Improving Cardiac MRI Convolutional Neural Network Segmentation on Small Training Datasets and Dataset Shift: A Continuous Kernel Cut Approach

Fumin Guo[a,b,*], Matthew Ng[a,b], Maged Goubran[a,b], Steffen E. Petersen[c], Stefan K. Piechnik[d], Stefan Neubauer[d], Graham Wright[a,b]

[a]*Sunnybrook Research Institute, University of Toronto, Toronto, Canada.*
[b]*Department of Medical Biophysics, University of Toronto, Toronto, Canada*
[c]*NIHR Biomedical Research Centre at Barts, Queen Mary University of London, London, UK.*
[d]*Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK.*

Abstract

Cardiac magnetic resonance imaging (MRI) provides a wealth of imaging biomarkers for cardiovascular disease care and segmentation of cardiac structures is required as a first step in enumerating these biomarkers. Deep convolutional neural networks (CNNs) have demonstrated remarkable success in image segmentation but typically require large training datasets and provide suboptimal results that require further improvements. Here, we developed a way to enhance cardiac MRI multi-class segmentation by combining the strengths of CNN and interpretable machine learning algorithms. We developed a continuous kernel cut segmentation algorithm by integrating normalized cuts and continuous regularization in a unified framework. The high-order formulation was solved through upper bound relaxation and a continuous max-flow algorithm in an iterative manner using CNN predictions as inputs. We applied our approach to two representative cardiac MRI datasets across a wide range of cardiovascular pathologies. We comprehensively evaluated the performance of our approach for two CNNs trained with various small numbers of training cases, tested on the same and different datasets. Experimental results showed that our approach improved baseline CNN segmentation by a large margin, reduced CNN segmentation variability substantially, and achieved excellent segmentation accuracy with minimal extra computational cost. These results suggest that our approach provides a way to enhance the applicability of CNN by enabling the use of smaller training datasets and improving the segmentation accuracy and reproducibility for cardiac MRI segmentation in research and clinical patient care.

Keywords: Cardiac MRI segmentation, normalized cuts, continuous max-flow, convex optimization

1. Introduction

According to the World Health Organization, cardiovascular disease is a leading cause of death globally and accounted for 17.9 million or 31% of global deaths in 2016 (Organization, 2017). A major goal of cardiovascular disease research is to identify disease phenotypes, stratify disease risks, and develop therapeutic interventions with an ultimate objective of improving patient outcomes (Blankstein, 2012). Cardiac magnetic resonance imaging (MRI) provides unique advantages for non-invasive evaluation of cardiac structural-functional information with exquisite soft tissue contrast, high spatial/temporal resolution and no ionizing radiation (Wu, 2017). In particular, cardiac MRI provides a wealth of imaging biomarkers, including but not limited to ejection fraction, myocardium mass, chamber volume and wall thickness (Peng et al., 2016).

To generate these imaging biomarkers and facilitate clinical applications of cardiac MRI for patient care, automatic or semi-automatic segmentation of cardiac structures is required. However, cardiac MRI segmentation is particularly challenging (Zhuang, 2013) because of: 1) large shape variations of cardiac structures; 2) high anatomical and geometric complexity; 3) image signal intensity inhomogeneity and weak boundaries; 4) motion/blood flow artefacts and partial volume effects; and 5) unbalanced size between cardiac structures. A number of studies have contributed to single cardiac structure segmentation; here we aimed to simultaneously extract multiple structures from cardiac MRI, including the left ventricle cavity (LV), left ventricle myocardium (Myo.), and right ventricle cavity (RV).

1.1. Related Work

Expert manual segmentation is generally time-consuming, tedious, expensive, and prone to observer variability. In the past decades, a number of algorithms have been developed; most of them employed thresholding, statistical shape/appearance models, deformable models, graphical models, atlases, and learning-based methods, as previously reviewed (Zhuang, 2013; Petitjean and Dacher, 2011; Petitjean et al., 2015). Although promising, these methods provide limited generalization capability mainly because they rely on hand-crafted shallow image features with limited representation/discrimination capability (Shen et al., 2017; Dou et al., 2017; Litjens et al., 2017; Tran, 2016).

[*]Corresponding author: Sunnybrook Research Institute, University of Toronto, Toronto, Canada, M4N 3M5.
Email address:
fumin.guo@sri.utoronto.ca,fumin.guo@utoronto.ca (Fumin Guo)

Recently, deep convolutional neural networks (CNN) have achieved remarkable success (Shen et al., 2017; Litjens et al., 2017) in many medical image applications, including classification, segmentation, registration, and computer-aided diagnosis. The power of CNN mainly stems from the deep architecture that discovers highly discriminative features through hierarchical information abstraction (Shen et al., 2017; Litjens et al., 2017; Zeiler and Fergus, 2014). Regarding cardiac MRI segmentation, Wolterink et al. (2016) employed a dilated CNN trained on three orthogonal planes, and similar work was performed by Mortazi et al. (2017). Tran (2016) adopted a fully convolutional network (FCN) and explored transfer learning on multiple cardiac MRI datasets. Yang et al. (2016) embedded feature extraction and label fusion into a CNN for atlas-based segmentation. Payer et al. (2017) employed two CNNs: one for ROI cropping and the other for region configuration learning. Rupprecht et al. (2016) used a CNN to predict a flow vector to guide the evolution of an active contour. Avendi et al. (2016) utilized a CNN for LV localization and a stacked autoencoder for LV shape inference followed by level-set refinement. Similarly, Ngo et al. (2017) employed deep belief networks for LV detection and LV endo/epicardium rough segmentation with level-set post-processing. Zheng et al. (2018) combined pre/post-processing and a segmentation consistency prior with CNN. Other studies (Dou et al., 2017; Bai et al., 2018; Khened et al., 2018; Oktay et al., 2018) performed cardiac MRI segmentation by training CNNs in an end-to-end manner.

Deep learning using CNN presents several unique challenges (Shen et al., 2017; Litjens et al., 2017), including the requirements of large and diverse datasets, high quality expert manual annotations, computationally-intensive hardware resources, and a lack of reproducibility and clear interpretation of the learned models. Numerous efforts have attempted to address these issues by designing CNN with advanced architecture, incorporating expert domain knowledge and data pre/post-processing techniques. Previous studies (Zheng et al., 2015; Johnson et al., 2016; Chen et al., 2018) have highlighted the advantages of combining CNN with straightforward machine learning techniques and task-specific knowledge to boost overall performance, and some of them have achieved great success. For example, Chen et al. (2018) applied a Conditional Random Field (CRF) on top of a CNN and achieved remarkable improvements in natural image segmentation. Zheng et al. (2015) formulated CRF as a recurrent layer as a part of a single unified CNN. In medical image segmentation, researchers have investigated the combination of CNN with active contours (Rupprecht et al., 2016), level-sets (Avendi et al., 2016; Ngo et al., 2017), CRF (Dou et al., 2017; Wang et al., 2018), graph cut (Dangi et al., 2018; Li et al., 2019) and continuous max-flow (Guo et al., 2018).

1.2. Contributions

In this work, we proposed a way to address some of the issues associated with CNN for cardiac MRI segmentation. We summarize our contributions as follows:

- We developed a continuous kernel cut segmentation method combining normalized cut and image-grid continuous regularization to improve coarse and suboptimal segmentation provided by CNNs on cardiac MRI data. Normalized cut provides balanced partitioning without a shrinking bias that is difficult to achieve using traditional Potts or CRF models. Continuous regularization demonstrates sub-pixel segmentation accuracy with low computational burden and high computational efficiency without metrication errors. However, the unique properties of normalized cut and continuous regularization have not been studied for medical imaging applications. To the best of our knowledge, this is the first investigation of continuous kernel cut for cardiac MRI segmentation.

