Source-Free Active Domain Adaptation for Efficient Medical Video Polyp Segmentation

Jialu Li[1], Hongqiu Wang[1], Weiming Wang[4], Jing Qin[5], Qiong Wang[3]✉, and Lei Zhu[1,2]✉⋆

[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China leizhu@ust.hk
[2] The Hong Kong University of Science and Technology, Hong Kong, China
[3] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China wangqiong@siat.ac.cn
[4] The Hong Kong Metropolitan University, Hong Kong, China
[5] The Hong Kong Polytechnic University, Hong Kong, China

Deep learning models have shown remarkable performance in medical video object segmentation. However, addressing the cross-center domain issue is crucial for achieving consistent performance across different medical facilities. Emerging Source-Free Active Domain Adaptation (SFADA) techniques can enhance the performance of target domain segmentation models, ensuring data privacy and security. While current approaches primarily focus on image-level tasks and mainly emphasize intra-frame pixel correlations, they overlook temporal correlations, which restricts their performance in video frame recommendation. Consequently, this paper proposes the first video-level SFADA method and evaluates it on video polyp segmentation across different data centers. Specifically, the Spatial-Temporal Active Recommendation (STAR) strategy is devised to recommend a few highly valuable frames for annotation by comprehensively evaluating the object spatial correlation and temporal movement density across different video frames, along with a Passive Phase Correction (PPC) module is proposed to suppress the noisy source disruptions of the remaining unlabeled data during the fine-tuning stage. Experimental results demonstrate that with a tiny quantity of annotation, our method significantly improves performance over the lower bound and achieves better performance than existing SOTA methods, which is valuable for practical clinical employment (link).

Keywords: Source-free active domain adaptation · domain adaptation · multi-center dataset· video polyp segmentation

1 Introduction

Medical segmentation is crucial for clinical diagnosis and treatment, as it automates lesion identification, thereby enhancing healthcare efficiency [18,22,24,26]. As an important early detection and treatment technique, deep learning-based

⋆ Q. Wang and L. Zhu are joint corresponding authors.

video polyp segmentation [12] could help reduce the risk of colorectal cancer. However, during clinical diagnosis and treatment, the diversity of imaging devices and patient populations may lead to obvious domain shifts [27], which may result in accuracy degradation. Moreover, the per-frame pixel-wise annotation for polyp videos could be costly and inefficient.

Unsupervised Domain Adaptation (UDA) [27] can partially alleviate the domain gap using labeled source-domain data and unlabeled target-domain data. However, it overlooks critical issues of data privacy and security. Although Source-Free Domain Adaptation (SFDA) [28] can alleviate the aforementioned bottleneck, the accuracy improvement on target domain data could be constrained due to the lack of real clinician labels. These challenges hinder the practical employment in clinical diagnosis and treatment.

Recently, the emerging Source-Free Active Domain Adaptation (SFADA) paradigm [13,14,21,23] could mitigate domain shift while ensuring data privacy by annotating only a small set of actively selected target-domain samples. This approach achieves superior performance compared to traditional UDA methods. By minimizing manual annotations, SFADA could reduce clinicians' workload and annotation costs, making it highly practical for real-world medical applications. However, previous SFADA methods [14,21,23] primarily focus on image-level tasks and mainly emphasize intra-frame pixel correlations but overlook temporal correlations, restricting their performance in video frame recommendation.

Therefore, we propose the first SFADA method specifically designed for medical video object segmentation. To evaluate our method, we assembled a multi-center video polyp segmentation (MC-VPS) dataset by leveraging and integrating existing open-source medical imaging resources [1,2,7].

The main contributions of this work can be summarized as follows:

- To the best of our knowledge, we propose the first SFADA method for medical video object segmentation and conduct a leading exploration of multi-center video polyp segmentation scenarios.
- We devise a Spatial-Temporal Active Recommendation (**STAR**) strategy that systematically evaluates the object's spatial correlation and temporal movement density across spatial and temporal dimensions. By doing so, we can actively recommend and annotate the most unreliable video frames, thereby broadening the knowledge boundary of the target model.
- We further propose the Passive Phase Correction (**PPC**) module to collaboratively leverage the rest of the unlabeled video frames by suppressing the noisy source disruptions. This synergizes with STAR's active annotation to ensure comprehensive utilization of both labeled and unlabeled data.
- We organize the first multi-center video polyp segmentation dataset (**MC-VPS**) to conduct research on this topic. Extensive experimental results demonstrate that our method achieves better segmentation performance than existing methods, which is valuable for clinical practice.

