HRVVS: A High-resolution Video Vasculature Segmentation Network via Hierarchical Autoregressive Residual Priors

Xincheng Yao[1], Yijun Yang[1,†], Kangwei Guo[2], Ruiqiang Xiao[1], Haipeng Zhou[1], Haisu Tao[2], Jian Yang[2], and Lei Zhu[1,3] ✉

[1] The Hong Kong University of Science and Technology (Guangzhou), China
yyang018@connect.hkust-gz.edu.cn
[2] Southern Medical University, Guangzhou, China
[3] The Hong Kong University of Science and Technology, Hong Kong, China

Abstract. The segmentation of the hepatic vasculature in surgical videos holds substantial clinical significance in the context of hepatectomy procedures. However, owing to the dearth of an appropriate dataset and the inherently complex task characteristics, few researches have been reported in this domain. To address this issue, we first introduce a high quality frame-by-frame annotated hepatic vasculature dataset containing 35 long hepatectomy videos and 11442 high-resolution frames. On this basis, we propose a novel high-resolution video vasculature segmentation network, dubbed as HRVVS. We innovatively embed a pretrained visual autoregressive modeling (VAR) model into different layers of the hierarchical encoder as prior information to reduce the information degradation generated during the downsampling process. In addition, we designed a dynamic memory decoder on a multi-view segmentation network to minimize the transmission of redundant information while preserving more details between frames. Extensive experiments on surgical video datasets demonstrate that our proposed HRVVS significantly outperforms the state-of-the-art methods. The source code and dataset will be publicly available at https://github.com/scott-yjyang/HRVVS.

Keywords: Video Vasculature Segmentation · High-resolution · Visual Autoregressive Modeling.

1 Introduction

Hepatectomy is a set of surgical procedures for local liver lesions, such as liver tumors, liver injuries, liver abscesses, and etc.. Given the rich blood supply in the liver, effective control of bleeding during surgery is pivotal for the success of liver resection [2,20]. Specifically, during the operation, surgeons need to focus on two types of blood vessels, the Glisson sheath and the hepatic vein. The segmentation of the hepatic vasculature in surgical videos can provide precise positioning for

[†] Project Lead.

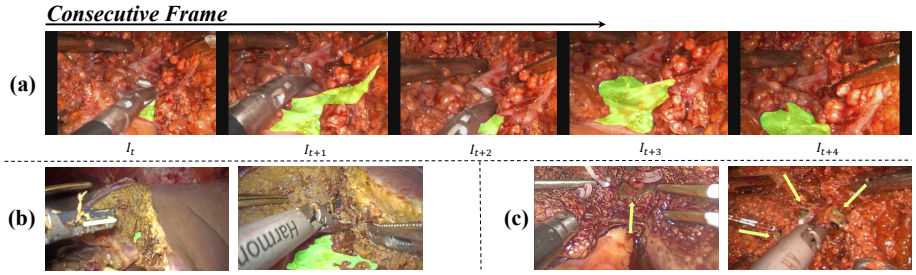

Fig. 1. Main challenges in hepatic vasculature segmentation. Fluorescent green shadow is used to show the location of vasculature in **(a)** and **(b)**, the arrow and green outline are used to point to the corresponding location in **(c)**. **(a)** Discontinuities between frames and abrupt positional transformations. **(b)** Significant variations of vessels in different contexts. **(c)** Similarities in the outline of vessels and surrounding tissue and segments by the surrounding tissue.

surgeons to prevent surgical bleeding by hemoclips during hepatectomy, which makes it has great clinical significance. Nevertheless, the fat and muscle around the vasculature generate significant redundant information, making the model difficult to segment the correct tissue in the video. Previous works [9, 24, 25] in hepatic vascular segmentation mainly focused on medical images from CT or MRA before surgery. However, these methods can not directly pinpoint the location of vasculature during the actual surgical procedure. Therefore, we are considering building a high-resolution video segmentation network to address this issue. Through collaboration with the hospital, we collect a dataset containing 35 long high-resolution videos with a total of 11442 frames, which provided the foundation for our model training. To the best of our knowledge, ours is the first work dedicated to this task.

