2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Segmentation of Cardiac Structures via Successive Subspace Learning with Saab Transform from Cine MRI

Xiaofeng Liu[1], Fangxu Xing[1], Hanna K. Gaggin[2], Weichung Wang[3], C.-C. Jay Kuo[4], *Fellow, IEEE,* Georges El Fakhri[1], *Fellow, IEEE,* Jonghye Woo[1], *Member, IEEE*

***Abstract*— Assessment of cardiovascular disease (CVD) with cine magnetic resonance imaging (MRI) has been used to non-invasively evaluate detailed cardiac structure and function. Accurate segmentation of cardiac structures from cine MRI is a crucial step for early diagnosis and prognosis of CVD, and has been greatly improved with convolutional neural networks (CNN). There, however, are a number of limitations identified in CNN models, such as limited interpretability and high complexity, thus limiting their use in clinical practice. In this work, to address the limitations, we propose a lightweight and interpretable machine learning model, successive subspace learning with the subspace approximation with adjusted bias (Saab) transform, for accurate and efficient segmentation from cine MRI. Specifically, our segmentation framework is comprised of the following steps: (1) sequential expansion of near-to-far neighborhood at different resolutions; (2) channel-wise subspace approximation using the Saab transform for unsupervised dimension reduction; (3) class-wise entropy guided feature selection for supervised dimension reduction; (4) concatenation of features and pixel-wise classification with gradient boost; and (5) conditional random field for post-processing. Experimental results on the ACDC 2017 segmentation database, showed that our framework performed better than state-of-the-art U-Net models with 200× fewer parameters in delineating the left ventricle, right ventricle, and myocardium, thus showing its potential to be used in clinical practice.**

***Clinical relevance*— Delineation of the left ventricular cavity, myocardium, and right ventricle from cardiac MR images is a common clinical task to establish diagnosis and prognosis of CVD.**

I. Introduction

Cardiovascular disease (CVD) continues to be the cause of the largest portion of morbidity and mortality globally, accounting for over 18 million deaths globally [1]. Assessment of CVD with cine magnetic resonance imaging (MRI) has been shown to provide a non-invasive way to evaluate the detailed morphology and function of the heart. In particular, cine MRI is considered to be the most accurate imaging modality for assessing various quantitative parameters with important prognostic implications.

[1]X. Liu, F. Xing, G. El Fakhri, and J. Woo are with the Gordon Center for Medical Imaging, Dept. of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[2]H. Gaggin is with Division of Cardiology, Corrigan Minehan Heart Center and Dept. of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[3]W. Wang is with Institute of Applied Mathematical Sciences, National Taiwan University, Taipei, Taiwan

[4]C.-C. J. Kuo is with Dept. of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA.

Segmentation of the left ventricle (LV), right ventricle (RV), and myocardium (MYO) from cardiac cine MR images plays an important role in characterizing clinically important parameters [2], such as ejection fraction (EF), end diastolic volume (EDV), end systolic volume (ESV), and myocardial mass. These parameters, in turn, can be used to identify disease phenotypes, stratify disease risks, and develop diagnostic and prognostic tools [3]. In clinical practice, semi-automated segmentation is still predominantly used, partly due to the lack of fully-automated and accurate segmentation tools [4], which is time-consuming and suffers from inter-observer variability.

With the recent progress of deep learning [5], numerous convolutional neural networks (CNN) models, e.g., U-Net [6], have been developed, demonstrating their accuracy in many medical image analysis tasks [7]. While deep learning has achieved impressive results for segmentation and classification, a number of challenges arise in developing and deploying deep learning models for clinical applications [7]. First, CNN models typically require a large number of labeled training datasets [5]. Sparse and inaccurate labels caused by privacy issues and the high cost of labeling, however, lead to difficulty in collecting sufficient and high-quality training sample datasets [5]; with the limited training datasets, an accurate model fitting at the training stage is challenging. Recently, to address this, efforts have been made to generate samples using data augmentation or adversarial training [8], which, however, results in an unavoidable problem of appearance shift between real and generated data. Second, importantly, many CNN models are seen as a "black-box" model [9], [5]. Accordingly, CNN models remain largely elusive how a particular CNN model makes a decision and when it can be trusted. Therefore, it is crucial to develop an explainable model that works with a limited number of datasets for clinical applications.

To address the aforementioned challenges, in this work, we propose to develop a lightweight, interpretable, and fully-automated segmentation framework with successive subspace learning (SSL) [10]. Specifically, our framework is comprised of the following steps: (1) sequential expansion of near-to-far neighborhood at different resolutions; (2) channel-wise subspace approximation using the subspace approximation with adjusted bias (Saab) transform for unsupervised dimension reduction; (3) a novel class-wise entropy guided feature selection for supervised dimension reduction; (4) concatenation of features and pixel-wise classification with gradient boost; and (5) conditional random field for post-

processing.