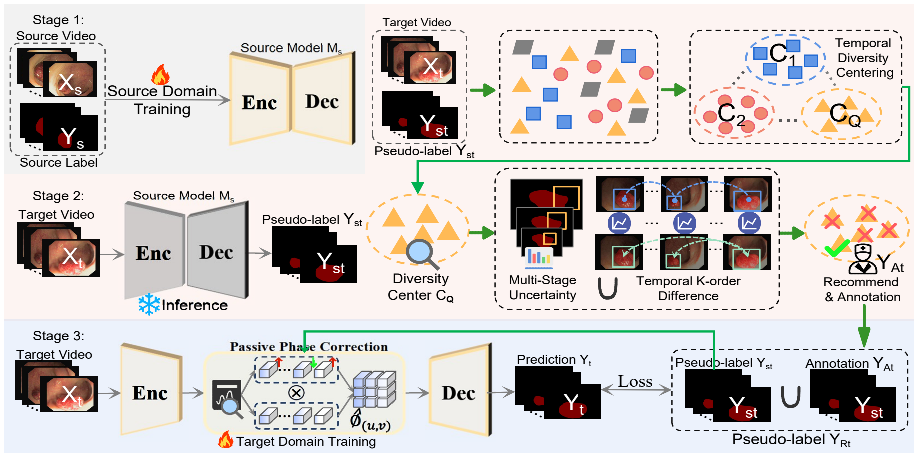

Fig. 1. Overview of the proposed video-based SFADA method. The first row of the gray area represents the source domain training stage. Then the second row of red area represents our proposed STAR strategy that recommends and annotates valuable video frames. The third row of the blue area represents the target domain fine-tuning using the combined pseudo-label.

2 Method

2.1 Problem setting

Given a sequence of video dataset D which contains video frames X and their corresponding masks Y. To protect data privacy and security, Source Free Active Domain Adaptation (SFADA) methods prohibit the accessibility of source domain datasets $\mathbf{D_s}$ to generate target domain prediction $\mathbf{Y_t}$, which only selects a few highly valuable target domain frames $\mathbf{X_t}$ for annotation to achieve an ideal segmentation performance. Ideal SFADA methods should achieve better target domain segmentation results with a lower selection quantity.

2.2 Pipeline of our SFADA

Recent SFADA methods [14, 21, 23] primarily focus on image-level tasks and mainly emphasize intra-frame pixel correlations but overlook temporal correlations, restricting their performance in video frame recommendation. This not only wastes valuable annotation resources but also ultimately degrades the target domain performance.

Therefore, we propose the first video-level SFADA method to select tiny valuable video frames $\mathbf{X_{At}}$ and then integrate their actively annotated labels $\mathbf{Y_{At}}$ into refined pseudo-labels $\mathbf{Y_{Rt}}$ for target model $\mathbf{M_t}$ fine-tuning. **1)** As depicted in the first row of the gray area in Fig. 1, the segmentation model is first trained on the source domain dataset $[\mathbf{X_s}, \mathbf{Y_s} \in \mathbf{D_s}]$ to obtain the source domain model $\mathbf{M_s}$. **2)** For the second row of the red area, the frozen source model

M_s is utilized to predict target domain pseudo-label $\mathbf{Y_{st}}$. Then, the Spatial-Temporal Active Recommendation (**STAR**) strategy is proposed to annotate a few highly valuable video frames by systematically evaluating the object's spatial correlation and temporal movement density across spatial and temporal dimensions. Then, the Passive Phase Correction (**PPC**) module is proposed to collaboratively leverage the rest of the unlabeled frames by suppressing the noisy disruptions. This complements STAR's active annotation strategy, ensuring the comprehensive utilization of both labeled and unlabeled target domain data. **3)** For the third row of the blue area, the actively annotated labels Y_{At} and remaining pseudo-labels Y_{st} are combined as the refined pseudo-label $\mathbf{Y_{Rt}}$, which is then utilized to fine-tune the target domain model M_t.