- We utilized CNN coarse outputs to initialize the continuous kernel cut model and designed image features exploiting image signal intensity and spatial location information to facilitate the segmentation. We developed a way to efficiently optimize the high-order segmentation formulation through upper bound linearization and convex relaxation in the spatially continuous settings. We derived a novel CNN-guided continuous max-flow model under a primal-dual perspective and developed a computationally efficient iterative continuous max-flow numerical solver with guaranteed algorithm convergence.

- We applied our approach for a standard U-net and a state-of-the-art CNN trained with various small numbers of subjects and tested it on the same and different datasets from two representative cardiac MRI databases. We comprehensively evaluated the performance of our approach and observed much improved segmentation accuracy and reduced segmentation variability compared with the baseline CNN and other post-processing methods. Results suggest that our approach provides a way to enhance the applicability of CNN by utilizing smaller dataset/expert annotations and improving the generalizability, which are critically required in research and clinical studies.

2. Methods

Figure 1 provides the workflow of the proposed cardiac MRI multi-class segmentation algorithm. Briefly, we trained a U-net and Isensee's network (Isensee2017) (Isensee et al., 2017) that provided excellent performance in a recent cardiac MRI segmentation challenge (Bernard et al., 2018), and applied the trained models to the test datasets to generate the labeling probability maps for region R_l, $l \in L = \{LV, Myo., RV\}$. The probability maps were used to initialize the continuous kernel cut segmentation framework that is comprised of normalized cut and image-grid continuous regularization. We linearized the high-order normalized cuts term through upper bound relaxation. The upper bound relaxed normalized cuts formulation and continuous regularization were solved using a continuous max-flow framework, which involved iterative maximization/minimization of max-flow variables and the labeling function. Upon max-flow convergence, the optimal labeling function was saved in a solution queue and was looped back to re-initialize the continuous kernel cut formulation. The upper bound relaxation was iterated until convergence and the last

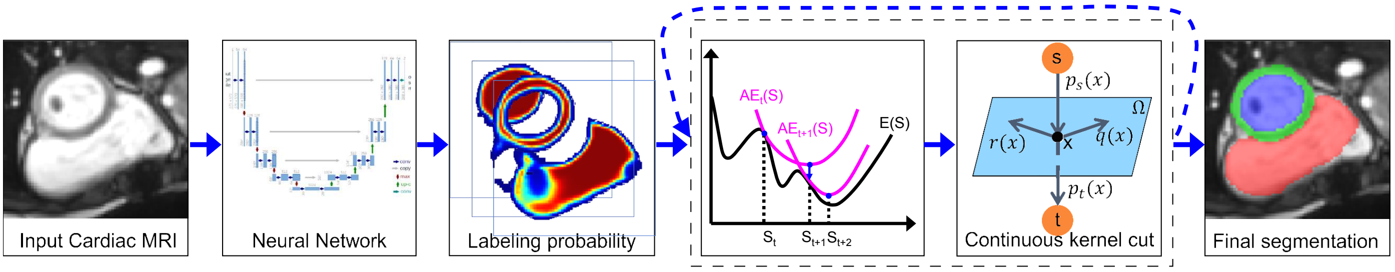

Figure 1: Cardiac MRI multi-class segmentation workflow. Cardiac MR images were entered into a trained U-net that generated labeling probability maps for each structure. The probability maps were used to initialize and guide a kernel cut segmentation module in an iterative continuous max-flow framework.

labeling function in the solution queue was used as the final segmentation. The upper bound relaxation and continuous max-flow global optimization scheme provide guaranteed algorithm convergence. In the following sections, we provide the details of the main components of the segmentation algorithm in Fig. 1.

2.1. *Baseline Convolutional Neural Network*

While many CNNs have been developed for medical image segmentation, here we employed a widely-used standard 2D U-net (Ronneberger et al., 2015) for the cardiac MRI multi-class segmentation task. The contracting path consists of five blocks each comprising two layers of 3×3 convolution filters and an exponential linear unit, followed by 2×2 max-pooling (stride = 2). The expanding path is symmetric with the contracting path and consists of five blocks each comprising 2×2 up-sampling (stride = 2), feature map concatenation from the corresponding contracting path, and two 3×3 convolutional layers with exponential linear activation. The concatenation procedure provides a way to combine abstract context information and corresponding local image features to aid feature classification. The final feature maps are processed by two 1×1 convolutional layers followed by a softmax operation, resulting in pixel-wise probability maps $\chi_l(x) \in [0, 1]$ for each class $l \in L$. In addition, we employed the Isensee2017 (Isensee et al., 2017) that utilized an ensemble of 2D and 3D U-nets with a deep supervision mechanism. The 2D and 3D U-nets were trained with five-fold cross validation, and the resulting 10 models were used to generate 10 sets of segmentation predictions, which were ensembled to provide the final segmentation and probability maps.

2.2. *CNN-guided Kernel Cut Segmentation Model*

2.2.1. *Continuous Kernel Cut Segmentation Formulation*

We aimed to segment a cardiac MR image $I(x)$, $x \in \Omega$, into mutually disjoint regions R_l, $l \in L$. By virtue of graph theory, the multi-class segmentation task can be achieved by way of normalized cuts that minimize the fraction of cut cost to the total edge connection between the segmented region and the entire graph (Shi and Malik, 2000). Alternatively, normalized cuts can be equivalently interpreted as the ratio of total edge connections within each segmentation region to that between the segmentation and the entire image domain, as reflected by the first term in Eq. (1). Previous studies have shown that normalized cuts provide segmentation with balanced partitioning and are able to avoid isolated small segmentation regions (Shi and Malik, 2000) with superior performance to graph cut and GrabCut (Tang et al., 2018). Here we proposed a continuous kernel cut segmentation model comprised of normalized cuts and spatially continuous regularization for cardiac MRI multi-class segmentation as follows:

$$E(R) := \sum_{l\in L} -\frac{Assoc(R_l, R_l)}{Assoc(\Omega, R_l)} + \alpha\,|\partial R_l| \,, \tag{1}$$

where the first term represents normalized cuts with $Assoc(X, Y) = \int_X \int_Y A(x, y)\, dydx$, A is a symmetric matrix and $A(x, y)$ measures the similarity between the features at voxels x and y; $\alpha\,|\partial R_l|$ measures the total surface area of each segmentation region weighted by $\alpha \geq 0$. Previous studies (Tang et al., 2018) employed normalized cuts for natural image segmentation and the optimization problem was solved using graph cut in the discrete domain. Although promising, it is well known that discrete graph cut methods are limited by grid bias or metrication error (Boykov and Kolmogorov, 2003) and extra computational load is required to alleviate these issues (Klodt et al., 2008). Here, for the first time we proposed a continuous kernel cut segmentation model that combines normalized cuts and spatially continuous regularization for cardiac MRI segmentation. The proposed approach is fundamentally different from the previous work (Tang et al., 2018) in that we investigated the optimization problem in the continuous settings and solved the complicated problem by means of convex relaxation. In addition, we derived a CNN-guided continuous max-flow model that differs from previous studies, for which we developed a computationally efficient continuous max-flow numerical solver under a primal and dual perspective. Compared with discrete graph cut, the convex continuous optimization methods provide sub-pixel segmentation accuracy with lower computational burden and higher computational efficiency without metrication error (Yuan et al., 2010; Nieuwenhuis et al., 2013).