In recent years, numerous video segmentation methods have emerged for medical imaging, exemplified by approaches such as Vivim [27, 28], which utilizes a state space model, and Ji *et al.* introduce SUN-SEG dataset for polyp segmentation in colonoscopy videos, alongside the PNS+ algorithm [12]. Furthermore, the advent of SAM2 [19] has inspired a series of video segmentation techniques [15, 31], demonstrating remarkable efficacy in medical video segmentation. However, these methods and their associated datasets are not optimized for high-resolution tasks, and their performance is often compromised in complex surgical scenarios. In our dataset, the segmentation of hepatic vasculature presents specific challenges, as illustrated in Fig 1. These include *frame discontinuities* and *abrupt positional changes* (Fig 1 (a)), *significant variations in vessel appearance* due to differing anatomical contexts and imaging conditions (Fig 1 (b)), and the *similarity between vessel outlines and surrounding tissue*, which can lead to segmentation errors (Fig 1 (c)). These factors complicate the task of maintaining segmentation continuity and accuracy across frames.

To address these issues, our approach in designing a high-resolution surgical video segmentation model focuses on two critical aspects: preserving detailed features within the current frame and minimizing computational load and cumulative errors from redundant frame-to-frame propagation. For the former, we employ a pretrained VAR model [21] as residual priors within a hierarchical encoder framework, refining it through adapter-based training. For the latter, we introduce a dynamic memory decoder featuring a Multi-view Spatiotemporal Interaction Module (MSIM) and a Dynamically Weighted Fusion Module (DWFM). Our method demonstrates state-of-the-art performance when benchmarked against the latest segmentation approaches, effectively overcoming the identified challenges associated with high-resolution tasks in complex surgical environments.

In conclusion, our contributions are: (1) **We develop a high-resolution video segmentation model for hepatic vasculature**, demonstrating the effectiveness of VAR in segmentation tasks. (2) **We introduce the first high-resolution video hepatic vasculature segmentation dataset under surgical scene**, which can be seen as a benchmark dataset for a completely new task. (3) Extensive experiments have been conducted on our dataset, **demonstrating the superiority of our proposed method**.

2 Method

2.1 Overview

Fig 2 shows the overall framework of the proposed segmentation model. For a high-resolution frame input, we extract its multi-level features by a dual-branch encoder based on VAR and Swin Transformer [16]. To tackle the aforementioned challenges (Fig 1), we proposes a memory-augmented decoder that integrates long short-term memory architecture, comprising a Multi-view Spatiotemporal Interaction Module (MSIM) and a Dynamic Weights Fusion Module (DWFM). MSIM preliminarily updates the local, global, and historical features through multi-dimensional spatiotemporal feature interaction mechanisms. The updated local and global features from MSIM will be sent into the multi-level decoder, which will also have the residual input from the corresponding layers of the multi-view encoder. Then we get the local and global features before the last layer of decoder, and fuse them together with the global feature of the previous frame from the memory bank as a reference of weights in DWFM. The final prediction will be obtained from the fused feature of DWFM. The details of each module are described in the following subsections.

2.2 Dual-branch Residual Prior Encoder

Visual auto-regressive modeling (VAR) [21] is renowned for its scalable auto-regressive generation capability. Its multi-scale unified quantization enables consistent image representation across different scales, effectively capturing both