2.3 Spatial-Temporal Active Recommendation (STAR) strategy

Here we will elaborate on our STAR strategy that comprehensively evaluates the object's spatial correlation and temporal movement density across spatial and temporal dimensions to recommend a few highly valuable video frames.

Cascaded Convincing Prediction Considering that medical datasets often require comprehensive evaluation from multiple experienced physicians, we leverage the multi-layer decoder output features to produce more convincing target domain pseudo-labels $\mathbf{Y_{st}}$ along with their uncertainty maps $\mathbf{U_{st}}$ that is more consistent with the actual annotation method of medical data:

$$Y_{st} = [\sum_{g=1}^{G}(Y_{st}^{g} \otimes U_{st}^{g})]/G, \quad U_{st} = [\sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{w=1}^{W}(U_{st}^{ghw})]/G. \tag{1}$$

where $\otimes$ denotes the pixel multiplication, G is the number of multi-layer decoder output features, Y_{st}^{g} and U_{st}^{g} are their output features and uncertainty maps.

Temporal Diversity Centering Recent SFADA [14, 21, 23] methods mainly focus on image-level tasks, which may lead to sub-optimal frame recommendations due to the lack of spatial-temporal representations, ultimately undermining segmentation performance.

In order to break through the spatial-temporal limitation aforementioned, we propose to first cluster the target domain samples $\mathbf{X_t}$ into $\mathbf{Q}$ clusters, then recommend the most valuable frame from each clustering center C_q:

$$\{C_1, C_2, \ldots, C_Q\} = \sum_{q=1}^{Q}\sum_{x_t \in C_q} |x_t - \mu_q|^2, \tag{2}$$

where Q denotes the number of clusters, C_q denotes the q-th cluster, x_t denotes the target domain video frame, and μ_q denotes the centroid of the q-th cluster [4].

K-order Spatial-Temporal Reliability To systematically evaluate the object's spatial correlation and temporal movement density across spatial and temporal dimensions, we propose to calculate the Spatial-Temporal Reliability $\Delta_k({}^{n}R_{st})$ and recommend the most unreliable video frames, thereby broadening the knowledge boundary of the target model.

Given the target domain frame ${}^{n}X_t$ with its frame number n in the video sequence N, along with its uncertainty map ${}^{n}U_{st}$. The Spatial-Temporal Reliability $\mathbf{{}^{n}R_{st}}$ for the target domain frame ${}^{n}X_t$ can be calculated as:

$$ {}^{n}R_{st} = [\sum_{h=1}^{H}\sum_{w=1}^{W}({}^{n}U_{st} + |{}^{n}X_t - {}^{n-1}X_t|)]/(H \times W). \tag{3} $$

where ${}^{n}U_{st}$ denotes the uncertainty map of the n-th frame , ${}^{n}X_t$ denotes the n-th frame, and ${}^{n-1}X_t$ denotes the (n-1)-th frame.

Although $\mathbf{{}^{n}R_{st}}$ can evaluate the quality of each individual frame along both the spatial and temporal dimensions, we consider it may lack the ability to evaluate the fluctuation degrees of these segmentation qualities from the temporal perspective, as frames with larger fluctuation degrees may exhibit controversy. Hence, we further apply the differential operator to the Spatial-Temporal Reliability ($\mathbf{\Delta_k}(\mathbf{{}^{n}R_{st}})$) and quantify the fluctuation degree [9, 20]:

$$ \Delta_k({}^{n}R_{st}) = \begin{cases} {}^{n}R_{st}, & \text{if } k = 0 \\ \Delta_k({}^{n}R_{st}) - \Delta_{k-1}({}^{n}R_{st}), & \text{if } k > 0 \end{cases} \tag{4} $$

where Δ_k denotes the K-order Difference calculation.

2.4 Passive Phase Correction (PPC) module

Although afore proposed STAR strategy could bridge the domain gap by recommending a few valuable frames, the remaining unlabeled samples still contain the source domain knowledge bias.

Hence, we propose the PPC module to collaboratively leverage the rest of the unlabeled video frames by suppressing the noisy disruptions. This synergizes with STAR's active annotation to ensure comprehensive utilization of both labeled and unlabeled data.

