8.4 mm. Experienced clinical doctors use GE Logiq E9 and GE Logiq E10 color Doppler

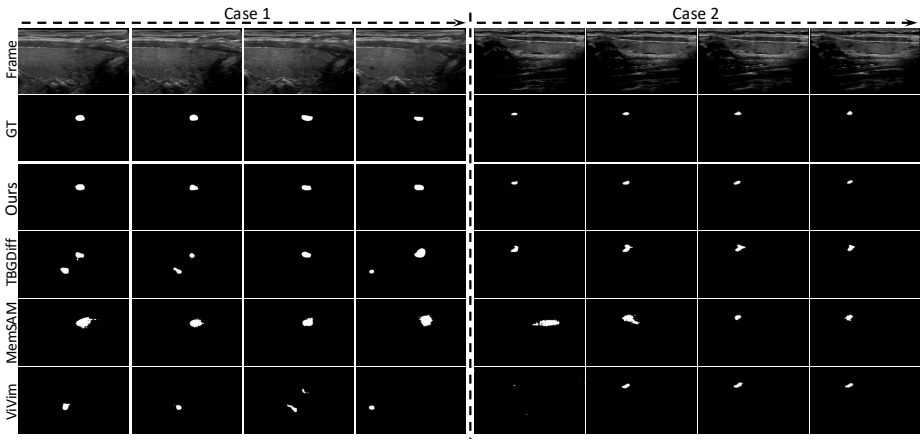

Fig. 3. Visual comparison of US thyroid nodule videos with small nodules.

Table 1. Quantitative comparison with different methods on the ultrasound video thyroid nodule segmentation dataset.

Methods	Type	Venue	Year	Jaccard	Dice	Precision	Recall
OSVOS [17]	Video	CVPR	2017	51.07	60.88	72.56	60.19
STM [15]	Video	ICCV	2019	56.18	66.64	72.38	69.36
DPSTT [8]	Video	MICCAI	2022	59.16	69.51	74.19	72.58
FLANet [11]	Video	MICCAI	2023	62.81	72.09	76.53	74.69
MemSAM [1]	Video	CVPR	2024	62.43	74.03	80.32	75.28
TBGDiff [26]	Video	ACM MM	2024	65.39	74.80	80.07	77.79
ViVim [25]	Video	TCSVT	2025	63.16	73.59	82.45	72.22
Ours	Video	—	2025	**68.50**	**78.75**	**84.22**	**79.20**

ultrasound scanners. Our dataset covers patients aged 25–78 years, with an average age of 49.8 years, and diverse thyroid diseases. 5 clinicians annotated using ITK-SNAP, with each frame annotated 3 times and pixel-wise majority voting to reduce observer variation. We employ the Segformer [23] as the diffusion backbone. We set the batch size and learning rate to 2 and 1×10^{-4}. All input frames are resized to 240×240 [3]. The model is trained using the BCE and Dice losses on 1 RTX 4090 GPU. No data augmentation is used except for the pixel normalization. We further follow original experimental settings [13, 25] to conduct experiments on the DPSTT [8] and SUN-SEG [5] medical video datasets.

Experimental Results Table 1 reports the quantitative segmentation results on the video thyroid nodule segmentation dataset. We can find that TBGDiff could achieve the Jaccard, Dice, Precision, and Recall scores of 65.39 %, 74.80%, 80.07%, and 77.79%. While our model could achieve better Jaccard, Dice, Precision, and Recall scores of 68.50%, 78.75%, 84.22%, and 79.20%. As shown in

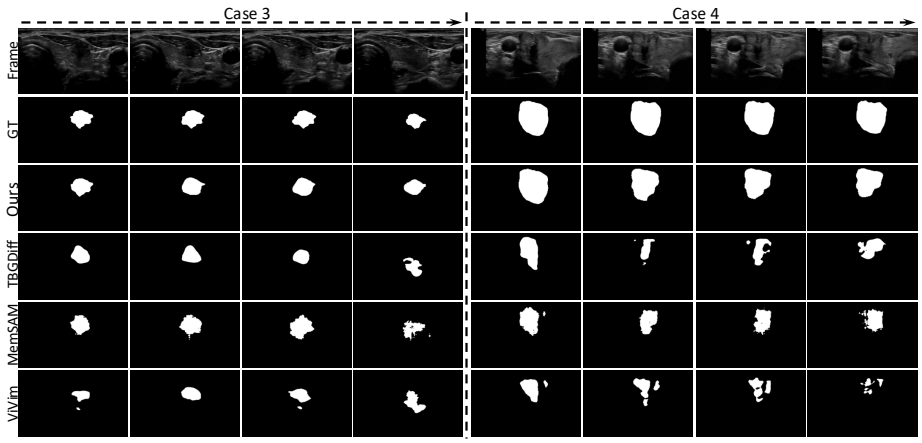

Fig. 4. Visual comparison of US thyroid nodule videos with blurred tumor boundaries.

