Spatiotemporal-Sensitive Network for Microvascular Obstruction Segmentation from Cine Cardiac Magnetic Resonance

Yang Yu[1], Christopher Kok[1], Jiahao Wang[2], Jun Cheng[1], Shuang Leng[3], Ru San Tan[3], Liang Zhong[3,4], and Xulei Yang[1]

[1] Institute for Infocomm Research (I^2R), A*STAR, Singapore
[2] Mechanobiology Institute (MBI), National University of Singapore, Singapore
[3] National Heart Centre Singapore (NHCS), Singapore
[4] Duke-NUS Medical School, Department of Biomedical Engineering, National University of Singapore (NUS), Singapore
zhong.liang@duke-nus.edu.sg, yang_xulei@i2r.a-star.edu.sg

Abstract. Accurate diagnosis of microvascular obstruction (MVO) in acute myocardial infarction (AMI) patients typically relies on Cine Cardiac Magnetic Resonance Imaging (CMR) (video sequences) and Late Gadolinium Enhancement (LGE) CMR (images). However, LGE imaging is contraindicated in approximately 20% of AMI patients with chronic kidney disease, underscoring the need for Cine CMR as a standalone diagnostic alternative. Although recent advancements in deep learning have improved video data processing, current methods fail to adequately capture complementary temporal motion features. This limits their efficacy and poses significant challenges for MVO segmentation with Cine CMR, as MVO regions are defined by dynamic motion rather than clear boundaries or contrast on Cine CMR. To address this limitation, we propose a Spatiotemporal-Sensitive Network that integrates static and motion encoders to effectively process Cine CMR. Further through a guided decoder utilizing the rich spatiotemporal information and an uncertainty-driven refinement leveraging uncertainty maps and low-level features, our method enhances segmentation accuracy and refines boundary delineation. Extensive experiments on 621 Cine CMR demonstrate superior performance over competing methods with a Dice score of 0.56 in Cine CMR-based MVO identification and highlight its potential to advance video analysis in clinical settings. The code is available at https://github.com/MICCAI25-MVO-Segmentation/miccai25-mvo-seg.

Keywords: Spatiotemporal Analysis · Video Segmentation · Cardiac Magnetic Resonance Imaging · Diagnostic Radiographs.

1 Introduction

Ischemic heart disease remains one of the leading causes of mortality worldwide, with microvascular obstruction (MVO) in acute myocardial infarction (AMI)

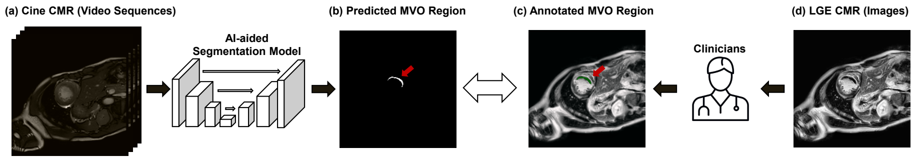

Fig. 1. Demonstrated diagnostic procedures through Cine CMR and LGE CMR. AI-assisted models are designed to predict MVO segmentation from Cine CMR, closely aligning with clinician-annotated regions using LGE CMR.

contributing significantly to global death rates [10]. Accurate identification and diagnosis of MVO typically depend on Cine Cardiac Magnetic Resonance Imaging (CMR) (video sequences) and Late Gadolinium Enhancement (LGE) CMR (images) [1]. Cine CMR provides dynamic imaging of the cardiac cycle and often serves as the first diagnostic step, followed by LGE CMR, which enhances the accurate visualization of myocardial tissue characteristics [11]. However, LGE imaging is contraindicated in approximately 20% of AMI patients with chronic kidney disease due to the risks associated with gadolinium-based contrast agents (CAs) [13]. This limitation underscores the need to explore Cine CMR, a contrast-free imaging technique that captures myocardial motion across multiple frames, as a standalone diagnostic tool for assessing myocardial damage [24] (Fig. 1).