2.2.2. *CNN-guided Continuous Kernel Cut Segmentation*

We introduced the indicator function $u_l(x)$ for region R_l with value of “1” for $x \in R_l$ and “0” otherwise, i.e., $u_l(x) \in \{0, 1\}$, for the kernel cut-based segmentation formulation (1). In addition, we utilized the baseline CNN segmentation probability maps $\chi_l(x)$ to guide the continuous kernel cut segmentation model by enforcing the “similarity” between the two sets of segmentation, e.g., $u_l(x) \approx \chi_l(x)$, for $\forall x \in R_l$. To this end, we derived a deep learning-guided continuous kernel cut (DLKC) segmentation formulation as follows:

$$E(u) := \sum_{l \in L} -\frac{u_l^T A u_l}{\mathbf{1} A u_l} + \int_\Omega g(x)\,|\nabla u_l(x)|\,dx + \beta \int_\Omega |u_l(x) - \chi_l(x)|\,dx\,, \quad s.t. \sum_{l \in L} u_l(x) = 1\,, \qquad (2)$$

where u_l and $\mathbf{1}$ (all-ones matrix) are vector indicator functions of region R_l and the entire image domain, respectively; $g(x)$ weights the total variation-measured surface area of each region based on image contrast edges, e.g., $g(x) := \lambda_1 + \lambda_2 * \exp(-\lambda_3 * |\nabla I(x)|)$, where $\lambda_{1,2,3} \geq 0$ are constants.

The DLKC segmentation formulation Eq. (2) gives rise to a challenging optimization problem because of the high-order normalized cuts, non-smooth total variation formulation and the absolute function terms (Yuan et al., 2010). In the next section, we provide a way to efficiently solve the multi-class segmentation problem Eq. (2).

2.3. *Efficient Optimization of Continuous Kernel Cut* (2)

2.3.1. *Upper Bound Relaxation of Continuous Kernel Cut* (2)

A non-linear function, including distribution matching, moment constraints and shape modeling, demonstrates outstanding performance and has been widely used in various research areas including medical imaging. However, one of the main challenges associated with these non-linear functions is the resulting difficult optimization. Bound optimization provides a generalized way to efficiently optimize any high-order non-linear functions, including but not limited to the continuous kernel cut model in Eq. (2), by alternatively tackling a simpler auxiliary function of the original formulation, assuming that the auxiliary formulation is easier to optimize (Ayed et al., 2013). In general, given any high-order formulation $D(x)$, $AD(x, x_t)$ is the upper bound (auxiliary) function of $D(x)$ at given solution x_t if:

$$D(x_t) = AD(x_t, x_t) \qquad (3a)$$

$$D(x) \leq AD(x, x_t)\,. \qquad (3b)$$

The upper bound relaxation scheme (3a) and (3b) provides a way to indirectly minimize $D(x)$ by dealing with the simpler auxiliary function $AD(x, x_t)$ instead of $D(x)$. For example, the optimal solution $x_{t+1} = \arg\min_x AD(x, x_t)$ also decreases the energy of $D(x)$! The proof follows that: $AD(x_{t+1}, x_t) \leq AD(x_t, x_t) = D(x_t)$ and $D(x_{t+1}) \leq AD(x_{t+1}, x_t)$ following (3a) and (3b), respectively. By implementing these steps iteratively, we can obtain a series of solutions $\{x_t, x_{t+1}, \ldots, x_{t+n}\}$ that progressively decrease the energy of the original high-order formulation $D(x)$.

Proposition 1. *Given the segmentation solution $\hat{u}_l$, $l \in L$, the normalized cuts term in Eq.* (2) *is upper-bounded by:*

$$AD_l(u_l, \hat{u}_l) = \left\langle u_l(x), A\mathbf{1}\frac{\hat{u}_l^T A \hat{u}_l}{(\mathbf{1}A\hat{u}_l)^2} - \frac{2A\hat{u}_l}{\mathbf{1}A\hat{u}_l} \right\rangle. \qquad (4)$$

Proof. The normalized cuts term in Eq. (2) is concave with respect to u_l by virtue of the negative definiteness of its Hessian matrix. Therefore, given solution $\hat{u}_l$, the normalized cuts term is upper bounded by its tangent function with a slope of $A\mathbf{1}\frac{\hat{u}_l^T A\hat{u}_l}{(\mathbf{1}A\hat{u}_l)^2} - \frac{2A\hat{u}_l}{\mathbf{1}A\hat{u}_l}$. Through simple re-organization, we have: $AD_l(u_l, \hat{u}_l) \geq -\frac{u_l^T A u_l}{\mathbf{1}A\hat{u}_l}$ and $AD_l(\hat{u}_l, \hat{u}_l) = -\frac{\hat{u}_l^T A\hat{u}_l}{\mathbf{1}A\hat{u}_l}$ that satisfy (3a) and (3b). $\langle , \rangle$ denotes the inner product of two functions. Therefore, Proposition 1 is proved. □

Based on Proposition 1, we can conclude that the DLKC segmentation Eq. (2) is upper-bounded by:

$$AE(u, \hat{u}) = \sum_{l \in L} AD_l(u_l, \hat{u}_l) + \int_\Omega g(x)\,|\nabla u_l(x)|\,dx + \beta \int_\Omega |u_l(x) - \chi_l(x)|\,dx\,, \quad s.t. \sum_{l \in L} u_l(x) = 1\,, \qquad (5)$$

2.3.2. *Dual Optimization of the Upper-bound Relaxed Continuous Kernel Cut* (5)

Eq. (5) provides a way to deal with the simpler auxiliary function $AE(u, \hat{u})$ instead of $E(u)$ in Eq. (2) to solve the multi-class segmentation problem. However, direct optimization of Eq. (5) is challenging because of the non-convex function $u(x) \in \{0, 1\}$. By means of convex relaxation (Yuan et al., 2010), the non-convex segmentation problem in Eq. (5) can be solved by relaxing $u(x) \in \{0, 1\}$ to convex sets $[0, 1]$ as follows:

$$\min_{u_l \in [0,1]} AE(u, \hat{u})\,, \quad s.t. \sum_{l \in L} u_l(x) = 1\,. \qquad (6)$$

In the following section, we provided an approach to efficiently solve the convex relaxed continuous min-cut segmentation formulation Eq. (6).

Based on the well established max-flow/min-cut theories (Ford and Fulkerson, 1962), we proposed a CNN-guided continuous max-flow segmentation model that is mathematically equivalent to the convex-relaxed min-cut segmentation formulation in Eq. (6) without explicitly involving the non-smooth total variation or the absolute difference terms. The proposed continuous max-flow network is based on the previous flow configuration $\{ps, pt_l, q_l\}(x)$, which was described by Guo et al. (2015, 2019). The CNN segmentation prediction *prior* was encoded in our flow network by adding an extra flow $r_l(x)$ (Fig. 1) at each location x in each image domain Ω_l, $l \in L$. We sought to maximize the flow that is allowed to send through the flow-network as follows:

$$\max_{ps, pt_l, q_l, r_l} \int_\Omega ps\,dx - \sum_{l \in L} \int_\Omega r_l \cdot \chi_l\,dx\,, \qquad (7)$$

subject to:

- Flow capacity constraints:
  $pt_l(x) \leq AD_l(x)$, $|q_l(x)| \leq g(x)$ and $|r_l(x)| \leq \beta$;
- Flow conservation constraints:
  $G_l(x) = (\operatorname{div} q_l - ps + pt_l + r_l)(x) = 0\,;$

for $\forall x \in \Omega$, $l \in L$.

