

Reconstruction-free segmentation from undersampled k-space using transformers

Yundi Zhang^{1,2}, Nil Stolt-Ansó^{1,3}, Jiazen Pan^{1,2}, Wenqi Huang^{1,2}, Kerstin Hammernik^{1,4}, and Daniel Rueckert^{1,2,3,4}

¹ School of Computation, Information and Technology, Technical University of Munich, Germany

² School of Medicine, Klinikum Rechts der Isar, Technical University of Munich, Germany

³ Munich Center for Machine Learning, Technical University Munich

⁴ Department of Computing, Imperial College London, UK
yundi.zhang@tum.de

Keywords: Cardiac CINE MRI · Unsupervised learning · Representation learning · K-space measurements

1 Synopsis

Motivation: High acceleration factors place a limit on MRI image reconstruction. This limit is extended to segmentation models when treating these as subsequent independent processes.

Goal: Our goal is to produce segmentations directly from sparse k-space measurements without the need for intermediate image reconstruction.

Approach: We employ a transformer architecture to encode global k-space information into latent features. The produced latent vectors condition queried coordinates during decoding to generate segmentation class probabilities.

Results: The model is able to produce better segmentations across high acceleration factors than image-based segmentation baselines.

Impact: Cardiac segmentation directly from undersampled k-space samples circumvents the need for an intermediate image reconstruction step. This allows the potential to assess myocardial structure and function on higher acceleration factors than methods that rely on images as input.

2 Introduction

In cardiac magnetic resonance (CMR) imaging, an abundance of quantitative clinical metrics (such as ejection fraction, strain, etc.) are derived from segmentation-based modeling of the myocardium. Image reconstruction and segmentation are typically thought of as independent serial processes. In order to reduce acquisition time, k-space data is usually undersampled and reconstruction techniques are employed. These approaches attempt to recover the pixel-level detail

lost during this process. However, accurate segmentation does not strictly benefit from this level of precision, often relying on high level information about the overall content of the image.

While segmentation of cardiac images predominantly takes place on clean images [1], previous works have attempted to tackle higher accelerations by performing segmentation directly on unrefined images [6]. Formulating the task as an end-to-end learning problem has shown further improvements [2].

We hypothesize that the process of magnetic resonance (MR) image reconstruction requires larger amounts k-space samples than what theoretically would be required to extract a segmentation signal from the raw data. Under this assumption, direct segmentation from k-space has the potential to allow the quantification of relevant clinical metrics under higher acceleration factors, while further decreasing acquisition time.

3 Method

In this work, we demonstrate that Transformers [7] are capable of employing global attention to leverage all available k-space measurements to predict accurate segmentation maps. The architecture is able to perform this task directly from a set of sampled k-space points, without a need for zero-filling or interpolating the k-space, and without any form of intermediate reconstruction step. We postulate that multi-headed attention, unlike convolutional approaches, offers the necessary properties to appropriately process the nature of k-space: (1) the mechanism considers global correlations, (2) feature extraction should be insensitive to the relative order in which the same samples are presented, (3) inputs of arbitrary sparsity are supported.

An overview of the architecture is presented in Fig. 1. Our architecture’s encoder extracts features from the sparse input k-space samples into a latent space over the course of 4 layers. In order to efficiently handle hundreds of thousands of k-space samples while avoiding the $\mathcal{O}(N^2)$ memory complexity of naive self-attention in standard transformers, our encoder utilizes alternating cross-attention (CA) and self-attention (SA) blocks as proposed by Perceiver [3]. The CA blocks project global k-space information into a fixed bottleneck of latent vectors, while the SA blocks contextualize features between latent vectors.

The decoder consists of 4 cross-attention blocks, which use the extracted latent information to condition any queried image-domain coordinate into producing segmentation class probabilities. The segmentation output is supervised on Dice and binary cross-entropy losses.

4 Implementation and Results

Our training dataset is comprised of 1200 mid-ventricular slices of cardiac short-axis scans from the UK-Biobank dataset [4]. The dataset is divided into training, validation, and testing sets containing 1000, 100, and 100 samples, respectively.

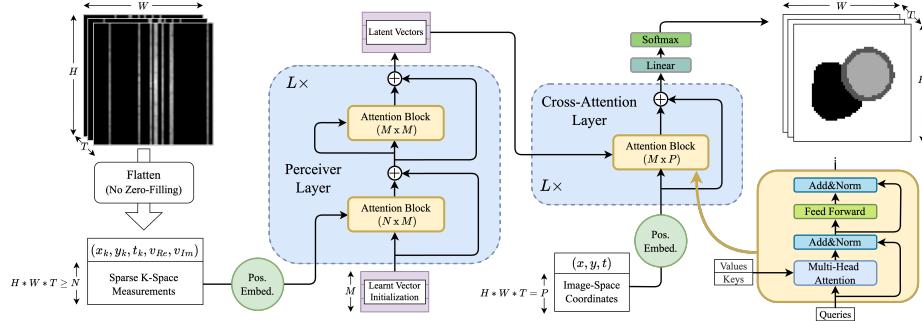


Fig. 1: Overview of DiSK architecture. N , M , and P represent the number of sampled points in k-space, latent vectors, and query points in image domain, respectively. The encoder is made up of L Perceiver layers [3] which alternates cross-attention and latent self-attention blocks. Specifically, cross-attention blocks project global features of the set of input k-space samples K into a fixed-dimensional latent bottleneck of M vectors. Self-attention between latent vectors contextualizes the extracted global features between the vectors. The decoder is made up of L cross-attention layers which condition queried image-domain coordinates with the encoder’s latent vectors into segmentation class probabilities.