Recent advancements in video segmentation deep learning methods have yielded substantial improvements. AFB-URR adopts a feature-matching approach by encoding object masks from previous frames using an adaptive feature bank, complemented by a region refinement mechanism [15]. DCFNet targets video salient object detection by generating dynamic convolution kernels capable of extracting temporal context features across frames [23]. DPSTT proposed a dynamic parallel spatiotemporal Transformer with an efficient dynamic memory selection [14]. PNS+ processes the initial anchor frame and subsequent frames within a sliding window using separate encoders, followed by normalized self-attention over the embeddings [9]. FLA-Net addresses video segmentation through a frequency-based feature aggregation module to capture temporal relations across spatial features [16]. Vivim leverages a spatiotemporal Mamba [5] encoder to construct a medical video segmentation model [21].

Despite the remarkable performance of these state-of-the-art (SOTA) techniques, they predominantly rely on clear spatial image features. This reliance is inadequate for segmenting MVO regions in Cine CMR, where the segmentation targets lack distinct regions of contrast or clear boundaries. Unlike LGE CMR, Cine CMR provides motion features that are critical for capturing dynamic phenomena such as MVO. This underscores a significant gap in existing approaches and highlights the need for methods that effectively incorporate temporal motion features for accurate segmentation of such regions. While Cine CMR has demonstrated significant advancements in identifying certain myocardial lesions, its potential for detecting MVO remains largely untapped [4, 17].

To this end, we introduce a novel framework incorporating both static and motion encoders to achieve non-contrast MVO segmentation on Cine CMR. The static encoder captures structural details of the myocardium by focusing on local features, while the motion encoder extracts complementary motion dynamics by analyzing changes between consecutive frames. These spatial and temporal features are seamlessly fused within a guided decoder, enabling the model to effectively leverage the rich spatiotemporal information inherent in Cine CMR. Building on prior research that utilizes uncertainty maps for fine-grained segmentation [8,12,15,20], an uncertainty-driven refinement is also adopted for further optimization on the boundary regions. This framework addresses existing challenges by dynamically integrating the complex relationships between spatial and temporal information, thereby advancing the comprehensive understanding of medical video data. Our contributions can be summarized as follows:

1. We propose an innovative framework designed to learn fused representations of spatiotemporal information videos through a guided decoder that effectively segments static and motion-related features.
2. We incorporate a local refinement mechanism leveraging uncertainty maps and low-level feature maps to enhance segmentation accuracy.
3. We perform a series of experiments to showcase the performance enhancements over SOTA methods. We highlight the framework's adaptability to practical scenarios involving varying numbers of frames in video inputs.

2 Methods

2.1 Problem Formulation

Our proposed method illustrated in Fig. 2 adopts a CA-free approach to identifying MVO regions in Cine CMR of AMI patients, and to widen the approach to patients who may have contraindications for LGE CMR procedures but may be able to undergo Cine CMR procedures. We consider a video sequence of Cine CMR $X \in \mathbb{R}^{H \times W \times D}$ and its corresponding segmentation mask on MVO $Y \in \mathbb{R}^{H \times W \times K}$ acquired from paired LGE CMR image, where H, W, D, K are the height, width, number of frames, and number of classes, respectively. We also compute the residual sequence from this video sequence denoted as $\tilde{X}$ where $\tilde{X}_i = X_i - X_{(i+1)}$. The last residual frame takes the difference from the last and first video frames to embed the cyclic relation of the Cine CMR.

2.2 Static and Motion Feature Extraction

For the static encoder on video frames, we employ a ResNet50 [7] backbone to compute various spatial feature maps $F^{\{l\}} \in \mathbb{R}^{H/2^l \times W/2^l \times D \times 2^{l-1}C}$ including scenes and objects depicted in the video from input frames X, where l is the layer index and C is the number of channels in the first feature map. For the motion encoder on residual frames, we also employ a ResNet50 [7] backbone to compute various temporal feature maps $\tilde{F}^{\{l\}}$ for strengthening the movement of the objects from residual frames $\tilde{X}$ such that $\tilde{F}^{\{l\}}$ is of the same shape as $F^{\{l\}}$.