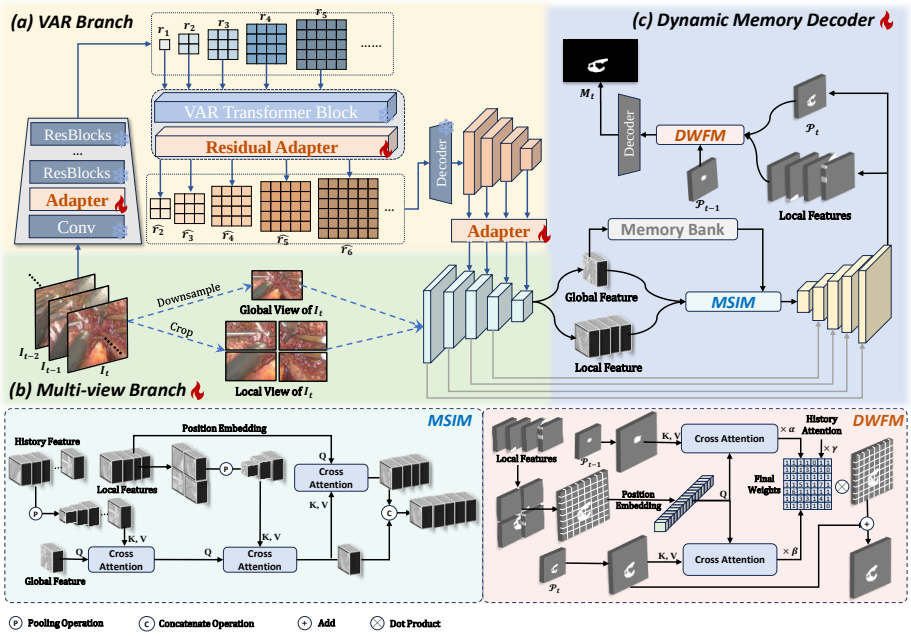

Fig. 2. Pipeline of the proposed HRVVS. Our method comprises three main components: (a) "VAR Branch", a multi-scale generation branch based on visual autoregressive modeling equipped with adapters. (b) "Multi-view Branch" is based on a hierarchical encoder with five different views of the current frame. (c) the "Dynamic Memory Decoder" is a decoder of our network, which includes a multi-scale decoder, a memory bank, a Multi-view Spatiotemporal Interaction Module (MSIM), and a Dynamic Weights Fusion Module (DWFM). Below the pipeline we show the detailed structure of the MSIM and DWFM.

global context and fine-grained details. By extracting hierarchical features from the VAR branch and incorporating them as residual priors into the downsampling layers of the multi-view branch, VAR enhances the information flow within the multi-view branch, improving feature representation.

Given a specific frame $I_t \in \mathbb{R}^{3\times H\times W}$ in an HR video $\mathcal{V} = \{I_i \mid i = 1, 2, \cdots, n\}$, we extract its hierarchical features separately through the VAR branch and multi-view branch, and add the multi-scale features from VAR branch as residual priors to the downsampling layers of multi-view branch. In the multi-view branch, we process high-resolution images by performing center quartering operations and downsampling on I_t respectively to obtain the local view $\{L_m\}_{m=1}^4 \in \mathbb{R}^{3\times h\times w}$ and global view $G \in \mathbb{R}^{3\times h\times w}$ of I_t, respectively, where $(H, W) = (2h, 2w)$. Then we get the features from the encoder $\{\mathcal{F}_m^i\}_{m=1}^5 = \{\{\mathcal{L}_m^i\}_{m=1}^4, \mathcal{G}^i\}$, where $i \in [1, 5]$ represents features extracted from the i-th layer of encoder. In the VAR branch, by using the pre-trained VQ-VAE and VAR weights, we freeze the backbone network, and finetune adapters incor-

porated between the VAE encoder and the VAR transformer blocks. As shown in Eq. 1, $\mathcal{A}$ and $\mathcal{T}$ represent the VAR adapter and VAR transformer block respectively, respectively, while $\theta_{\mathcal{A}}$ represents the learnable parameters in $\mathcal{A}$:

$$r_k = \mathcal{A}\left(\mathcal{T}([s], r_1, r_2, \cdots, r_{k-1}); \theta_{\mathcal{A}}\right). \tag{1}$$

Another adapter is adopted to project VAR features into multi-view latent space. We denote the fused features as $\mathcal{F'}^i \in \{\mathcal{L'}^i, \mathcal{G'}^i\}$ and store $\mathcal{G'}^i$ in the memory bank as the current frame global feature.