As shown in Fig. 1, the encoder feature map is first projected into the frequency domain to obtain its corresponding phase and amplitude spectrum, $\Phi(u, v)$ and $\mathcal{M}(u, v)$. The phase feature and amplitude feature hold the structural prior and texture information [3, 5] of the image feature as understood by the source model M_s, respectively. Hence, a learnable matrix W_Φ is utilized during the target domain fine-tuning that can suppress negative components and emphasize valuable components related to the target domain [8]:

$$ \hat{\Phi}(u, v) = \Phi(u, v) \otimes S\big(W_\Phi\big). \tag{5} $$

where S denotes the sigmoid operation, $\otimes$ denotes the multiply operation and W_Φ denotes the learnable weight matrix. Then frequency components are passed through the $\mathcal{IFFT}$ [17] to generate the spatial domain feature $\hat{X}$ for the decoder:

$$ \hat{X} = \mathcal{IFFT}(\mathcal{M}(u, v)e^{i\hat{\Phi}(u,v)}). \tag{6} $$

Table 1. Quantitative analysis of the Multi-Center Domain Adaptation Video Polyp Segmentation dataset (MC-VPS).

Source/ Target	Dataset	Minimum Resolution	Maximum Resolution	Pixel-wise Annotation
Source	CVC-ColonDB [2]	574x500	574x500	√
Center A	CVC-ClinicDB [1]	384x288	384x288	√
Center B	SUN-SEG [7]	1158x1008	1240x1080	√

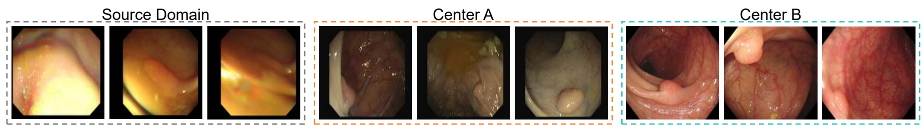

Fig. 2. Visualization of the domain gap between different data centers.

3 Experiments

3.1 Dataset & Implementation

Due to the lack of a public multi-center video polyp segmentation dataset, to validate our method, we organize a multi-center video polyp segmentation dataset (MC-VPS) by integrating existing open-source medical imaging resources [1, 2, 7]. As shown in Table 1, this dataset includes CVC-ClinicDB [1, 15], CVC-ColonDB [2], and SUN-SEG [7]. Fig. 2 further visualizes the domain gap between different centers. Then, following previous SFADA works [14, 21, 23], we compare with the SFADA and SFDA methods for a comprehensive comparison and further report their segmentation performance in Table 2 and Table 3. CVC-ColonDB is used as the source domain dataset D_s. CVC-ClinicDB and SUN-SEG are used as the Center A and Center B target domain datasets. Following previous video polyp segmentation works [7], four popular evaluation metrics are used to evaluate the performance of these methods, including S_α, E_θ^{mn}, Jaccard, and Dice. For consistency, our method and compared SFADA methods are re-implemented using the popular STM memory block [16] for the Segformer backbone [25] with the normal DDIM process [11, 19, 29] as the segmentation architecture to evaluate these methods. We implement our method using PyTorch and an RTX 3090 GPU. Image size and learning rate are set to 240×240 and 1e-4. Adam optimizer is used to minimize the Dice and BCE loss.

3.2 Experimental Results

Table 2 and Table 3 quantitatively report the experimental results across different video polyp data centers, including the lower bound (source model without fine-tuning) and upper bound (source model fine-tuned with all target-domain labels), and various state-of-the-art methods. Fig. 3 qualitatively compares the segmentation results of our method with recent state-of-the-art methods. It is obvious from Table 2 and Table 3 that notable performance disparities exist

Table 2. Quantitative comparison on Dice and Jaccard of our method and other state-of-the-art methods on the MC-VPS.