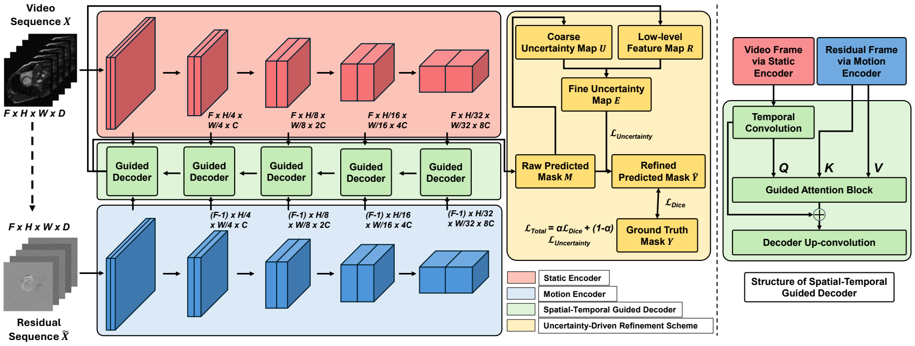

Fig. 2. The workflow of proposed Spatiotemporal-Sensitive Network. The static encoder captures the intricate structural details by emphasizing local features and the motion encoder extracts dynamic motion patterns by analyzing changes between consecutive frames. These features are harmoniously integrated within a guided decoder to effectively harness the rich spatiotemporal information embedded in Cine CMR. Additionally, uncertainty estimation is employed to refine the segmentation along the boundary regions, ensuring improved precision and reliability.

2.3 Guided Decoder for Spatiotemporal Analysis

Unlike the conventional two-stream architectures which process spatiotemporal information through late fusion [3, 19, 22], we propose a novel guided decoder to refine feature extraction along the temporal dimension of video frames while optimizing these features using motion information derived from residual frames. The continuous movements of the target object within a confined spatial region are effectively enhanced, with minimized interference caused by redundant temporal features. To achieve this, we first convolve the embedding F from the static encoder over its temporal dimension to produce a spatiotemporal embedding $\hat{F}^{\{l\}} \in \mathbb{R}^{H/2^l \times W/2^l \times 2^{l-1}C}$ where l is the layer index. Further, we compute cross-attention between the spatiotemporal embeddings $\hat{F}$ and each residual embedding $\tilde{F}_i$ to produce a final embedding $Z^{\{l\}} \in \mathbb{R}^{H/2^l \times W/2^l \times 2^{l-1}C}$:

$$Z^{\{l\}} = \sum_{i=1}^{D} \mathrm{GuidedAttentionBlock}\left(\hat{F}^{\{l\}}, \tilde{F}_i^{\{l\}}, \tilde{F}_i^{\{l\}}\right) \tag{1}$$

where $\hat{F}$ is the query tensor, and $\tilde{F}$ is both the key and value tensors. Lastly, we fuse spatiotemporal $\hat{F}$ and cross-attention embeddings Z by adding them to compute the final embeddings $\hat{Z}$ for leveraging complementary information from both encoders:

$$\hat{Z} = \hat{F} + Z \tag{2}$$

These final embeddings are propagated up through the deconvolutional layers of the decoder to produce an initial segmentation mask $M \in \mathbb{R}^{H \times W \times K}$.