2.3 Multi-view Spatiotemporal Interaction Module

To update the multi-view features and the global features from the memory bank, we have designed a Multi-view Spatiotemporal Interaction Module (MSIM). Inspired by [13, 23], we update the local and global features by a multi-head cross-attention (MHCA). Before the multi-view interaction process, we introduce the multi-scale memory features as the reference to update the global features. As shown in Fig 2, the MSIM module effectively combines $\{\mathcal{L'}^5_m\}^4_{m=1}$ and $\mathcal{G'}^5$ using a MHCA mechanism. During the forward propagation, $\{\mathcal{L'}^5_m\}^4_{m=1}$ are first rearranged and position-encoded. When the number of historical frames reaches the upper bound, the historical frames are downsampled and position-encoded to generate multi-scale memory concatenated tokens $\mathcal{H}_n$. MHCA is then employed to compute the attention for the historical frames, thereby updating the global features $\mathcal{G}^5$ as Eq. 2, where Q are the tokens from $\mathcal{G'}^5$, and K and V are $\mathcal{H}_n$:

$$\mathcal{G}_h = \mathcal{G'}^5 + \text{Dropout}\left(\text{MHCA}\left(Q, K, V\right)\right), \tag{2}$$

Next, the local features of the current frame are pooled to generate features at different scales. We also use MHCA to further update the global feature $\mathcal{G}_h$ and get $\mathcal{G}_{msim}$, which is updated by $\{\mathcal{L'}^5_m\}^4_{m=1}$. In another branch, we concatenate $\{\mathcal{L'}^5_m\}^4_{m=1}$ with position encoding p_{poses} and calculate MHCA on the $\mathcal{G}_{msim}$ to obtain the updated local feature (equation 3). The updated local features $\{\mathcal{L}_{mism}\}^4_{m=1}$ are then rearranged and concatenated with the updated global features to be the input of the decoder:

$$\{\mathcal{L}_{mism}\}^4_{m=1} = \{\mathcal{L}^5_m\}^4_{m=1} + \text{Dropout}\left(\text{MHCA}\left(Q_m, K + p_{\text{poses}}, V\right)\right). \tag{3}$$

2.4 Dynamic Weights Fusion Module

Inspired by [14, 22, 26], we proposed DWFM. Specifically, we further divide the 4 local features into 4×16 small patches, and assign corresponding weights to each patch, to reduce the boundary fragmentation caused by local attempts to directly aggregate. We take each local patch as Q and calculate MHCA with the current global feature and the last global feature separately, obtaining the corresponding importance weights $\mathcal{W}^i_{final}$.

Excessive focus on the global features of the current frame can lead to a loss of optical flow connection in adjacent frames. We compute Weights A and

Weights B from the current global feature $\mathcal{P}_t$ and the previous global feature $\mathcal{P}_{t-1}$, labeling them as $\mathcal{W}_g^t$ and $\mathcal{W}_l^t$ for the t-th frame, respectively. For the first frame in each video, only its global weight $\mathcal{W}_g^0$ is calculated, and it is stored as the historical weight $\mathcal{W}_h^1$ for the next frame's prediction. For each subsequent frame's patches, the weight score of each patch is continuously updated based on the historical weight $\mathcal{W}_h^t$, the current global feature weight $\mathcal{W}_g^t$, and the previous frame's weight $\mathcal{W}_l^t$. The expressions for updating historical weights and calculating current weights are shown in equations 4 and 5, respectively.

The prediction result of the previous frame should be given attention because there is often a clear optical-flow connection between adjacent frames, and excessive attention to the global features of the current frame can cause the segmentation result to lose this connection. The Weights A and Weights B are calculated from the current global feature $\mathcal{P}_t$ and the last global feature $\mathcal{P}_{t-1}$, and marked as $\mathcal{W}_g^t$ and $\mathcal{W}_l^t$ for t-th frame respectively. For the patches of the first frame in each video, we only calculate its global weights $\mathcal{W}_g^0$ and save it as the history weights $\mathcal{W}_h^1$ of the next frame prediction. For patches of each following frame, we continuously update the weight score of each patch based on the historical weight score $\mathcal{W}_h^t$, the current weight score of global feature $\mathcal{W}_g^t$, and the weight score of the last frame $\mathcal{W}_l^t$. The historical weights update and current weights calculation expressions are respectively shown in Eq. 4 and 5

$$\mathcal{W}_{final}^t = \begin{cases} \mathcal{W}_g^0, & t = 0 \\ \alpha \times \mathcal{W}_l^t + \beta \times \mathcal{W}_g^t + \gamma \times \mathcal{W}_h^t, & t > 0 \end{cases}, \tag{4}$$

$$\mathcal{W}_h^{t+1} = \delta \times \mathcal{W}_h^t + (1 - \delta) \times \mathcal{W}_{final}^t. \tag{5}$$