Proposition 2. *The CNN-guided continuous max-flow model* (7) *is equivalent to the convex relaxed CNN-guided continuous min-cut model* (6)*:*

$$(7) \iff (6)\,.$$

Proof. We multiplied the labeling function $u_l(x)$ and the respective flow conservation constraints $G_l(x)$ associated with Eq. (7), and added this term to Eq. (7). This gives the Lagrangian function of Eq. (7) and through simple re-organization, we have:

$$\max_{ps,pt_l,q_l,r_l} \min_{u_l} \left\langle 1 - \sum_{l \in L} u_l, ps \right\rangle + \sum_{l \in L} \Big\{ \langle u_l, pt_l \rangle + \langle u_l, \operatorname{div} q_l \rangle + \langle u_l, r_l \rangle - \langle \chi_l, r_l \rangle \Big\}, \tag{8}$$

subject to the flow capacity and conservation constraints associated with Eq. (7). Similar to the previous analyses (Guo et al., 2016, 2017), maximization over unconstrained source flow $ps(x)$, constrained sink flow $pt_l(x)$, and spatial flow $q_l(x)$ leads to the region layout constraint: $\sum_l u_l(x) = 1$, the first and second terms in Eq. (6) (derived from Eq. (5)), respectively. The third term in Eq. (6), by virtue of conjugate representation of an absolute function, can be re-written as follows:

$$\beta \int_\Omega |u_l - \chi_l| \, dx = \max_{|r_l| \le 1 \cdot \beta} \int_\Omega r_l \cdot (u_l - \chi_l) dx \, .$$

Based on the above analyses, we have: Eqs. (7) ⇔ (8) ⇔ (6). Therefore, Proposition 2 is proved. □

2.3.3. A CNN-guided Kernel Cut-based Continuous Max-flow Segmentation Algorithm

Proposition 2 shows that, instead of dealing with the challenging continuous min-cut segmentation problem in Eqs. (5) and (6), we can alternatively tackle the simpler continuous max-flow formulation in Eq. (7) subject to a series of linear and convex constraints. Following the Lagrangian theories (Bertsekas, 1999), we derived an efficient numerical solver of Eq. (7) by maximizing over $\{ps, pt, q, r\}(x)$ and minimizing over $u(x)$ as follows:

$$\int_\Omega ps \, dx - \sum_{l \in L} \Big\{ \langle r_l, \chi_l \rangle - \langle u_l, G_l \rangle + \frac{c}{2} \|G_l\|^2 \Big\}, \tag{9}$$

where $c > 0$ is a constant (Yuan et al., 2010).

The augmented Lagrangian function in Eq. (9) can be optimized by splitting the whole optimization problem into a series of sub-problems. Each of the sub-problems tackles a single variable of $(ps, pt_l, q_l, r_l)(x)$, and can be efficiently solved following the steps 7-11 in Algorithm 1. The details of the numerical implementation of Eq. (9) are similar to previous work (Guo et al., 2016, 2017). The max-flow algorithm was implemented in parallel on a graphics processing unit (GPU) for speedup.

2.4. CNN-guided Continuous Kernel Cut Algorithm

The upper bound relaxation of Eq. (2) involves iterative implementation of Eq. (5). The continuous max-flow numerical solver in Eq. (9) provides a way to solve the upper-bound relaxed kernel cut-based segmentation model in Eqs. (5) and (6) with given solution $\hat{u}$. With the help of Proposition 1, we derived a CNN-guided continuous kernel cut segmentation Algorithm 1 for the multi-class segmentation problem Eq. (2) as follows:

Algorithm 1 CNN-guided continuous kernel cut algorithm

```
1: Train CNN and generate segmentation probability maps $\chi$;
2: Let $j = 1$ and start the $j^{th}$ upper-bound iteration as follows:
3: repeat
4:     Compute the upper bound $AD_l(u_l, \hat{u}_l)$ following Eq. (4) with given $\hat{u}$ ($\hat{u} = \chi$ for $j = 1$);
5:     Initialize $(ps, pt_l, q_l, r_l)(x)$ for $k = 1$ and start the $k^{th}$ max-flow iteration as follows:
6:     repeat
7:         Maximize $q_l(x)$ using Chambolles gradient projection: $q_l^{k+1} = \mathrm{Proj}_{|q_l(x)| \le g(x)}(q_l^k + \tau \nabla(\operatorname{div} q_l^k + G_l^k - \operatorname{div} q_l^k - u_l^k / c))$, where $\tau$ is the projection step size (Chambolle, 2004);
8:         Maximize $pt_l(x)$ by: $pt_l^{k+1} = \min(pt_l^k - G_l^k + u_l^k / c, AD_l)$;
9:         Maximize $r_l(x)$ by: $r_l^{k+1} = \min(\max(r_l^k - G_l^k + u_l^k / c - \chi_l / c, \ -\beta), \beta)$;
10:        Maximize $ps(x)$ by: $ps^{k+1} = \frac{1}{4}(1/c + \sum_{l=1}^4 G_l^k + ps^k - u_l^k / c)$;
11:        Update $u_l(x)$ by: $u_l^{k+1} = u_l^k - c \cdot G_l^{k+1}$;
12:        $k \leftarrow k + 1$;
13:    until Convergence
14:    $\hat{u} \leftarrow u^*$;
15:    $j \leftarrow j + 1$;
16: until Convergence
```

3. Experiments

3.1. Cardiac MRI Datasets

We applied the developed segmentation algorithm to two representative cardiac MRI datasets:

3.1.1. Automated Cardiac Diagnosis Challenge (ACDC)

The ACDC dataset (Bernard et al., 2018) comprises 100 patients evenly distributed in 5 categories of well-defined pathologies: normal condition, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle. 2D short axis cine images were acquired from the base to the apex using a bSSFP sequence at 1.5T or 3.0T under breath-hold conditions in a gated manner and stacked to generate a 3D image (inplane voxel size = 0.7-1.9 mm^2, slice thickness = 5-10 mm, inter-slice gap = 5 mm (sometimes), 6-18 slices, 28-40 phases). For each patient, the 3D images at end-diastole (ED) and end-systole (ES) were manually segmented for LV, Myo., and RV by a single clinical expert as the reference segmentation.

3.1.2. UK BioBank (UKBB)

The UKBB dataset (Petersen et al., 2015) consists of a large population of mainly healthy volunteers. 3D images covering the LV and RV were generated by stacking a series of 2D cine images acquired using a bSSFP sequence at 1.5T under breath-hold and gated conditions (inplane voxel size = 1.8-2.3 mm^2, slice thickness = 8 mm, inter-slice gap = 2 mm (sometimes), ∼10 slices, 50 phases). Manual segmentation of LV, Myo., RV was performed at ES and ED phases by 8 observers guided by 3 principal investigators. One hundred subjects were selected for our algorithm segmentation (access application #2964).

3.2. Algorithm Implementation

3.2.1. CNN Training and Testing

Figure 1 shows a schematic of the cardiac MRI multi-region segmentation pipeline. We randomly selected 50, 10, and 40 subjects out of the 100 subjects in the UKBB datasets as the entire training, validation and testing datasets, respectively. The ACDC dataset was divided in the same manner and the training, validation and testing datasets consist of 10, 2 and 8 subjects respectively from each of the 5 pathology categories. U-net and Isensee2017 were trained from scratch on 6 occasions using 5, 10, 20, 30, 40, and 50 subjects randomly selected (equal number in each pathology category for the ACDC dataset) from the entire training dataset, while validation and testing were performed on the same subjects in each dataset. We optimized the configuration of the U-net in this work by varying: the number of convolution levels (n=4, 5), batch normalization of the input image (True, False), batch normalization of intermediate features (True, False), activation function (ReLU, ELU), and training loss (DSC, cross entropy). Among the resulting 32 models, we determined an optimal U-net with 5 convolution levels using batch normalization of the image and feature layers, ELU activation, and cross entropy loss. CNN training, validation and testing were performed on a NVIDIA GPU (Tesla P100, NVIDIA Corp., Santa Clara, CA, USA) and the trained U-net model with the highest dice-similarity-coefficient (DSC) on the validation dataset was selected. For U-net training, the initial number of feature channels was 32. The feature maps were doubled and halved for the next layer in the contracting and expanding paths, respectively. The network was trained for 300 epochs each comprising 100 updates with a batch size of 2 subjects. Data augmentation including random rotation (-60 to 60 degrees), translation (-60 to 60 pixels), scaling (0.7 to 1.3 times), and intensity alternations (0.7 to 1.3 times) was performed. The network parameters were optimized using the ADAM solver with a learning rate of 1×10^{-4}. Isensee2017 was implemented using the default settings (Isensee et al., 2017).