Table 2. Quantitative comparison with different methods on the DPSTT dataset [8].

Methods	Type	Venue	Year	Jaccard	Dice	Precision	Recall
OSVOS [17]	Video	CVPR	2017	56.74	70.98	77.78	64.04
STM [15]	Video	ICCV	2019	68.58	78.62	82.01	79.10
DPSTT [8]	Video	MICCAI	2022	73.64	82.55	83.89	84.55
CIOCR [10]	Video	JBHI	2024	73.84	83.23	84.76	84.49
TBGDiff [26]	Video	ACM MM	2024	74.31	82.96	84.51	85.02
ViVim [25]	Video	TCSVT	2025	74.64	83.72	84.67	85.71
Ours	Video	—	2025	**76.12**	**85.28**	**87.19**	**86.28**

Table 2 and Table 3, we also report the segmentation performance of our method on two public medical video datasets. We can find that our method could achieve better segmentation performance than recent SOTAs on the DPSTT and the SUN-SEG datasets. Fig. 3 and Fig. 4 visualize the US thyroid nodule cases with small nodules and blurred tumor boundaries, respectively. We can find that our method could achieve better segmentation performance and is more consistent with the GT than recent methods.

Ablation Study As depicted in Table 4, we conduct a quantitative ablation study to verify the effectiveness of the Retentive Movement Perception (RMP), the Discriminative Texture Perception (DTP), and the Synergistic Controller (SC). Model 'Baseline' is constructed by replacing all modules with the original STM memory block (Memory). We observe that incorporating RMP and DTP consistently improves segmentation performance. Moreover, the proposed SC further enhances performance compared to M3 and M2. As shown in Fig. 5, we can find that after introducing the proposed SC, our model can generate more

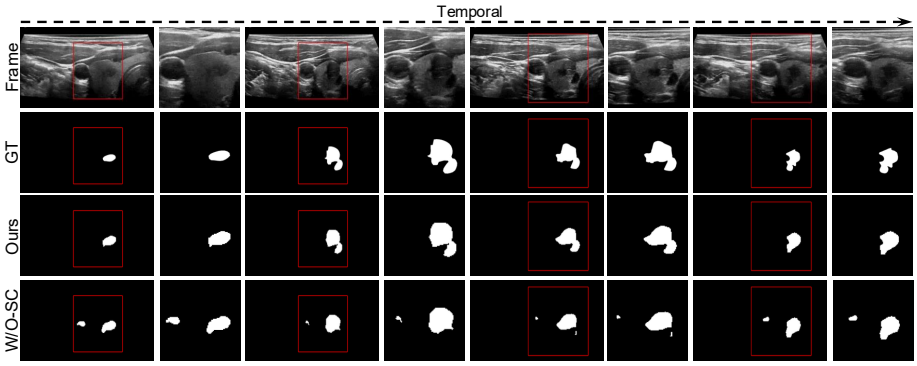

Fig. 5. Ablation visualization of the Synergistic Controller (SC) that can generate more precise tumor boundaries and suppress non-tumor areas with the Motion Perception Guidance G_M and Texture Perception Guidance G_B.

Table 3. Quantitative comparison on the SUN-SEG video polyp dataset (seen) [5].

Methods	Easy/Hard	Year	Dice	S_α	E_θ^{mn}
PNSNet [4]	Easy/Hard	2021	86.1/82.3	90.6/87.0	91.0/89.2
PNS+ [5]	Easy/Hard	2022	88.8/85.5	91.7/88.7	92.4/92.9
ICBNet [22]	Easy/Hard	2023	84.6/82.7	88.7/86.2	92.1/88.7
PVT-CASCADE [19]	Easy/Hard	2023	85.5/83.2	78.9/86.9	93.0/89.1
CTNet [21]	Easy/Hard	2024	86.7/84.0	89.9/87.3	93.9/89.5
SSTFB [24]	Easy/Hard	2024	91.1/81.1	93.5/90.9	96.8/94.2
Diff-VPS [13]	Easy/Hard	2024	90.8/86.8	93.0/90.0	96.6/94.7
Ours	Easy/Hard	2025	**92.8/88.9**	**93.8/91.2**	**98.7/97.3**

precise tumor boundaries and further suppress non-tumor areas with the Motion Perception Guidance G_M and the Texture Perception Guidance G_B.