3 Experiments

3.1 Dataset and Experimental Settings

Hepa-SEG Dataset. We introduce **Hepa-SEG**, the first vasculature segmentation dataset for hepatectomy. The dataset consists of 35 hepatectomy videos, totaling 11,442 frames with a resolution of 1080×1920. Each video contains approximately 8 minutes of continuous frames from the liver transection stage, where every frame is manually annotated. The dataset includes two vasculature types: the Glisson sheath and the hepatic vein. The data is randomly split into training, validation, and test sets with a ratio of 7:1:2.

Implementation Details. All experiments are conducted on a single NVIDIA A800 GPU. Our model is trained for 15 epochs with a batch size of 32. A sliding window sampler is used to ensure that each batch contains consecutive frames. We optimize the model using Adam with an initial learning rate of 1×10^{-5}, which is decayed using a polynomial scheduler with a decay rate of 0.9.

Table 1. Quantitative Comparison with Different Methods on Hepa-SEG. The best values are highlighted in bold. ↑ denotes that a higher score is better.

Methods	Venue	Type	Jaccard ↑	Dice ↑	S_α ↑	F_β^ω ↑	E_ϕ^{mn} ↑
PraNet [7]	$MICCAI_{20}$	image	0.3569	0.4586	0.6875	0.5124	0.8135
LDNet [30]	$MICCAI_{22}$	image	0.2322	0.2929	**0.8355**	0.2798	0.8331
ISNet [18]	$ECCV_{22}$	image	0.1982	0.2576	0.7854	0.2710	0.8103
HitNet [11]	$AAAI_{23}$	image	0.4481	0.5700	0.4851	0.5434	0.8276
SLT-Net [3]	$CVPR_{22}$	video	0.2825	0.4904	0.6521	0.4097	0.6729
Vivim [28]	$TCSVT_{24}$	video	0.4480	0.5801	0.7511	0.5801	0.8380
Med-SAM2 [31]	$Arxiv_{24}$	video	0.3470	0.4555	0.6728	0.4552	0.5268
SALI [10]	$MICCAI_{24}$	video	0.5239	0.6424	0.7748	0.6496	0.8405
MemSAM [4]	$CVPR_{24}$	video	0.1337	0.2126	0.4642	0.2369	0.4683
HRVVS(Ours)	– –	video	**0.5405**	**0.6532**	0.7878	**0.6769**	**0.8711**

3.2 Comparisons with State-of-the-Arts

Baselines and Metrics. We evaluate HRVVS against nine state-of-the-art segmentation methods on the **Hepa-SEG dataset**, including four image-level and five video-level approaches. Specifically, the baselines comprise two high-resolution segmentation methods (i.e., HitNet [11] and ISNet [29]), six medical image segmentation methods (i.e., PraNet [7], LDNet [30], Vivim [28], Med-SAM2 [31], SALI [10], and MemSAM [4]), and one general segmentation method (i.e., SLT-Net [8]). For quantitative evaluation, we adopt five commonly used metrics [17]: Jaccard index, Dice coefficient, Structure-measure ($\mathcal{S}_\alpha$) [5], F-measure (F_β^ω) [1], and Enhanced-alignment measure (E_ϕ^{mn}) [6].

Quantitative Comparison. As shown in Table 1, our proposed HRVVS achieves state-of-the-art performance on the Hepa-SEG dataset, outperforming all baselines across most metrics. Specifically, compared to the best-performing baseline, HRVVS achieves a relative improvement of +3.16% in Jaccard index, +1.68% in Dice coefficient, +4.20% in F-measure, and +3.60% in Enhanced-alignment measure. The only exception is the S-measure, where LDNet achieves a slightly higher score (0.8355 vs. 0.7878). However, LDNet exhibits a significantly lower Dice coefficient (0.2929), indicating that while it maintains high local consistency, it struggles to segment the complete target region (see Fig. 3). Additionally, methods such as MemSAM and LDNet, which are optimized for ultrasound image segmentation, perform poorly on Hepa-SEG. This highlights the unique challenges posed by our dataset, where both spatial continuity and fine-grained vessel structures are critical for accurate segmentation.