3.2.2. CNN Coarse Outputs Post-processing

For each CNN (U-net and Isensee2017), the resulting 12 sets (6 sets for CNN trained on ACDC and tested on ACDC, and 6 sets for CNN trained on UKBB and tested on UKBB) of segmentation and the associated inference probability maps were post-processed using Matlab 2013a (The Mathworks, Inc., Natick, MA) and CUDA (CUDA v8.0, NVIDIA Corp., Santa Clara, CA, USA) on a Linux Desktop (Ubuntu 14.04, Intel(R) CPU i7-7770K, 4.2 GHz, 16G RAM) with a NVIDIA GPU (GeForce, GTX TITAN X, NVIDIA Corp., Santa Clara, CA, USA).

We explored a number of different post-processing methods all initialized with U-net outputs, including: 1) continuous max-flow (+CMF) as a counterpart of graph cut, 2) deep learning-guided continuous kernel cut without the similarity prior (+nDLKC: $\beta = 0$ in Eq. (2)), 3) deep learning-guided continuous kernel cut (+DLKC) in Eq. (2), 4) morphological refinements (+Morph.) by keeping the largest connected components and filling small holes, 5) fully connected CRF (+CRF) (Kamnitsas et al., 2017). For +CMF, the U-net output probability maps $\chi_l(x)$, $x \in \Omega$, $l \in L$, were used to generate the data term in forms of $\int_\Omega -log(\chi_l) \cdot u_l dx$; the regularization term was based on image edge contrast $g(x)$ as in Eq. (2) (please see Guo et al. (2016) for CMF setup details). For +nDLKC, the probability maps were used in the same manner as +DLKC in Eq. (2). For both +DLKC and +nDLKC, the similarity matrix A in Eq. (2) was generated using the k-nearest neighbour (kNN) (i.e., $A(x, y) = 1$ if x is among the first K nearest neighbours of y and 0 otherwise, where K is a positive integer). For Isensee2017, we implemented +DLKC and +nDLKC to refine the baseline segmentation. For both U-net and Isensee2017, the post-processing methods were independently optimized for the 12 sets of baseline segmentation using another 10 subjects in each dataset that were not used for algorithm segmentation performance evaluation. We note that K and β represent the important parameters of the +DLKC algorithm in Eq. (2). In general, greater K captures a larger neighbourhood and *vice versa*; greater β leads to higher similarity between the initial (χ) and final segmentation (u). For baseline CNN with high segmentation accuracy (e.g., CNN trained and tested on the same dataset with NTrS=20, 30, 40, 50), we utilized $K = 5$ to capture a relatively small neighbourhood and $\beta = 10$ to enforce a high similarity between the baseline and final segmentation. For moderate and fairly low baseline CNN accuracy (e.g., CNN trained with NTrS=5, 10 and tested on the same dataset), we used $K = 10$ to capture a larger neighbourhood and $\beta = 5$ to enforce a lower similarity between the baseline and final segmentation. In cases of fairly low baseline CNN performance (e.g., CNN trained with NTrS=5, 10, 20, 30, 40, 50 and tested on different datasets), we used $K = 10$ and $\beta = 0$ to evolve the initial CNN masks. Other parameters, including λ_1, λ_2, and λ_3 that are associated with $g(x)$ in Eq. (2), are related to image edge contrasts and have less influence on algorithm optimization. For +DLKC/nDLKC and +CMF, we optimized the edge regularization term $g(x)$ in Eq. (2) with $\{\lambda_1, \lambda_2, \lambda_3\} = \{0.1, 10, 200\}$ for the ACDC and $\{\lambda_1, \lambda_2, \lambda_3\} = \{0.05, 5, 200\}$ for the UKBB dataset. Other ways of using the continuous kernel cut module include manual and atlas-based initializations. However, these methods are generally labor intensive, time-consuming, and provide initializations that are inferior to CNN.

We also explored the generalizability of +nDLKC by training a U-net and Isensee2017 on one dataset and testing on the other dataset. For both U-net and Isensee2017, we applied the 6 models trained on the ACDC dataset to the UKBB test dataset. We refined the 6 sets of baseline CNN predictions using the +nDLKC algorithms optimized for CNN training and testing on the UKBB datasets. The same procedures were applied to the 6 sets of predictions obtained by applying the 6 CNN models trained on the UKBB dataset to the ACDC test dataset.

3.3. Evaluation

3.3.1. Segmentation Accuracy

Algorithm segmentation masks provided by baseline CNN (U-net and Isensee2017), and the combination of baseline CNN and different post-processing methods (Sec. 3.2.2) were compared with expert manual outputs using DSC, average symmetric surface distance (ASSD), and Hausdorff distance (HD) (Bai et al., 2018) for LV, Myo., and RV. DSC measures the overlap ratio of algorithm and manual segmentation with 0 denoting no overlap and 1 perfect overlap, i.e., DSC $\in [0,1]$. ASSD is calculated as: $\frac{1}{2}\{\frac{1}{|\partial R_a|}\sum_{p\in\partial R_a} d(p,\partial R_m) + \frac{1}{|\partial R_m|}\sum_{p\in\partial R_m} d(p,\partial R_a)\}$, where $d(p,\partial R_m)$ represents the minimal distance from a point p to surface ∂R_m, and $\left|\partial R_{a,m}\right|$ is the number of points that comprise $\partial R_{a,m}$. HD is defined as: $\max(\max_{p\in\partial R_a} d(p,\partial R_m), \max_{p\in\partial R_m} d(p,\partial R_a))$, which measures the extreme distance between two surfaces.

3.3.2. Continuous Kernel Cut Segmentation Improvements

For each set of the baseline U-net (12 sets for training and testing on the same dataset and 12 sets for training and testing on different dataset) and Isensee2017 (12 sets for training and testing on the same dataset and 12 sets for training and testing on different dataset) segmentation/predictions, we calculated the percent improvement (*PI*) in the mean and standard deviation (SD) of segmentation accuracy provided by +DLKC/nDLKC. We calculated *PI* as: $PI = \frac{M_{+pp}-M_{cnn}}{M_{cnn}} * 100\%$, where M_{cnn} and M_{+pp} represent the mean and SD of segmentation accuracy provided by baseline CNN and +DLKC/nDLKC, respectively. Paired t-tests were performed for statistical comparison of +DLKC/nDLKC *vs* baseline U-net and Isensee2017. Normality of data was determined using Shapiro-Wilk tests and when significant, the Mann-Whitney U tests for nonparametric data were performed. In addition, the influence of baseline U-net and Isensee2017 outputs on the continuous kernel cut (+DLKC/nDLKC) refinement was determined using Pearson correlation coefficient (r) between baseline CNN and +DLKC/nDLKC post-processing accuracy measurements. All statistics were performed using GraphPad Prism v7.00 (GraphPad Software, San Diego, CA). Results were considered statistically significant when the probability of making a type I error was less than 5% ($p < 0.05$).

3.3.3. Computation Time

The computational efficiency of our segmentation framework was determined using the runtime required by CNN training, CNN testing, and post-processing.

4. Results

4.1. Image Features for Continuous Kernel Cut

Figure 2 provides examples for visualization of the voxel-to-voxel similarity matrix A in Eq. (2) using features of image signal intensity only, e.g., $f(x) = \{I(x)\}$, and the combination of signal intensity and pixel coordinate information, e.g., $f(x) = \{I(x), \omega X(x), \omega Y(x), \omega Z(x)\}$, where ω was 0.02 for the ACDC and 0.2 for the UKBB dataset. The initial segmentation was provided in the forms of user seeds (for illustration only) and CNN predictions in cyan, and the voxels that were "similar" to the initial segmentation under the kNN criteria were highlighted in yellow. Figure 2 shows that image features $f(x) = \{I(x)\}$ resulted in scattered neighbours that were used to generate the final segmentation and a large portion of the neighbours were outside the myocardium. In contrast, features $f(x) = \{I(x), \omega X(x), \omega Y(x), \omega Z(x)\}$ compensated for image signal intensity inhomogeneity and led to a spatially more compact distribution of the neighbours.