4 Conclusion

In this paper, we construct a large-scale US thyroid nodule video dataset to facilitate future research on this topic. Then, we propose a novel Discriminative Synergistic Temporal Diffusion (DSTD) framework to dynamically adjust different types of guidance cues at different stages of the denoising process. Experimental results show that our method outperforms existing SOTA methods.

Acknowledgment. This work was supported by Guangdong Science and Technology Department (2024ZDZX2004), the Natural Science Foundation of China under Grant U23A20391, and Shenzhen Science and Technology Program (JCYJ20241202152803005, NO.KJZD20240903102717023).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Table 4. Quantitative ablation with different modules on our collected dataset.

Model	Type	Memory	RMP	DTP	SC	Jaccard	Dice	Precision	Recall
Baseline	Video	√	–	–	–	59.01	70.11	80.44	73.29
M1	Video	√	√	–	–	61.36	72.19	82.93	73.82
M2	Video	√	–	√	–	62.91	73.54	80.85	76.21
M3	Video	√	√	√	–	66.58	76.63	83.44	76.69
Ours	Video	√	√	√	√	**68.50**	**78.75**	**84.22**	**79.20**

References

1. Deng, X., Wu, H., Zeng, R., Qin, J.: Memsam: Taming segment anything model for echocardiography video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9622–9631 (2024)
2. Heil, C.: Ten lectures on wavelets (ingrid daubechies). SIAM Review **35**(4), 666–669 (1993)
3. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural networks **121**, 74–87 (2020)
4. 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)
5. 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)
6. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
7. Lee, H., Lee, H., Gye, S., Kim, J.: Beta sampling is all you need: Efficient image generation strategy for diffusion models using stepwise spectral analysis. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4215–4224. IEEE (2025)
8. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV. pp. 391–400. Springer (2022)
9. Li, J., Zhu, L., Shen, G., Zhao, B., Hu, Y., Zhang, H., Wang, W., Wang, Q.: Liver lesion segmentation in ultrasound: A benchmark and a baseline network. Computerized Medical Imaging and Graphics **123**, 102523 (2025)
10. Li, J., Zhu, L., Xing, Z., Zhao, B., Hu, Y., Lv, F., Wang, Q.: Cascaded inner-outer clip retformer for ultrasound video object segmentation. IEEE Journal of Biomedical and Health Informatics (2024)
11. 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)

12. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov–arnold networks. In: The Thirteenth International Conference on Learning Representations
13. Lu, Y., Yang, Y., Xing, Z., Wang, Q., Zhu, L.: Diff-vps: Video polyp segmentation via a multi-task diffusion network with adversarial temporal reasoning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 165–175. Springer (2024)
14. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE transactions on pattern analysis and machine intelligence **11**(7), 674–693 (1989)
15. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9226–9235 (2019)
16. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
17. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2663–2672 (2017)
18. Qian, Y., Cai, Q., Pan, Y., Li, Y., Yao, T., Sun, Q., Mei, T.: Boosting diffusion models with moving average sampling in frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8911–8920 (2024)
19. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6222–6231 (2023)
20. Singla, R., Ringstrom, C., Hu, G., Lessoway, V., Reid, J., Nguan, C., Rohling, R.: The open kidney ultrasound data set. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 155–164. Springer (2023)
21. Xiao, B., Hu, J., Li, W., Pun, C.M., Bi, X.: Ctnet: Contrastive transformer network for polyp segmentation. IEEE Transactions on Cybernetics (2024)
22. Xiao, Y., Chen, Z., Wan, L., Yu, L., Zhu, L.: Icbnet: Iterative context-boundary feedback network for polyp segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1297–1304. IEEE (2022)
23. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
24. Xu, Z., Rittscher, J., Ali, S.: Sstfb: Leveraging self-supervised pretext learning and temporal self-attention with feature branching for real-time video polyp segmentation. arXiv preprint arXiv:2406.10200 (2024)
25. Yang, Y., Xing, Z., Yu, L., Fu, H., Huang, C., Zhu, L.: Vivim: a video vision mamba for ultrasound video segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2025)
26. Zhou, H., Wang, H., Ye, T., Xing, Z., Ma, J., Li, P., Wang, Q., Zhu, L.: Timeline and boundary guided diffusion network for video shadow detection. In: ACM Multimedia 2024 (2024)