Qualitative Comparison. In Fig. 3, we visualize the segmentation results of HRVVS alongside state-of-the-art methods on Hepa-SEG. HRVVS effectively captures fine details of hepatic vasculature, demonstrating superior segmentation accuracy and robustness in complex surgical scenes.

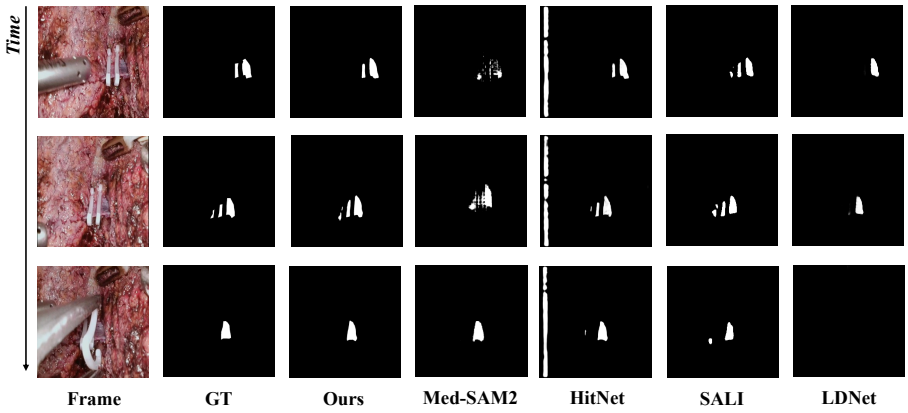

Fig. 3. Visualization results of different methods on a challenge clip.

3.3 Ablation Experiments

We conduct the ablation study to evaluate the effectiveness of three main modules (i.e., the VAR branch, the MSIM module, and the DWFM module), and report the results in Tab 2.

Table 2. Ablation study on Hepa-SEG dataset. "VAR" denotes the VAR branch, MSIM and DWFM are two modules introduced above.

Design	VAR	MSIM	DWFM	Jaccard ↑	Dice ↑	S_α ↑	F_β^ω ↑	E_ϕ^{mn} ↑
basic	-	-	-	0.4938	0.6122	0.7515	0.6311	0.8189
M1	✓	✓	-	0.4994	0.6233	0.7603	0.6307	0.8222
M2	✓	-	✓	0.5332	0.6442	0.7771	0.6712	0.8615
M3	-	✓	✓	0.5242	0.6384	0.7757	0.6613	0.8619
Ours	✓	✓	✓	**0.5405**	**0.6532**	**0.7878**	**0.6769**	**0.8711**

In this ablation study, we evaluate the impact of key components in our model on the Hepa-SEG dataset, specifically the VAR branch, Multi-scale Integration Module (MSIM), and Dynamic Weighted Feature Module (DWFM).

The baseline model, which excludes all three components, achieves a Jaccard index of 0.4938, a Dice coefficient of 0.6122, an S_α score of 0.7515, an F_β^ω score of 0.6311, and an E_ϕ^{mn} score of 0.8189. Adding only the VAR branch and MSIM (Model M1) slightly improves performance (Jaccard: 0.4994, Dice: 0.6233), suggesting their individual contributions are modest.

Incorporating VAR with DWFM (Model M2) leads to more substantial improvements (Jaccard: 0.5332, Dice: 0.6442), emphasizing DWFM's effectiveness in feature refinement. Similarly, using MSIM and DWFM together (Model M3)

enhances performance, though slightly less than M2. Finally, integrating all three components in the full model achieves the highest performance (Jaccard: 0.5405, Dice: 0.6532), demonstrating their complementary roles in improving segmentation accuracy.

4 Conclusion

This paper presents the first hepatic vasculature segmentation dataset under surgical video scenes, and a matching method based on hierarchical autoregressive residual priors. To address challenges in high-resolution surgical hepatectomy videos, our method proposes a VAR branch and a dynamic memory mechanism to embed them into a multi-view segmentation network. Experiments demonstrate that our HRVVS is capable of state-of-the-art results on Hepa-SEG and can be a critical baseline for video vasculature segmentation.