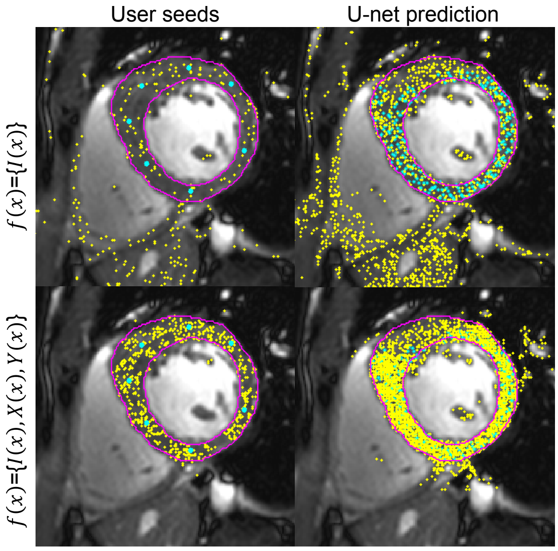

Figure 2: Pixel labeling probability with given initial segmentation using K-nearest neighbour kernel. Top and bottom rows shows neighbouring pixels (yellow) of initial segmentation (cyan) in the forms of user seeds (left column) and U-net predictions (right column) using features of $f(x) = \{I(x)\}$ and $f(x) = \{I(x), \omega X(x), \omega Y(x), \omega Z(x)\}$, respectively. U-net prediction (cyan dots) was generated by down-sampling the original U-net prediction by a factor of five for better visualization

.

4.2. Representative Segmentation and Algorithm Refinements

Figure 3A shows a representative cardiac MR image with heterogeneous image signal intensities and the continuous kernel cut segmentation iterations (Iter.). Intermediate segmentation results are shown for myocardium in Fig. 3B-F with kernel cut inputs (CNN prediction used for Iter.=1 in B) in cyan and outputs in yellow. Similarly, Fig. 3H-L shows RV segmentation (U-net results used for Iter.=1 in H) and the kernel cut segmentation energy through iterations in G). For this example, segmentation converged within 30 iterations. Figure 4 shows representative segmentation of basal to apical slices in 2D and the entire volume rendered in 3D for a patient in the ACDC dataset. Figure 5 provides examples of problematic U-net initial segmentation, as indicated by white circles, and the improved outcomes achieved using +DLKC. Figure 5A shows the LV cavity was substantially under-segmented, which would affect LV cavity volume and dimension quantification. Although there was decent overlap between baseline U-net and manual segmentation of the myocardium as shown in Fig. 5B, the edge localization error may affect regional myocardium wall thickness measurements. However, these issues were much improved by +DLKC that explicitly enforces the similarity of image features within each class and encourages the alignment between segmentation boundaries and image edges. The U-net segmentation leakage in Fig. 5C may be partially improved by removing the small isolated islands but the large false positive region connected to the RV cannot be easily removed using morphological operations. In addition, it is not uncommon to observe false positive regions that are larger than the target objects and morphological operations do not generalize well for these situations. In contrast, our approach provides a generalized way that integrates human prior and well-defined mathematical formulations to refine and improve CNN coarse outputs.

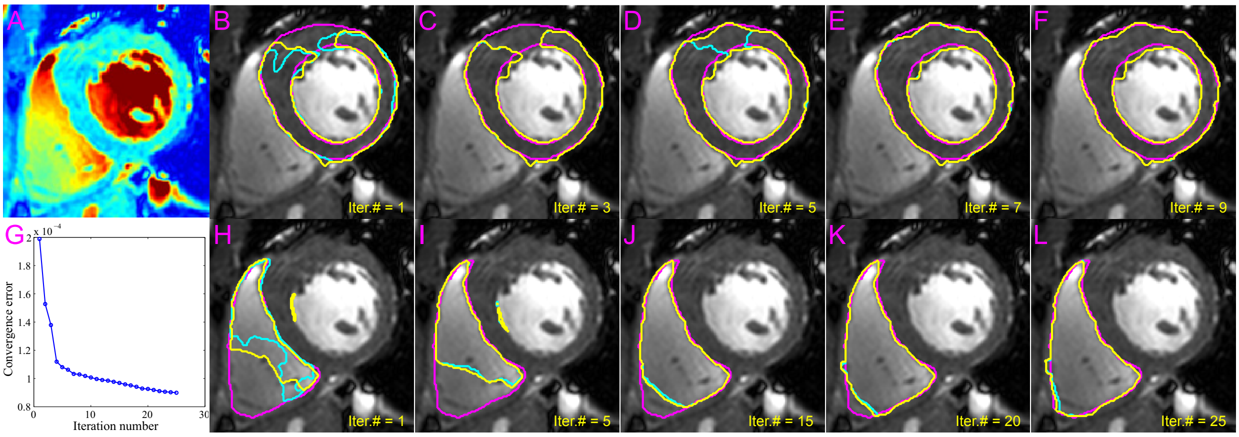

Figure 3: Cardiac MR image signal intensity inhomogeneity and +DLKC segmentation iterations. A) Cardiac MR image signal intensity inhomogeneity. B-F) +DLKC myocardium segmentation initialized by U-net prediction (cyan in B) and refined by DLKC (yellow) through iterations. G) +DLKC RV segmentation energy through iterations (H-L) using U-net initialization (cyan in H) and DLKC refinement (yellow). Purple contours represent manual segmentation and cyan contours indicate +DLKC refinements of previous outcomes (yellow) at each iteration (Iter.).

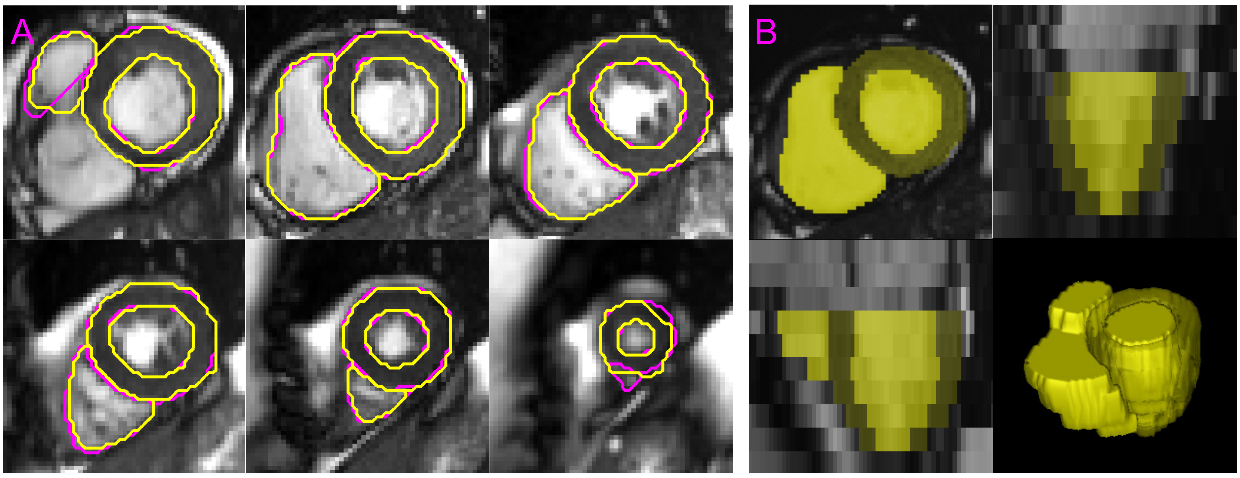

Figure 4: Representative cardiac MRI LV, Myo., and RV segmentation results. +DLKC segmentation results are shown in 2D basal-to-apical slices in A) and rendered in 2D and 3D views in B). Purple and yellow contours represent manual and +DLKC segmentation, respectively.