To the best of our knowledge, this is the first attempt at exploring the SSL framework with the Saab transform for a segmentation task. Our framework is lightweight and interpretable, yet achieving a superior segmentation performance with 200× fewer parameters, compared with state-of-the-art U-Net models.

II. Methodology

A. Fundamentals of SSL and Saab Transform

Inspired by the recent stacked design of CNN models, the SSL principle [10] has been targeted for classifying 2D natural images (e.g., PixelHop [11], [12]), 3D MR images [13], and point clouds (e.g., PointHop [12]). In each layer of SSL, the Saab transform [14], a variant of Principal Component Analysis (PCA), is used as an alternative to nonlinear activation, thereby alleviating the sign confusion problem [9]. Furthermore, the Saab transform is deemed more interpretable than nonlinear activation functions in CNNs [14], [15], as the model parameters are computed stage-by-stage in a feedforward manner, without backpropagation. Accordingly, the training of our SSL-based method is more efficient and interpretable than that of CNN models [11].

B. Our Saab-based SSL segmentation Framework

In this work, we have a 2D MR image $\mathbf{x} \in \mathbb{R}^{H\times W\times 1}$ and its corresponding label $\mathbf{y} \in \mathbb{R}^{H\times W\times 4}$, where H and W denote the horizontal and vertical dimensions, respectively. The channel of the gray-value sample is 1, and the label of each pixel is encoded as a four-dimensional one-hot vector for four tissue classes. The architecture of our framework is illustrated in Fig. 1, as detailed below.

1) Module 1: Unsupervised Feature Selection: We first construct I cascade SSL units and $I-1$ max-pooling operations to extract the attributes at different spatial scales in the unsupervised Module 1. Similar to PixelHop [11], in each SSL unit, we construct the neighboring region on the $H\times W$ plane. For instance, in the first SSL unit, for the single-channel data, we construct the 3×3 region for each pixel position. Each of them is then flattened to a 9-dimensional vector. With a padding operation, $\mathbf{x}$ is transformed to a cubic with the size of $H\times W\times 9$. Then, the Saab transform is used for unsupervised dimension reduction in the channel direction. Each 9-dimensional vector is mapped to a F_1-dimensional feature vector, where F_1 is a hyperparameter to control the output dimension of the first PixelHop unit.

Specifically, the terms, direct current (DC) and alternating current (AC), are adopted from the circuit theory. In the first Saab transform, we configure one DC and F_1-1 AC anchor vectors with the size of $H\times W\times 1$. Then, the c-th dimension of f can be an affine transform of x, i.e.,

$$f_c = a_c^T x + b_c, \quad c = 0, 1, \cdots, F_1 - 1, \tag{1}$$

and the Saab transform has a special design of the anchor vector $a_c \in \mathbb{R}^{1\times(H\times W\times 1)}$ and the bias term $b_c \in \mathbb{R}$ [14].

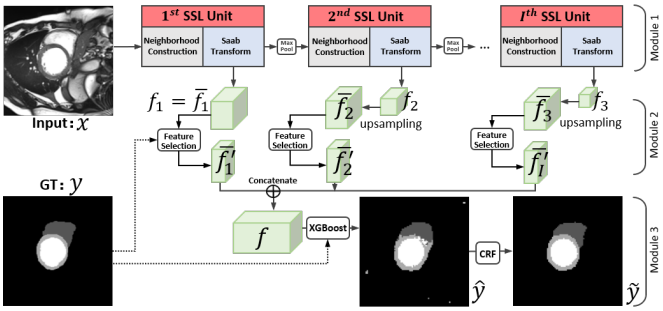

Fig. 1. Illustration of our proposed framework using the Saab transform, which consists of 3 modules.

Similar to [14], we can set $b_c \equiv d\sqrt{F_1}, d \in \mathbb{R}$, and divide the anchor vector into two categories:

$$\begin{aligned} &\bullet \text{ DC anchor vector } \quad a_0 = \frac{1}{\sqrt{H\times W\times 1}}(1,\cdots,1)^T, \\ &\bullet \text{ AC anchor vector } \quad a_c, \quad c = 1,\cdots,F_1-1. \end{aligned} \tag{2}$$

After computing $f_1 \in \mathbb{R}^{H\times W\times F_1}$, we half its spacial size with the max-pooling operation to $\frac{H}{2}\times\frac{W}{2}\times F_1$ and send to the next SSL unit. With the multi-channel input, the neighborhood construction involves $3\times 3\times F_1$ region at each pixel position. Then, the neighborhood union is flattened to a vector, which is further processed by the Saab transform for dimension reduction. The detailed structure of our module 1 is provided in Table I.