Acknowledgments. This work is supported by the Guangdong Science and Technology Department (No. 2024ZDZX2004) and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007).

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

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1597–1604. IEEE (2009)
2. Ali, S., Espinel, Y., Jin, Y., Liu, P., Güttner, B., Zhang, X., Zhang, L., Dowrick, T., Clarkson, M.J., Xiao, S., et al.: An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion from the miccai2022 challenge. Medical image analysis **99**, 103371 (2025)
3. Cheng, X., Xiong, H., Fan, D.P., Zhong, Y., Harandi, M., Drummond, T., Ge, Z.: Implicit motion handling for video camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13864–13873 (2022)
4. 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)
5. 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)
6. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. Scientia Sinica Informationis **6**(6), 5 (2021)
7. 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)

8. Feng, K., Ren, L., Wang, G., Wang, H., Li, Y.: Slt-net: A codec network for skin lesion segmentation. Computers in Biology and Medicine **148**, 105942 (2022)
9. Guo, X., Xiao, R., Zhang, T., Chen, C., Wang, J., Wang, Z.: A novel method to model hepatic vascular network using vessel segmentation, thinning, and completion. Medical & biological engineering & computing **58**, 709–724 (2020)
10. Hu, Q., Yi, Z., Zhou, Y., Peng, F., Liu, M., Li, Q., Wang, Z.: Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–541. Springer (2024)
11. Hu, X., Wang, S., Qin, X., Dai, H., Ren, W., Luo, D., Tai, Y., Shao, L.: High-resolution iterative feedback network for camouflaged object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 881–889 (2023)
12. 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)
13. Liang, H., Wen, G., Hu, Y., Luo, M., Yang, P., Xu, Y.: Mvanet: Multi-task guided multi-view attention network for chinese food recognition. IEEE Transactions on Multimedia **23**, 3551–3561 (2020)
14. Liang, Y., Li, X., Chen, X., Chen, H., Zheng, Y., Lai, C., Li, B., Xue, X.: Global semantic-guided sub-image feature weight allocation in high-resolution large vision-language models. arXiv preprint arXiv:2501.14276 (2025)
15. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. arXiv preprint arXiv:2408.07931 (2024)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. 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)
18. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Van Gool, L.: Highly accurate dichotomous image segmentation. In: European Conference on Computer Vision. pp. 38–56. Springer (2022)
19. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
20. Smyrniotis, V., Farantos, C., Kostopanagiotou, G., Arkadopoulos, N.: Vascular control during hepatectomy: review of methods and results. World journal of surgery **29**, 1384–1396 (2005)
21. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. arXiv preprint arXiv:2404.02905 (2024)
22. Yang, Y., Aviles-Rivero, A.I., Fu, H., Liu, Y., Wang, W., Zhu, L.: Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13200–13210 (2023)

23. Yang, Y., Fu, H., Aviles-Rivero, A.I., Xing, Z., Zhu, L.: Diffmic-v2: Medical image classification via improved diffusion network. IEEE Transactions on Medical Imaging (2025)
24. Yang, Y., Wang, S., Liu, L., Hickman, S., Gilbert, F.J., Schönlieb, C.B., Aviles-Rivero, A.I.: Mammodg: Generalisable deep learning breaks the limits of cross-domain multi-center breast cancer screening. arXiv preprint arXiv:2308.01057 (2023)
25. Yang, Y., Wang, S., Zhu, L., Yu, L.: Hcdg: A hierarchical consistency framework for domain generalization on medical image segmentation. arXiv preprint arXiv:2109.05742 (2021)
26. Yang, Y., Wu, H., Aviles-Rivero, A.I., Zhang, Y., Qin, J., Zhu, L.: Genuine knowledge from practice: Diffusion test-time adaptation for video adverse weather removal. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25606–25616. IEEE (2024)
27. 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)
28. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)
29. Zhang, M., Zhang, R., Yang, Y., Bai, H., Zhang, J., Guo, J.: Isnet: Shape matters for infrared small target detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 877–886 (2022)
30. 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)
31. Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)