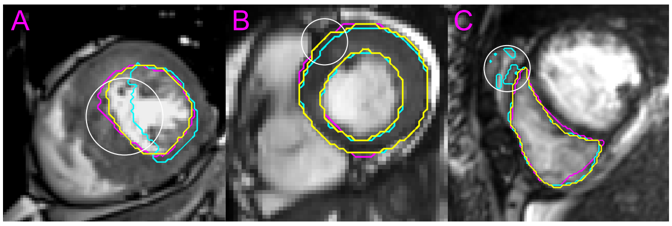

Figure 5: Sub-optimal U-net segmentation (cyan), +DLKC improvements (yellow), and manual segmentation (purple). White circles indicate problematic U-net coarse segmentation and improvements yielded by +DLKC.

4.3. *Continuous Kernel Cut Overall Accuracy Improvements: CNN Trained and Tested on The Same Datasets*

Table 1 shows the percent improvements (*PI*) in the mean of the segmentation accuracy achieved by +DLKC *vs* baseline U-net for U-net trained with various number of training subjects (NTrS). For U-net trained (NTrS=5) and tested on the ACDC dataset, +DLKC improved baseline U-net DSC by {1.1%, 1.4%, and 1.3%}, ASSD by {47.2%, 54.0%, and 49.7%}, and HD by {67.6%, 73.2%, and 60.4%} for {LV, Myo., and RV}. For U-net trained (NTrS=5) and tested on the UKBB dataset, +DLKC slightly decreased DSC by 0.1-0.4% but substantially improved ASSD by 7.0-19.5% and HD by 11.6-47.8%. For both cases, we observed greater algorithm improvements for moderate baseline U-net results and lower improvements when the baseline U-net accuracy was high (Table 1 and supplementary Table S5). For example, +DLKC improved DSC by -0.3-0.5%, ASSD by 7.1-23.8%, and HD by 27.3-32.7% on the ACDC dataset for NTrS=20; +DLKC slightly decreased DSC by 0.3-0.7%, ASSD by 3.7-4.0%, but improved HD by 2.6-8.4% on the ACDC dataset for NTrS=50.

Table 2 provides the improvements of using +DLKC for Isensee2017 trained with various NTrS. For the ACDC test dataset, +DLKC improved DSC by 1.1-9.0%, ASSD by 11.1-32.8%, and HD by 16.2-28.3% for NTrS=5; +DLKC improved DSC by 0.0-0.1%, ASSD by -0.3-1.2%, and HD by 0.3-2.1% for NTrS=20; +DLKC improved DSC by 0.0-0.1%, ASSD by 0.1-2.5%, and HD by 0.1-1.6% for NTrS=50. Similar results were obtained for Isensee2017 trained and tested on UKBB.

Table 1: Improvements (%) in the **mean** of segmentation accuracy using +DLKC/nDLKC *vs* baseline **U-net** for a **U-net** trained with 5, 20 and 50 subjects from the same and different datasets (NTrS: number of training subjects)

		ACDC (Trained on ACDC)			UKBB (Trained on UKBB)			ACDC (Trained on UKBB)			UKBB (Trained on ACDC)		
NTrS	Metrics	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV
5	DSC	1.1	1.4	4.3	-0.1	-0.4	-0.3	15.0	5.0	15.6	1.2	2.5	1.8
	ASSD	47.2	54.0	49.7	17.7	19.5	7.0	20.1	24.7	14.6	60.9	68.1	50.0
	HD	67.6	73.2	60.4	44.4	47.8	11.6	36.2	12.7	16.6	75.1	83.1	46.4
20	DSC	0.1	-0.3	0.5	0.0	-0.5	0.0	2.6	3.4	3.6	-0.2	0.1	0.0
	ASSD	9.7	7.1	23.8	10.3	28.1	29.7	28.5	27.6	18.1	25.0	35.0	26.2
	HD	31.7	27.3	32.7	23.5	33.7	24.6	27.2	19.2	22.7	42.0	52.7	30.6
50	DSC	-0.3	-0.7	0.5	0.0	-0.6	-0.1	6.5	2.2	5.0	-0.1	0.2	0.9
	ASSD	-4.0	-3.7	12.6	33.7	17.6	7.5	27.0	11.0	32.2	30.8	25.7	37.7
	HD	2.6	8.4	18.4	60.3	45.0	20.7	30.4	4.7	15.0	43.6	39.4	36.0

Table 2: Improvements (%) in the **mean** of segmentation accuracy using +DLKC/nDLKC *vs* baseline **Isensee2017** for **Isensee2017** trained with 5, 20 and 50 subjects from the same and different datasets (NTrS: number of training subjects)

		ACDC (Trained on ACDC)			UKBB (Trained on UKBB)			ACDC (Trained on UKBB)			UKBB (Trained on ACDC)		
NTrS	Metrics	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV
5	DSC	1.1	1.3	9.0	-0.1	-0.3	0.9	2.9	5.6	12.6	1.1	0.6	19.6
	ASSD	11.1	19.3	32.8	-1.2	3.5	10.0	17.3	26.7	29.3	9.7	11.8	34.6
	HD	16.2	22.9	28.3	-1.2	1.0	4.7	21.4	24.9	16.3	17.6	17.6	30.3
20	DSC	0.0	0.0	0.1	0.0	0.0	0.0	1.8	4.4	5.9	-0.1	-0.7	4.6
	ASSD	-0.3	0.4	1.2	0.5	1.7	0.7	12.5	24.9	19.0	1.6	6.8	24.1
	HD	0.3	2.1	0.7	0.8	1.6	0.5	17.3	23.8	10.7	4.9	4.1	19.1
50	DSC	0.0	0.0	0.1	0.0	0.1	0.1	1.8	3.4	2.4	-0.2	-0.7	2.4
	ASSD	0.1	0.2	2.5	0.4	2.0	0.5	12.6	19.7	11.8	1.0	1.7	16.7
	HD	0.1	1.4	1.6	0.4	0.8	0.4	17.9	21.8	3.6	1.4	2.2	14.1

Supplementary Table S1 provides the details of the segmentation accuracy achieved by baseline U-net and Isensee2017 and various post-processing methods for CNN trained (NTrS=5, 10, 20, 30, 40, 50) and tested on the ACDC dataset. For example, for U-net trained with 20 subjects, +DLKC yielded DSC of {0.931±0.06, 0.863±0.06, 0.870±0.10}, ASSD of {1.27±0.82 mm, 1.37±0.64 mm, 2.35±1.98 mm}, and HD of {4.0±2.7 mm, 5.6±3.4 mm, 9.8±7.1 mm} for {LV, Myo., RV}. In contrast, the baseline U-net generated DSC of {0.930±0.06, 0.866±0.06, and 0.865±0.11}, ASSD of {1.41±1.05 mm, 1.47±0.94 mm, and 3.09±4.38 mm}, and HD of {5.8±6.8 mm, 7.8±8.3 mm, and 14.6±18.4 mm}, for {LV, Myo., RV}, respectively. As shown in supplementary Table S1 and S2, we obtained DSC of {0.928±0.03, 0.856±0.03, 0.891±0.04}, ASSD of {1.27±0.46 mm, 1.21±0.34 mm, 2.06±0.95 mm}, and HD of {3.6±1.7 mm, 4.4±1.4 mm, 7.9±3.3 mm}, for {LV, Myo., RV} using U-net+DLKC for U-net trained (NTrS=50) and tested on the UKBB dataset. We achieved DSC of {0.935±0.05, 0.879±0.04, 0.888±0.09} for {LV, Myo., RV} using U-net+DLKC for training (NTrS=50) and testing on the ACDC dataset. These accuracy results are higher than baseline U-net. In most but not all the experiments in supplementary Table S1 and S2, we observed higher segmentation accuracy using +DLKC *vs* +nDLKC, +CMF, and U-net alone. We did not observe much difference between +Morph. and baseline U-net, and in some cases, we obtained worse results. Examples include incorrect removal of target objects that are smaller than false positive regions outside and the inability to remove false positive areas that are connected to the target object as shown in Fig. 5C. We obtained overall segmentation accuracy boosts with +CRF but the improvements were much lower than +DLKC in general as shown in supplementary Table S1 and summarized in supplementary Table S11. For example, +DLKC yielded DSC of 0.797±0.16, ASSD of 3.48±2.76 mm, and HD of 13.3±9.6 mm for RV *vs* DSC=0.760±0.21, ASSD=4.25±4.00 mm, and HD=14.6±11.1 mm provided by +CRF for U-net trained (NTrS=10) and tested on the ACDC dataset.