2.4 Uncertainty-Driven Refinement

We further incorporate a confidence loss to quantify the ambiguity in segmentation results and a refinement module to tackle these ambiguous regions. This design is particularly suited to the characteristics of Cine CMR, where the lack of distinct contrast regions and well-defined boundaries presents unique challenges. From the initial segmentation mask M, we generate a coarse uncertainty map U given the largest two likelihood values $\hat{M}^1$ and $\hat{M}^2$ for every pixel:

$$U = \exp\left(1 - \frac{\hat{M}^1}{\hat{M}^2}\right) \tag{3}$$

where $\frac{\hat{M}^1}{\hat{M}^2} \in [1, +\infty)$ and $U \in (0, 1]$. Next, we take the weighted average of the local spatiotemporal feature $\hat{F}^{\{1\}}$ from the first layer of the guided decoder that contains both spatiotemporal and cross-attention information to compute the reference feature R_i with the use of AvgPool and MaxPool operations:

$$R_i = \frac{\text{AvgPool}\left(M_i F^{\{1\}}\right)}{\text{AvgPool}\left(M_i\right)} \tag{4}$$

Embeddings are then obtained from a residual network module f_l which learns to predict local similarity, computing a local refinement mask E by comparing the similarity between $\hat{F}^{\{1\}}$ and R_i:

$$E_i = \text{MaxPool}\left(M_i\right) f_l\left(\hat{F}^{\{1\}}, R_i\right) \tag{5}$$

We compute the final segmentation mask $\hat{Y}$ by adding the local refinement mask E weighted by the uncertainty estimate U of the initial segmentation mask M and take the softmax of the output over the class dimension.

$$\hat{Y}(p) = \text{softmax}\left(M(p) + U(p)E(p)\right) \tag{6}$$

2.5 Loss Function

To facilitate the training of the model, we employ a loss function that combines Dice loss L_{dice} and a confidence loss $L_u(\hat{U}) = \|\hat{U}\|_2$ computed from $\hat{U}$—an uncertainty map taken from the final segmentation mask $\hat{Y}$ (see Eq. 3):

$$L_{\text{overall}} = \alpha L_{\text{dice}}(Y, \hat{Y}) + (1 - \alpha) L_u(\hat{U}) \tag{7}$$

where $\hat{Y}$ denotes the predicted logits for MVO presence and Y indicates the ground truth mask. Empirical tests revealed that setting $\alpha = 0.95$ yielded the best performance. During the inference phase, the model is provided with regular and residual frames X' and $\tilde{X}'$ from their respective datasets to produce a segmentation mask $\hat{Y}'$, which will be used to evaluate the model's performance.

3 Experiments

3.1 Datasets

Short-axis Cine and LGE CMR data were acquired using a 3T Siemens scanner, resulting in a dataset comprising 621 paired scans from 125 cases. All images encompassed the entire left ventricle. Each scan included a 30-frame sequence of Cine CMR capturing the full cardiac cycle and a corresponding image of LGE CMR obtained during the diastolic phase. This study was approved by SingHealth Centralised Institutional Review Board. Expert manual annotation was conducted where both the endocardium and epicardium were delineated after rigid registration of the Cine CMR and LGE CMR. Annotations were performed by a specialist with over 10 years of experience. MVOs were manually segmented from the myocardium on the LGE CMR, serving as the raw masks. Data partitioning follows a subject-wise stratification strategy, ensuring scans from the same subject are never included in both training and test sets.

3.2 Experiment Settings

Based on the filenames of the subjects in the original dataset, we divided the data into training, validation, and testing sets in a uniform ratio of 70%, 10%, and 20%, respectively. This resulted in a total of 443/63/115 scans distributed across the three sets. All images were cropped to a size of 224x224, and data augmentation was performed using random Flip/Rotation/Transform to enhance the variability of the dataset. We selected 10 evenly spaced frames from Cine CMR to serve as the primary focus of our analysis. For model training and inference, we utilized PyTorch Lightning as the deep learning framework. We adopt a U-Net [18] architecture with a modified ResNet-50 [7] backbone encoder. All experiments were conducted on a single NVIDIA A5000 GPU (24 GB of VRAM) with a training time of 3 hours. Key hyperparameters included the Adam optimizer, an initial learning rate of 1e-4, beta values set to [0.5, 0.99], a weight decay of 3e-4, and a cosine annealing scheduler for learning rate adjustment. The batch size was fixed at 8, and each experiment was run for 3200 iterations. The performance in the segmentation task was evaluated using several metrics: Dice coefficient, Jaccard index, Hausdorff Distance, pixel-based precision, and recall. Statistical significance was evaluated using Wilcoxon Signed-Rank Test.