With the cascaded SSL units, the neighborhood union is correlated with more pixels of $\mathbf{x}$ to extract global information. This process is similar to CNN models in that a larger reception field is achieved in the deeper layers.

2) Module 2: Supervised Feature Selection: In what follows, we resort to the supervised dimension reduction based on class-wise entropy-guided feature selection to tailor the discriminative feature for our segmentation task.

Because of the resolution deduction in each unit, we have different spatial size of the extracted features. The features in the later units correspond to a larger reception field (i.e., more pixels) in $\mathbf{x}$ and $\mathbf{y}$. To match the features in these units with the original pixels, we resize $f_2,\cdots,f_I$ to the size of f_1 and denote as $\overline{f_2},\cdots,\overline{f_I}$. Therefore, we have $\overline{f_1} \in \mathbb{R}^{H\times W\times C_1}$, $\overline{f_2} \in \mathbb{R}^{H\times W\times C_2}$, $\cdots$, $\overline{f_I} \in \mathbb{R}^{H\times W\times C_I}$.

Because of the disparate importance, depending on the different channels for the segmentation decision, it is necessary to make supervised feature selection. In related developments, PixelHop++ [16] proposes to classify each channel with the size of $\mathbb{R}^{H\times W}$ and select the channels with low cross-entropy score. However, it is not applicable to segmentation as a channel selection, since the label in the segmentation task is pixel-wise and the feature of a pixel in each channel is only a scalar, making it challenging to be used as a feature for a classifier.

Instead, we propose to select the channel with the small entropy of each class. Specifically, we would encourage the feature of a pixel in each channel to be similar, if the label of the corresponding pixels is the same class. We denote the

TABLE I

THE DETAILED STRUCTURE OF OUR 4 CONSECUTIVE SSL UNITS

Input Size	Type	Filter Shape
$[224 \times 224 \times 1]$	Saab Trans	F_1 kernels of 3×3
$[224 \times 224 \times F_1]$	MaxPool	(2×2)-(1×1)
$[112 \times 112 \times F_1]$	Saab Trans	F_2 kernels of 3×3 for F1 channels
$[112 \times 112 \times F_2]$	MaxPool	(2×2)-(1×1)
$[56 \times 56 \times F_2]$	Saab Trans	F_3 kernels of 3×3 for F2 channels
$[56 \times 56 \times F_3]$	MaxPool	(2×2)-(1×1)
$[28 \times 28 \times F_3]$	Saab	F_4 kernels of 3×3 for F3 channels

feature of a pixel in each channel p_i^c for the i-th pixel of a class in the c-th channel. The entropy of a sample can be:

$$H = \sum_{j=1}^{4} H_j, \quad H_j = -\sum_i p_i^c \log p_i^c, \tag{3}$$

where we use j to index the four classes in our segmentation task. After calculating the entropy of four classes for each channel, we rank the entropy in descending order. Then, we select the top 80% channels for the subsequent pixel-wise classification task.

3) Module 3: Information fusion for segmentation and post-processing: With the extracted features $\overline{f_2'}, \cdots, \overline{f_I'}$ with both the Saab transform and class-wise entropy guided selection, we concatenate them along with the channel dimension to get the feature $f \in \mathbb{R}^{H \times W \times C}$. The channel dimension C is the sum of all channels in $\overline{f_2'}, \cdots, \overline{f_I'}$. Each feature vector on the $H \times W$ plane of f corresponding to an original pixel in $\mathbf{x}$ or $\mathbf{y}$. Then, we carry out the pixel-wise classification for each of C dimensional features with a classifier. We empirically choose the extreme gradient boosting (XGBoost) [17], which is an optimized distributed gradient boosting library designed to be highly efficient and flexible. XGBoost is trained to learn the correlation of pixel-wise feature and ground truth pixel class label in our training set.

We note that with a limited number of SSL units, it is challenging to support the reception field to cover all of the pixels for global perception. In contrast, too many SSL units will lead to very low resolution in the later units, which is not sufficient to support the pixel-wise segmentation. In addition, the channel size of the later units will be very large, leading to a long and indiscriminative feature vector, which can distract the pixel-wise classification.

To balance this conflict, we propose to validate the most effective I and adopt the well-established post-processing tool of conditional random field (CRF) to further refine the segmentation results and get the final results of our framework $\tilde{y} = CRF(\hat{y})$.