	Dice (mean%±var)			Jaccard (mean%±var)		
Methods	Center A	Center B	Overall	Center A	Center B	Overall
Lower bound	67.39±0.029	58.30±0.041	62.85±0.035	60.29±0.044	52.85±0.036	56.57±0.040
Upper bound	79.44±0.005	71.33±0.060	75.39±0.033	69.17±0.009	62.54±0.071	65.86±0.040
FSM [28]	70.18±0.033	60.69±0.049	65.44±0.041	59.87±0.054	49.79±0.044	54.83±0.049
Random	69.82±0.041	60.22±0.046	65.02±0.044	59.18±0.042	48.79±0.059	53.99±0.051
LC [6]	71.12±0.017	61.06±0.043	66.09±0.030	60.30±0.034	49.15±0.039	54.73±0.037
SALAD [10]	71.80±0.013	63.79±0.042	67.80±0.028	61.10±0.030	52.05±0.037	56.58±0.034
UGTST [14]	74.46±0.014	63.35±0.039	68.91±0.027	64.34±0.025	52.48±0.054	58.41±0.040
CUP [23]	73.99±0.013	64.75±0.051	69.37±0.032	63.41±0.019	52.68±0.047	58.05±0.033
STDR [21]	72.98±0.011	62.53±0.035	67.76±0.023	61.27±0.018	49.98±0.025	55.63±0.022
Ours	76.42±0.010	66.42±0.042	71.42±0.026	66.00±0.014	56.33±0.043	61.17±0.029

Table 3. Quantitative comparison on S_α and E_θ^{mn} of our method and other state-of-the-art methods on the MC-VPS.

	S_α (mean%±var)			E_θ^{mn} (mean%±var)		
Methods	Center A	Center B	Overall	Center A	Center B	Overall
Lower bound	75.33±0.016	71.55±0.013	73.44±0.015	81.35±0.027	77.26±0.021	79.31±0.024
Upper bound	83.46±0.04	80.80±0.013	82.13±0.027	90.10±0.015	85.28±0.032	87.69±0.024
FSM [28]	77.54±0.018	73.04±0.018	75.29±0.018	81.82±0.026	77.33±0.025	79.58±0.026
Random	77.30±0.009	73.10±0.015	75.20±0.012	82.91±0.032	78.94±0.020	80.93±0.026
LC [6]	77.88±0.009	72.99±0.016	75.44±0.013	83.25±0.011	79.46±0.024	81.36±0.018
SALAD [10]	78.30±0.007	74.42±0.013	76.36±0.010	83.85±0.007	79.55±0.020	81.70±0.014
UGTST [14]	79.97±0.007	75.63±0.013	77.80±0.010	86.44±0.008	81.26±0.016	83.85±0.012
CUP [23]	80.19±0.006	75.48±0.019	77.84±0.013	85.46±0.006	78.18±0.027	81.82±0.017
STDR [21]	78.17±0.005	73.22±0.012	75.70±0.009	85.52±0.004	77.63±0.022	81.58±0.013
Ours	81.33±0.005	77.53±0.013	79.43±0.009	87.39±0.004	83.05±0.020	85.22±0.012

between the lower and upper bounds across various popular evaluation metrics. Especially for the case of Dice, the overall performance gap exists from 62.85% to 75.39%. We further evaluate our method against various recent state-of-the-art methods under identical video object segmentation (VOS) architectures and experimental conditions, all evaluated methods are assigned the same experimental setup with 5% target-domain labeled data. Compared with recent SOTA methods that focus mainly on spatial dimension selection, our strategy demonstrated better segmentation performance on popular evaluation metrics, all underscoring the efficiency of our spatial-temporal-based approach augmented by the STAR selection strategy and PPC module.

3.3 Ablation Studies

As shown in Table 5, we conduct various ablation experiments on the MC-VPS dataset to evaluate the effectiveness of each component in our proposed method. We consider four baseline networks: 1) M1 randomly selects video frames and then performs pixel-wise annotation, 2) M2 incorporates the STAR strategy

Table 4. Ablation study for our proposed modules on target domain Center B

Methods	Dice	Jaccard	S_α	E_θ^{mn}
M1	60.22±0.046	48.79±0.059	73.10±0.015	78.94±0.020
M2	64.19±0.035	53.99±0.034	76.15±0.012	80.59±0.020
M3	65.24±0.042	53.84±0.046	75.47±0.013	80.17±0.019
Ours	66.42±0.042	56.33±0.043	77.53±0.013	83.05±0.020

Table 5. Ablation study for the annotation percentage on target domain Center B.