3.3 Comparison Study

We evaluated the performance of our proposed method by comparing it with 7 SOTA methods, encompassing both image-based and video-based approaches. The image-based methods included UNet [18], UNet++ [25], TransUNet [2] and SwinUnetR [6], while the video-based methods comprised AFB-URR [15], DP-STT [14], PNS+ [9], and Vivim [21]. To ensure accurate and meaningful comparisons, the segmentation results of all SOTA methods were obtained using their

Table 1. Results on SOTA Comparison for the MVO segmentation task.

Methods and Tasks		MVO Segmentation				
SOTA methods	Image/Video	Dice ↑	Jaccard ↑	HSD ↓	Precison ↑	Recall ↑
UNet [18]	Image	0.4742	0.4603	10.4260	0.6351	0.6290
UNet++ [25]	Image	0.4893	0.4709	7.5933	0.7355	0.6290
TransUNet [2]	Image	0.4942	0.4878	9.9162	0.8145	0.5829
SwinUnetR [6]	Image	0.4481	0.4300	7.5491	0.6572	0.6541
AFB-URR [15]	Video	0.4372	0.4361	39.3599	0.7158	0.5872
DPSTT [14]	Video	0.5026	0.4895	8.7484	0.7655	0.5726
PNS+ [9]	Video	0.4957	0.4957	144.9119	0.8609	0.5565
Vivim [21]	Video	0.5264	0.5099	8.5704	0.6573	**0.6397**
Our Proposed Method	Video	**0.5556**	**0.5440**	**7.2389**	**0.8710**	0.5951

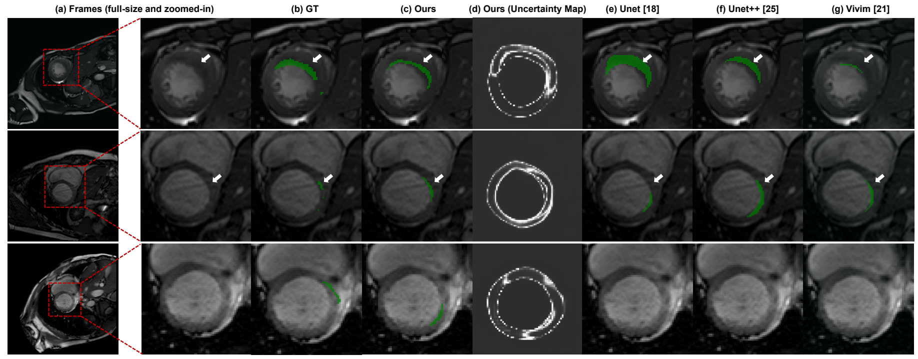
(a) Frames (full-size and zoomed-in)
(b) GT
(c) Ours
(d) Ours (Uncertainty Map)
(e) Unet [18]
(f) Unet++ [25]
(g) Vivim [21]

Fig. 3. Visual comparisons of MVO segmentation results produced by our framework and SOTA methods. "GT" denotes the ground truth.

publicly available implementations or our implementation when the code was not available.