III. EXPERIMENTS

To demonstrate the performance of our Saab transform-based SSL framework, we validated it on the Automated Cardiac Diagnosis Challenge (ACDC 2017) database, which contains 100 subjects. The cine MRI short-axis slices were acquired with 1.5T or 3.0T MRI scanners. The acquired cine MRI short-axis slices covered the LV, RV, and MYO from the base (upper slice) to the apex (lower slice), with 5–8 mm slice thickness, 5 or 10 mm inter-slice gap and the spatial resolution of 1.37–1.68 mm^2.

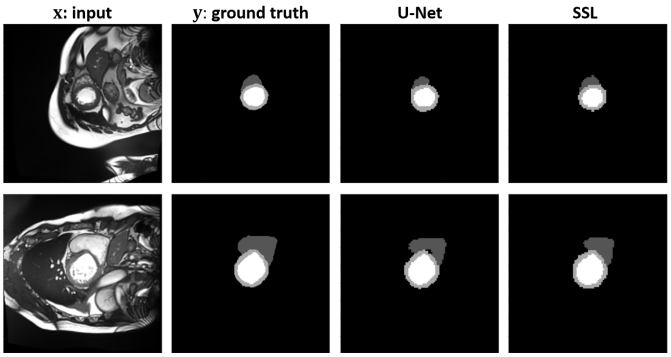

Fig. 2. Comparison of the segmentation results with 60 subjects for training.

TABLE II

COMPARISON OF THE DICE SCORE WITH 50 TRAINING SUBJECTS

Methods	**Parameters**	RV	MYO	LV	**Average**
U-Net	5.88 M	81.45%	78.54%	90.07%	83.35%
AttenUNet	6.40 M	81.02%	78.40%	89.32%	82.91%
SSL	**0.03 M**	**82.91%**	**81.83%**	**90.62%**	**85.12%**

TABLE III

COMPARISON OF THE DICE SCORE WITH 60 TRAINING SUBJECTS

Methods	**Parameters**	RV	MYO	LV	**Average**
U-Net	5.88 M	85.54%	78.81%	**92.15%**	85.50%
AttenUNet	6.40 M	**85.64%**	77.37%	91.29%	84.76%
SSL	**0.03 M**	83.89%	**82.57%**	91.62%	**86.03%**

For each patient, the delineations of the LV, RV, and MYO, were obtained by two clinical experts. On average, each subject had about 27 labeled slices. We reported the average Dice similarity score with 30 subjects for testing, 10 subjects for validation, and 50 or 60 subjects for training.

A. Implementation Details

All the experiments were implemented using Python on a server with a Xeon E5 v4 CPU/Nvidia Tesla V100 GPU with 128GB memory. We also used the widely adopted deep learning library, Pytorch, to implement U-Net [6] and AttnUNet [18]. For a fair comparison, we resized all of the slices to $224 \times 224 \times 1$, which was consistent with the input of the U-Net models.

We empirically used four SSL units and set F_1=5, F_2=10, F_3=30, and F_4=100. We note that the number of the Saab AC filters in the unsupervised dimension reduction procedure controls the preserved energy ratio.

B. Experimental Results

Fig. 2 shows the segmentation results of U-Net with ResNet50 backbone and our SSL framework. We can see that SSL is able to achieve comparable or even better performance than the widely used U-Net models.

For quantitative evaluation, we compared the Dice similarity score in Tables II and III, which used 50 or 60 subjects for training, respectively. Note that the larger Dice similarity score indicates the better segmentation performance. The

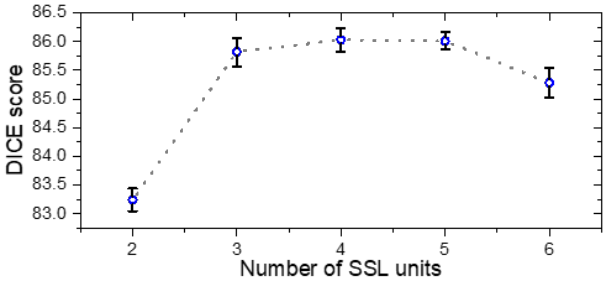

Fig. 3. Sensitivity analysis of the number of SSL units for the case of using 60 subjects in training.