Each scan has 50 frames with an average in-plane resolution of approximately 80x80 pixels per frame. We create synthetic undersampled k-space data on-the-fly for each 2D+time scan. Each frame, we apply additional Gaussian B0 variations (in order to remove the conjugate symmetry of k-space) and generate Cartesian undersampling masks by sampling normally distributed lines centered on the DC component. Implementation and training was performed using the Pytorch library on an NVIDIA A40 GPU.

We evaluate the performance of models trained on acceleration factors 4, 8, 16, 32, and 64. Our approach is compared to the performance of two image-based segmentation baselines. Following [6], we implement a model based on the U-Net [5] architecture (Syn-Net) and an autoencoder approach (LI-Net). Their work showed these models to be capable of producing segmentations on noisy reconstructions of undersampled images. As shown in Tab. 1 and Fig. 2, our model obtains higher segmentation Dice scores and lower Hausdorff distances than the proposed baselines across all tested accelerations.

5 Discussion

Due to the nature of short-axis acquisitions, the heart is consistently on the same general location and orientation across the dataset. It is therefore easy for the models to achieve a decent performance at high accelerations by memorizing a general shape and location. Despite this, our model appears to consistently

Table 1: Dice scores and Hausdorff distances over a testing set of 100 subjects.

Acc.		Syn-Net	LI-Net	Ours
4×	Dice	$\uparrow 0.749 \pm 0.260$	0.805 ± 0.198	0.902 ± 0.089
	HD	$\downarrow 6.809 \pm 2.776$	6.557 ± 3.160	4.797 ± 2.064
8×	Dice	$\uparrow 0.748 \pm 0.258$	0.809 ± 0.192	0.902 ± 0.089
	HD	$\downarrow 6.794 \pm 2.868$	7.019 ± 3.521	4.772 ± 2.253
16×	Dice	$\uparrow 0.742 \pm 0.264$	0.800 ± 0.197	0.903 ± 0.085
	HD	$\downarrow 6.792 \pm 2.818$	6.841 ± 2.971	4.509 ± 2.068
32×	Dice	$\uparrow 0.723 \pm 0.287$	0.752 ± 0.242	0.902 ± 0.086
	HD	$\downarrow 7.383 \pm 3.122$	7.531 ± 3.131	4.665 ± 2.054
64×	Dice	$\uparrow 0.733 \pm 0.261$	0.799 ± 0.190	0.902 ± 0.085
	HD	$\downarrow 7.543 \pm 2.972$	6.706 ± 2.567	4.911 ± 2.356

produce better approximations of the true anatomy. We suspect that our model’s ability to attend globally across all time frames plays a key role.

6 Conclusion

To the best of our knowledge, this is the first study that explores the prediction of cardiac segmentation maps directly from sparse under-sampled k-space measurements without an explicit image reconstruction step. Our results show that transformer architectures are capable of extracting global features from sparse k-space measurements and improve segmentation performance over image-based baselines at high acceleration factors.

References

1. Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., et al.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of cardiovascular magnetic resonance* **20**(1), 65 (2018)
2. Huang, Q., Yang, D., Yi, J., Axel, L., Metaxas, D.: FR-Net: Joint reconstruction and segmentation in compressed sensing cardiac MRI pp. 352–360 (2019)
3. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General Perception with Iterative Attention pp. 4651–4664 (18–24 Jul 2021)
4. Petersen, S.E., Matthews, P.M., Francis, J.M., Robson, M.D., Zemrak, F., Bouber-takh, R., Young, A.A., Hudson, S., Weale, P., Garratt, S., et al.: UK Biobank’s cardiovascular magnetic resonance protocol. *JCMR* pp. 1–7 (2015)
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation pp. 234–241 (2015)
6. Schlemper, J., Oktay, O., Bai, W., Castro, D.C., Duan, J., Qin, C., Hajnal, J.V., Rueckert, D.: Cardiac MR segmentation from undersampled k-space using deep latent representation learning pp. 259–267 (2018)

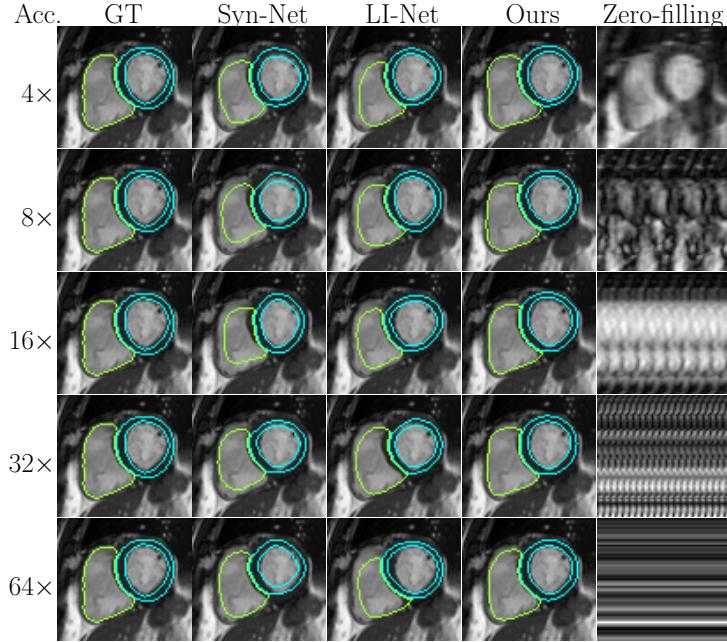


Fig. 2: Test set segmentations predicted from different models over varying acceleration factors. The last column visualizes the undersampled k-space measurements in a time frame.

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)