Methods	Dice	Jaccard	S_α	E_θ^{mn}
2%	62.46±0.041	50.93±0.038	73.91±0.014	79.23±0.019
5% (Ours)	66.42±0.042	56.33±0.043	77.53±0.013	83.05±0.020
10%	68.55±0.047	58.63±0.042	77.55±0.013	85.70±0.014
15%	70.35±0.062	61.71±0.054	80.65±0.016	85.35±0.031

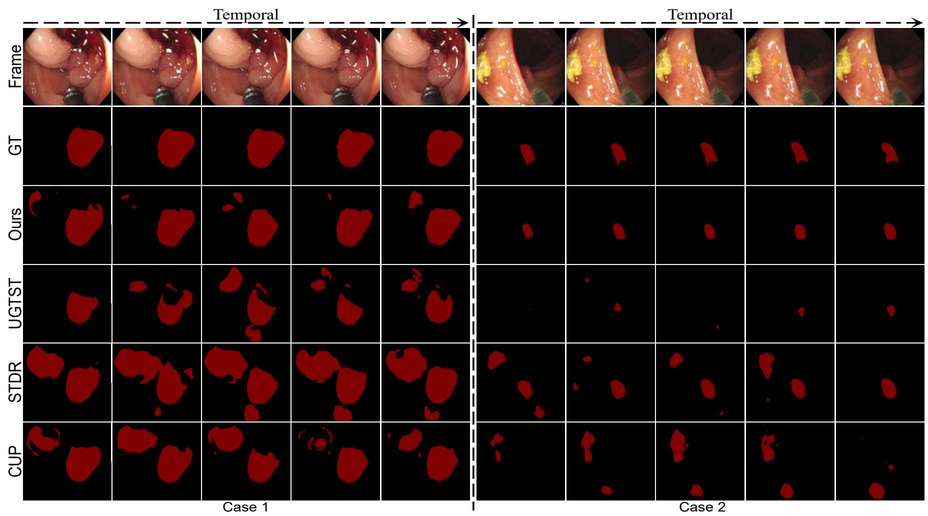

Fig. 3. Visual comparisons of our proposed method and recent SOTA methods. Our method could more accurately and consistently segment the video polyps.

without the Difference calculation. 3) M3 replaces the random selection strategy with STAR and Difference calculation. 4) M4 is constructed by further adding the PPC module. Note that M2 and M3 are parallel ablation settings based on M1, not progressive experimental settings. M3 introduces no new modules and aims to verify the impact of different Spatial-Temporal Reliability representations, rather than ablating the Difference operation alone. We can find that M2 and M3 outperform M1, and Model Ours achieves better performance than other ablation models. Note that SFADA is an offline task to recommend valuable frames for clinical annotation, and the recommendation speed of STAR is 20.8 FPS. Hence, it will not affect the practical annotation process, and has the potential to save 95% of the clinical annotation workload and achieve performance close to full annotation. Table 4 reports the impact of different active annotation ratios on the segmentation results. We can find that a larger active annotation ratio (15%) can improve the model segmentation results.

4 Conclusion

In this paper, we propose the first video-level SFADA method and evaluate it on video polyp segmentation across different data centers. Considering that recent SFADA methods mainly focus on image-level tasks, which may have the limitations of spatial-temporal representations. Hence, we propose the STAR strategy to efficiently recommend valuable video frames, along with the PPC module to suppress the source noisy component that is irrelevant to the target domain. Moreover, we built the MC-VPS dataset to facilitate related research. Experimental results demonstrate that our method achieves better performance than recent SOTA methods.

Acknowledgement. This work was supported by the National Key R&D Program of China (No.2023YFB4705700), the Natural Science Foundation of China (U24A20278), Shenzhen Science and Technology Program (JCYJ20241202152803 005), the Guangdong Science and Technology Department (No.2024ZDZX2004), Shenzhen High-tech Zone Development Special Plan Innovation Platform Construction Project, the proof of concept center for high precision and high resolution 4D imaging, and the Research Grants Council of the Hong Kong Special Administrative Region, China (No.UGC/FDS16/E02/23).

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

References

1. J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.
2. J. Bernal, J. Sánchez, and F. Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.
3. G. Chen, P. Peng, L. Ma, J. Li, L. Du, and Y. Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 458–467, 2021.
4. R. Fabbri, L. D. F. Costa, J. C. Torelli, and O. M. Bruno. 2d euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys (CSUR)*, 40(1):1–44, 2008.
5. R. C. Gonzales and P. Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.
6. T. He, X. Jin, G. Ding, L. Yi, and C. Yan. Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 1360–1365. IEEE, 2019.
7. G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022.
8. M. Jiang, P. Zeng, K. Wang, H. Liu, W. Chen, and H. Liu. Fecam: Frequency enhanced channel attention mechanism for time series forecasting. *Advanced Engineering Informatics*, 58:102158, 2023.
9. W. G. Kelley and A. C. Peterson. *Difference equations: an introduction with applications*. Academic press, 2001.