TABLE IV
RESULTS OF ABLATION STUDIES WITH 60 TRAINING SUBJECTS

Methods	Average DICE
SSL	**86.03%**
SSL without CRF	84.76%
SSL without entropy-guided feature selection	85.91%

best results are bolded. With 50 subjects for training, our SSL framework outperformed U-Net [6] and attention-based U-Net [18] in all of the three classes. We can observe that with relatively limited training datasets, the performance of the CNN models is inferior to our framework. In addition, the statistics of the network parameters are provided and compared in Table II. We can see that the number of parameters of our SSL framework was about 200 times fewer than the popular U-Net structures [6], [18]. The much fewer parameters can largely alleviate the difficulty of a small number of training datasets. In the case of using 60 subjects for training, our SSL framework achieved better performance than the U-Net based methods in the average Dice similarity score.

C. Sensitivity Analysis and Ablation Study

With four SSL units, we achieved a state-of-the-art Dice similarity score in both 50 and 60 training subjects settings. The number of SSL units is important for our segmentation framework to balance the efficiency and perception area. The low resolution can be challenging to provide accurate information for fine-grained pixel-wise classification. In Fig. 3, we have shown the detailed sensitivity study using different SSL units. The standard deviation was computed with five random choices of training and validation splits. The class-wise entropy-guided feature selection was developed to simplify the subsequent classification modules. In addition, CRF was applied as a post-processing step. To demonstrate their effectiveness, we provide the ablation study in Table IV and the effect of CRF in Fig. 4.

IV. CONCLUSION

In this work, we presented a lightweight, interpretable, and fully-automated SSL framework with the Saab transform to segment the LV, RV, and MYO from cine MRI. A novel class-wise entropy-guided feature selection was proposed to achieve accurate segmentation. Our thorough experiments carried out using the ACDC 2017 database with different number of training subjects demonstrated that our framework achieved a superior performance, compared with the U-Net-based approaches, with about 200× fewer parameters.

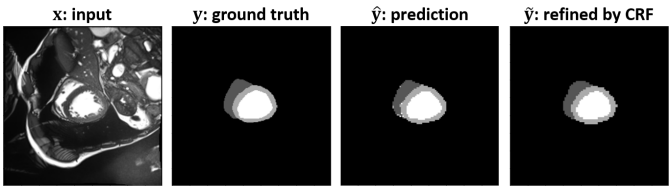

Fig. 4. Comparison of with or without CRF post-processing.

REFERENCES

[1] G. A. Roth, G. A. Mensah, and V. Fuster, "The Global Burden of Cardiovascular Diseases and Risks: A Compass for Global Action," *American College of Cardiology*, pp. 2980–2981, 2020.
[2] H. K. Gaggin and J. L. Januzzi Jr, "Biomarkers and diagnostics in heart failure," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1832, no. 12, pp. 2442–2450, 2013.
[3] A. Ammar, O. Bouattane, and M. Youssfi, "Automatic cardiac cine mri segmentation and heart disease classification," *Computerized Medical Imaging and Graphics*, vol. 88, p. 101864, 2021.
[4] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.
[5] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.
[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[7] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.
[8] X. Liu, B. V. Kumar, P. Jia, and J. You, "Hard negative generation for identity-disentangled facial expression recognition," *Pattern Recognition*, vol. 88, pp. 1–12, 2019.
[9] C.-C. J. Kuo, "Understanding convolutional neural networks with a mathematical model," *Journal of Visual Communication and Image Representation*, vol. 41, pp. 406–413, 2016.
[10] M. Rouhsedaghat, M. Monajatipoor, Z. Azizi, and C.-C. J. Kuo, "Successive subspace learning: An overview," *arXiv*, 2021.
[11] Y. Chen and C.-C. J. Kuo, "Pixelhop: A successive subspace learning (SSL) method for object recognition," *Journal of Visual Communication and Image Representation*, p. 102749, 2020.
[12] M. Zhang, H. You, P. Kadam, S. Liu, and C.-C. J. Kuo, "Pointhop: An explainable machine learning method for point cloud classification," *IEEE Transactions on Multimedia*, 2020.
[13] X. Liu, F. Xing, C. Yang, C.-C. J. Kuo, S. Babu, G. E. Fakhri, T. Jenkins, and J. Woo, "Voxelhop: Successive subspace learning for als disease classification using structural MRI," *arXiv preprint arXiv:2101.05131*.
[14] C.-C. J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen, "Interpretable convolutional neural networks via feedforward design," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 346–359, 2019.
[15] F. Fan, J. Xiong, and G. Wang, "On interpretability of artificial neural networks," *arXiv preprint arXiv:2001.02522*, 2020.
[16] Y. Chen, M. Rouhsedaghat, S. You, R. Rao, and C.-C. J. Kuo, "Pixelhop++: A small successive-subspace-learning-based (SSL-based) model for image classification," *arXiv preprint arXiv:2002.03141*, 2020.
[17] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, 2015.
[18] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.