Table 1 presents the quantitative results of our proposed network alongside those of the compared methods for MVO segmentation. Results reveal that most video-based methods consistently outperform image-based methods. We mainly focused on the Dice, Jaccard, and HSD values since they provide a clear and intuitive way to assess if the boundaries of MVO is being accurately identified. Among the evaluated approaches, Vivim [21] achieves the highest scores. Notably, our proposed method surpasses Vivim [21] in terms of Dice, Jaccard and HSD, exhibiting superior overall performance. Specifically, our method improves the Dice score from 0.5264 to 0.5556, the Jaccard score from 0.5099 to 0.5440, the HSD score from 8.5704 to 7.2389. The resulting p-values (0.006 for Dice, 0.009 for Jaccard, 0.001 for HSD) from the Wilcoxon Signed-Rank Test also indicate a statistically significant improvement over the best baseline vivim. Additionally, Fig. 3 provides a visual comparison of MVO segmentation outcomes

Table 2. Results on Ablation Study for the MVO segmentation task.

Methods and Tasks				MVO Segmentation				
Spatial Encoder	Motion Encoder	Guided Decoder	Uncertainty Refinement	Dice ↑	Jaccard ↑	HSD ↓	Precison ↑	Recall ↑
✓	×	×	×	0.3271	0.3023	10.1626	0.3690	**0.6932**
✓	✓	×	×	0.4311	0.4195	10.9859	0.6707	0.6223
×	✓	✓	×	0.4636	0.4589	26.5120	0.7381	0.5712
✓	✓	✓	×	0.5229	0.5134	**5.4888**	0.8434	0.5906
✓	✓	✓	✓	**0.5556**	**0.5440**	7.2389	**0.8710**	0.5951

Table 3. Results on Extended Study with various number of input frames.

Methods and Tasks	MVO Segmentation				
Number of Frames for Video Inputs	Dice ↑	Jaccard ↑	HSD ↓	Precison ↑	Recall ↑
5	0.5162	0.5116	23.6467	0.8701	0.5784
10	**0.5556**	**0.5440**	7.2389	**0.8710**	0.5951
15	0.4810	0.4708	8.1463	0.7518	0.6149
30	0.5143	0.4957	**6.4272**	0.7236	**0.6225**

produced by our network and other SOTA methods on randomly selected video frames. Remarkably, our method demonstrates the capability to accurately segment MVOs of varying sizes and diverse shapes from input CMR video frames. The refined uncertainty maps are included to demonstrate the effectiveness of the uncertainty scheme. We also include a case where our method encounters a challenge due to the subtle motion differences.

3.4 Ablation Study

To validate the contributions of each component in our approach, we conducted a comprehensive ablation study. The data presented in Table 2 clearly demonstrate that the model's segmentation performance improves when both the static and motion encoders are activated. These findings align with our hypothesis that Cine CMR lacks the contrast-enhancing benefits provided by LGE CMR and, as a result, depends predominantly on motion features for the accurate identification of MVO regions. Furthermore, when the model is combined with the uncertainty modelling scheme, the segmentation performance is further enhanced, as illustrated in Table 2. This highlights the complementary role of uncertainty modelling in optimizing segmentation accuracy.

3.5 Extended Study

To evaluate the clinical applicability of our proposed method in resource-efficient scenarios, we also examined the impact of varying the number (5/15/30) of input frames fed into the model. The corresponding results are presented in Table 3. A decline in performance was observed when the number of video frames was reduced to 5. We hypothesize that this decrease is attributed to the interplay between the number of frames and the magnitude of pixel differences in the residual frames, as well as the limited motion features available when fewer frames are used. Conversely, as the number of frames increases, the pixel magnitude in the residual frames also diminishes, with 15 frames yielding the smallest pixel differences in the residual frames.

4 Conclusion

In this study, we present a novel framework designed to effectively learn and integrate spatial and temporal information, as well as improve segmentation accuracy through a dual-branch encoder, guided decoder, and uncertainty-driven refinement by leveraging video sequences for MVO identification using Cine CMR. Comprehensive experiments conducted on clinical imaging datasets highlight the robustness and efficacy of the proposed framework in enhancing segmentation quality. These findings provide strong evidence for the potential of our method as a contrast-free imaging technique, offering a standalone diagnostic solution for evaluating myocardial damage.

