# Supplementary Material for Continuous Spatio-Temporal Memory Networks for 4D Cardiac Cine MRI Segmentation

Meng Ye[1], Bingyu Xin[1], Leon Axel[2], Dimitris Metaxas[1]
[1]Rutgers University, [2]New York University School of Medicine
{my389, bx64, dnm}@cs.rutgers.edu

## 1. Analysis of Computational Complexity

The computational complexity of pixel-level dense query-memory matching in STM [7] or STCN [4] is $\mathcal{O}(TH^2W^2)$. In our patch-level memory matching (PLMM), the main computational complexity consists of two parts: (1) $\mathcal{O}(TN^2)$ for the computation of patch-level affinity scores and (2) $\mathcal{O}(NKH_p^2W_p^2)$ for the computation of pixel-level query-memory matching weights within patches. Therefore, the overall computational complexity of PLMM is $\mathcal{O}(TN^2 + NKH_p^2W_p^2)$. Since $H_p \ll H$, $W_p \ll W$ and $K$ is usually small, the computational complexity of PLMM is greatly reduced, compared with the vanilla dense query-memory matching computational complexity. Furthermore, there is no `softmax` operation involved in the computation of patch-level affinity scores. Hence, our overall computational complexity is further reduced.

## 2. Dataset Details

The 4D cMR dataset used in our work is curated from three public 4D cMR datasets: ACDC [1], MnM [2], and MnM-2 [6]. For ACDC, there are 100 cases in the training set and 50 cases in the testing set. For MnM, there are 175 cases in the training set, 34 cases in the validation set, and 136 cases in the testing set. For MnM-2, there are 160 cases in the training set and 40 cases in the validation set; the testing set is not released publicly. Apart from the validation set of MnM-2, mask annotations on the ED and ES phases of each 4D cMR set in these datasets are presented. We trained and tested CSTM based on these sparsely annotated cMR data. We combined the training sets in the three datasets as our own training set, which gives 435 4D cMR cases. And in the main text, we have reported the testing results on the ACDC testing set (50 cases) and the MnM validation and testing sets (170 cases), respectively.

All the three cMR datasets are of multiple cardiovascular pathologies and healthy volunteers. While ACDC is a single-center dataset, MnM and MnM-2 are both multi-center and multi-vendor datasets. Of particular, in the validation and testing datasets of MnM, there exists an out-of-distribution subdataset (center-5, MRI scanner manufacturer-Canon), which is not presented in the training dataset. Therefore, MnM is a more heterogeneous cMR dataset compared with ACDC.

## 3. Training and Inference Details

Each training sample consists of three frames, either sampled along the temporal axis (ED-ES-ED) or spatially ordered along the $z$-axis with a maximum spatial sampling distance of 5. The online data augmentation exactly follows the strategies used in the main training stage of STCN [4]. Basically, random horizontal flip, random resized cropping (of size 384), color jitter, random grayscale, and random affine transformation were included in the data augmentation. More details could be found in STCN [4].

We took at most two areas of the heart presenting on the first frame as segmentation targets. The key encoder (ResNet-50) and value encoder (ResNet-18) were pre-trained on ImageNet [5]. The batch size was set as 4. Adam optimizer was used with default momentum $\beta_1 = 0.9$, $\beta_2 = 0.999$, a base learning rate of $10^{-5}$, and a $L_2$ weight decay of $10^{-7}$. The training was performed for 300K iterations.

During inference, we resized each input 2D cMR image to ensure the shorter side has a size of 384 pixels. The predicted segmentation mask was then resized to the original cMR image size.

## 4. Training and Inference Details for Baseline Methods

For baseline methods STM [7], HMMN [8], STCN [4], and XMem [3], we follow the same training and inference schemes used for natural scene videos. For each 4D cMR data, we sequentially combined all 2D cMR sequences along the $z$-axis as a single long temporal sample, which was used for training or testing the baseline methods. Note that, each segment of 2D cMR sequence at a specific $z$ position covers a full cardiac cycle. Therefore, the temporal continuity is still preserved in the long temporal samples.

We followed the original hyper-parameter settings to train these baseline models on the cMR dataset, with weights of their key and value encoders (ResNets) initialized from ImageNet [5] pre-training.

During inference, all baseline methods took the first frame (at the middle myocardium wall level) with annotation masks as the memory and propagated the masks bidirectionally towards the basal or apex regions. For subsequent query frames, both STM and HMMN took every fifth frame as a memory frame, and the immediately previous frame as a temporary memory frame. STCN took every fifth query frame as a memory frame. XMem has three memory stores: sensory memory, working memory and long-term memory. The sensory memory was updated every query frame. The working memory was updated every fifth query frame following the First-In-First-Out (FIFO) approach to ensure the total number of memory frames $T_{max} \leq 5$. For most cases, the long-term memory was disabled since these temporal samples were not that long (usually less than 1K frames). When the long-term memory was enabled, we followed the long-term memory generation method proposed in XMem.

## References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 1

[2] Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. 1

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

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[6] Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 2023. 1

[7] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 1

[8] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021. 1