10. D. Kothandaraman, S. Shekhar, A. Sancheti, M. Ghuhan, T. Shukla, and D. Manocha. Salad: Source-free active label-agnostic domain adaptation for classification, segmentation and detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 382–391, 2023.
11. M. Li, X. Lang, R. Gong, J. Zhou, X. Yang, and N. Sang. Tpsegmentdiff: An enhanced diffusion model for tactile paving image segmentation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, pages 1–6, 2024.
12. Y. Lu, Y. Yang, Z. Xing, Q. Wang, and L. Zhu. 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*, pages 165–175. Springer, 2024.
13. X. Luo, H. Wang, J. Xu, L. Li, Y. Zhao, Y. He, H. Huang, J. Xiao, T. Song, S. Zhang, et al. Generalizable magnetic resonance imaging-based nasopharyngeal carcinoma delineation: Bridging gaps across multiple centers and raters with active learning. *International Journal of Radiation Oncology* Biology* Physics*, 121(5):1384–1393, 2025.
14. Z. Luo, X. Luo, Z. Gao, and G. Wang. An uncertainty-guided tiered self-training framework for active source-free domain adaptation in prostate segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–117. Springer, 2024.
15. D. Ninja. Visualization tools for cvc-clinicdb dataset. `https://datasetninja.com/cvc-612`, jun 2025. visited on 2025-06-20.
16. S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9226–9235, 2019.
17. H. M. Ozaktas, M. A. Kutay, and D. Mendlovic. Introduction to the fractional fourier transform and its applications. In *Advances in imaging and electron physics*, volume 106, pages 239–291. Elsevier, 1999.
18. Y. Pang, Y. Li, T. Huang, J. Liang, Z. Ding, H. Chen, B. Zhao, Y. Hu, Z. Zhang, and Q. Wang. Efficient breast lesion segmentation from ultrasound videos across multiple source-limited platforms. *IEEE Journal of Biomedical and Health Informatics*, 2025.
19. J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*.
20. M. Taniguchi. *Higher order asymptotic theory for time series analysis*, volume 68. Springer Science & Business Media, 2012.
21. H. Wang, J. Chen, S. Zhang, Y. He, J. Xu, M. Wu, J. He, W. Liao, and X. Luo. Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. *IEEE Transactions on Medical Imaging*, 2024.
22. H. Wang, Y. Chen, W. Chen, H. Xu, H. Zhao, B. Sheng, H. Fu, G. Yang, and L. Zhu. Serp-mamba: Advancing high-resolution retinal vessel segmentation with selective state-space model. *arXiv preprint arXiv:2409.04356*, 2024.
23. H. Wang, X. Luo, W. Chen, Q. Tang, M. Xin, Q. Wang, and L. Zhu. Advancing uwf-slo vessel segmentation with source-free active domain adaptation and a novel multi-center dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 75–85. Springer, 2024.
24. H. Wang, G. Yang, S. Zhang, J. Qin, Y. Guo, B. Xu, Y. Jin, and L. Zhu. Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. *IEEE Transactions on Medical Imaging*, 2024.

25. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.
26. Z. Xing, L. Wan, H. Fu, G. Yang, Y. Yang, L. Yu, B. Lei, and L. Zhu. Diff-unet: A diffusion embedded network for robust 3d medical image segmentation. *Medical Image Analysis*, page 103654, 2025.
27. J. Yan, H. Zhu, T. Hou, N. Chen, W. Lu, Y. Wang, and B. Huang. Mbda-net: Multi-source boundary-aware prototype alignment domain adaptation for polyp segmentation. *Biomedical Signal Processing and Control*, 96:106664, 2024.
28. C. Yang, X. Guo, Z. Chen, and Y. Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79:102457, 2022.
29. H. Zhou, H. Wang, T. Ye, Z. Xing, J. Ma, P. Li, Q. Wang, and L. Zhu. Timeline and boundary guided diffusion network for video shadow detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 166–175, 2024.