Acknowledgments. This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021) and SingHealth Duke-NUS under its Academic Medicine Research Grant (AM/TP032/2020 (SRDUKAMR2032)).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Beijnink, C.W., van der Hoeven, N.W., Konijnenberg, L.S., Kim, R.J., Bekkers, S.C., Kloner, R.A., Everaars, H., El Messaoudi, S., van Rossum, A.C., van Royen, N., et al.: Cardiac mri to visualize myocardial damage after st-segment elevation myocardial infarction: a review of its histologic validation. Radiology **301**(1), 4–18 (2021)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Chen, J.C., Lee, C.Y., Huang, P.Y., Lin, C.R.: Driver behavior analysis via two-stream deep convolutional neural network. Applied Sciences **10**(6), 1908 (2020)
4. Gonzales, R.A., Lamy, J., Seemann, F., Heiberg, E., Onofrey, J.A., Peters, D.C.: Tvnet: Automated time-resolved tracking of the tricuspid valve plane in mri long-axis cine images with a dual-stage deep learning pipeline. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 567–576. Springer (2021)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Holder, C.J., Shafique, M.: Efficient uncertainty estimation in semantic segmentation via distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3087–3094 (2021)
9. 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)
10. Khan, M.A., Hashim, M.J., Mustafa, H., Baniyas, M.Y., Al Suwaidi, S.K.B.M., AlKatheeri, R., Alblooshi, F.M.K., Almatrooshi, M.E.A.H., Alzaabi, M.E.H., Al Darmaki, R.S., et al.: Global epidemiology of ischemic heart disease: results from the global burden of disease study. Cureus **12**(7) (2020)
11. Kramer, C.M., Barkhausen, J., Bucciarelli-Ducci, C., Flamm, S.D., Kim, R.J., Nagel, E.: Standardized cardiovascular magnetic resonance imaging (cmr) protocols: 2020 update. Journal of Cardiovascular Magnetic Resonance **22**(1), 17 (2020)
12. Krygier, M.C., LaBonte, T., Martinez, C., Norris, C., Sharma, K., Collins, L.N., Mukherjee, P.P., Roberts, S.A.: Quantifying the unknown impact of segmentation uncertainty on image-based simulations. Nature communications **12**(1), 5414 (2021)
13. Ledneva, E., Karie, S., Launay-Vacher, V., Janus, N., Deray, G.: Renal safety of gadolinium-based contrast media in patients with chronic renal insufficiency. Radiology **250**(3), 618–628 (2009)
14. Li, J., Zheng, Q., Li, M., Liu, P., Wang, Q., Sun, L., Zhu, L.: Rethinking breast lesion segmentation in ultrasound: a new video dataset and a baseline network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–400. Springer (2022)
15. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. Advances in Neural Information Processing Systems **33**, 3430–3441 (2020)
16. 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)
17. Meng, Q., Bai, W., Liu, T., O'regan, D.P., Rueckert, D.: Mesh-based 3d motion tracking in cardiac mri using deep learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 248–258. Springer (2022)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
19. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems **27** (2014)
20. Yang, H., Chen, C., Chen, Y., Yip, H.C., QI, D.: Uncertainty estimation for safety-critical scene segmentation via fine-grained reward maximization. Advances in Neural Information Processing Systems **36**, 36238–36249 (2023)

21. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)
22. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702 (2015)
23. Zhang, M., Liu, J., Wang, Y., Piao, Y., Yao, S., Ji, W., Li, J., Lu, H., Luo, Z.: Dynamic context-sensitive filtering network for video salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1553–1563 (2021)
24. Zhang, Q., Burrage, M.K., Shanmuganathan, M., Gonzales, R.A., Lukaschuk, E., Thomas, K.E., Mills, R., Leal Pelado, J., Nikolaidou, C., Popescu, I.A., et al.: Artificial intelligence for contrast-free mri: scar assessment in myocardial infarction using deep learning–based virtual native enhancement. Circulation **146**(20), 1492–1503 (2022)
25. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)