For Isensee2017, supplementary Table S1 shows that baseline DSC = {0.871±0.15, 0.816±0.12, 0.673±0.28}, ASSD = {1.71±1.67 mm, 1.82±1.42 mm, 4.51±3.85 mm}, and HD = {6.0±6.7 mm, 8.4±6.9 mm, 18.2±12.2 mm} for {LV, Myo., and RV} for training (NTrS=5) and testing on the ACDC dataset. With the help of +DLKC, the segmentation accuracy was increased to DSC = {0.881±0.13, 0.827±0.10, 0.733±0.23}, ASSD = {1.52±1.36 mm, 1.47±0.82 mm, 3.03±2.43 mm}, and HD = {5.0±5.5 mm, 6.5±5.3 mm, 13.1±8.4 mm} for {LV, Myo., and RV}. Similarly, for Isensee2017 trained (NTrS=10) and tested on the UKBB dataset as shown in supplementary Table S2, +DLKC boosted: DSC from {0.916±0.07, 0.852±0.06, 0.855±0.11} to {0.919±0.06, 0.850±0.05, 0.863±0.09}, ASSD from {1.40±0.85 mm, 1.41±0.98 mm, 2.48±1.51 mm} to {1.37±0.71 mm, 1.34±0.75 mm, 2.30±1.21 mm}, and HD from {3.9±1.8 mm, 5.3±3.0 mm, 10.0±5.7 mm} to {3.7±1.6 mm, 5.0±2.2 mm, 9.3±4.9 mm} for {LV, Myo., and RV}. Similar to U-net, we observed greater improvements for moderate baseline Isensee2017 accuracy (e.g., NTrS=5, 10), and lower improvements towards minimal differences for high (NTrS=20, 30, 40, 50) baseline Isensee2017 performance for Isensee2017 trained and tested on the same datasets.

We refer readers to Table 1 and 2 for the details of *PI* in the mean of segmentation accuracy provided by +DLKC *vs* baseline U-net and Isensee2017. Figure 6 and 7 depict the distribution of the 6 sets (NTrS=5, 10, 20, 30, 40, 50) of segmentation accuracy yielded by baseline U-net, U-net+DLKC, baseline Isensee2017, and Isensee2017+DLKC for CNN trained and tested on the same dataset. We observed general trend of higher baseline segmentation accuracy when NTrS was in-

Table 3: Improvements (%) in the **SD** of segmentation accuracy using +DLKC/nDLKC *vs* baseline **U-net** for a **U-net** trained with 5, 20 and 50 subjects from the same and different datasets (NTrS: number of training subjects)

		ACDC (Trained on ACDC)			UKBB (Trained on UKBB)			ACDC (Trained on UKBB)			UKBB (Trained on ACDC)		
NTrS	Metrics	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV
5	DSC	8.6	8.5	13.5	5.6	0.8	-1.9	12.1	30.2	14.4	22.1	17.0	11.0
	ASSD	53.0	77.1	57.8	59.5	60.0	25.0	-25.7	58.5	40.2	84.3	83.8	67.6
	HD	69.2	72.5	61.4	84.8	80.8	36.3	10.1	38.3	28.9	89.4	91.7	58.8
20	DSC	5.2	3.4	5.5	5.1	0.3	2.1	7.8	14.8	8.9	5.8	2.5	1.7
	ASSD	22.4	31.5	54.8	37.9	76.6	79.7	70.7	61.0	30.9	67.2	72.1	62.4
	HD	59.9	58.9	61.6	57.8	74.7	60.8	50.6	36.9	19.1	81.4	85.7	54.2
50	DSC	1.7	1.3	9.7	2.7	2.8	4.6	4.7	24.1	16.6	3.3	1.2	10.1
	ASSD	4.0	6.3	38.6	73.3	72.2	14.4	17.6	-4.6	68.6	69.1	72.0	74.5
	HD	27.7	47.3	26.0	85.0	85.1	51.4	41.5	17.3	41.4	75.3	78.9	68.4

Table 4: Improvements (%) in the **SD** of segmentation accuracy using +DLKC/nDLKC *vs* baseline **Isensee2017** for **Isensee2017** trained with 5, 20 and 50 subjects from the same and different datasets (NTrS: number of training subjects)

		ACDC (Trained on ACDC)			UKBB (Trained on UKBB)			ACDC (Trained on UKBB)			UKBB (Trained on ACDC)		
NTrS	Metrics	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV	LV	Myo.	RV
5	DSC	13.4	15.8	17.3	12.5	9.6	28.3	9.3	14.7	9.4	6.5	11.7	10.2
	ASSD	18.3	41.9	37.0	21.8	24.9	40.9	22.0	42.0	35.1	22.4	63.2	47.5
	HD	17.8	23.1	30.9	23.3	12.1	29.2	28.6	30.6	27.9	37.6	53.1	44.0
20	DSC	1.2	0.6	0.9	0.5	0.0	0.4	12.9	19.9	5.0	9.7	-3.6	24.0
	ASSD	2.5	3.1	2.5	4.0	6.0	1.3	33.0	55.3	30.2	10.6	37.4	42.2
	HD	-1.3	2.5	1.0	-0.2	7.7	3.7	40.7	38.5	28.6	18.0	21.9	34.0
50	DSC	0.5	0.0	1.4	0.5	0.5	1.5	13.7	20.8	6.0	1.3	-4.4	18.0
	ASSD	1.8	-1.2	6.6	1.6	4.2	0.3	13.1	38.0	26.9	5.2	14.5	29.9
	HD	1.4	1.5	1.6	0.2	-0.1	-0.3	15.1	24.1	24.3	4.8	16.5	32.7

creased. The *PI* in the mean of segmentation accuracy was greater for moderate (e.g., NTrS=5, 10) baseline CNN accuracy, and lower for high (e.g., NTrS=40, 50) baseline accuracy for U-net and Isensee2017 trained and tested on the same datasets.

4.4. Continuous Kernel Cut Overall Accuracy Improvements: CNN Trained and Tested on Different Datasets

Table 1 also shows the *PI* in the mean of the segmentation accuracy provided by +nDLKC for a U-net trained on one dataset and tested on the other dataset. For example for U-net trained on the UKBB and tested on the ACDC database, +nDLKC yielded *PI* of 5.0-15.6% in DSC, 14.6-24.7% in ASSD, and 12.7-36.2% in HD for NTrS=5. Slightly lower but substantial boosts were observed as NTrS increases, e.g., *PI* was 2.2-6.5% for DSC, 11.0-32.2% for ASSD, and 4.7-30.4% for HD for NTrS=50. Similar trend was observed for U-net trained on the ACDC (NTrS=5, 10, 20, 30, 40, 50) and tested on the UKBB dataset. We observed lower improvements in DSC (towards minimal *PI* for NTrS=20, 30, 40, 50) but substantially greater boosts in ASSD and HD compared with U-net trained on the UKBB and tested on the ACDC dataset. For example, +nDLKC improved baseline U-net by -0.1-0.9% in DSC, 25.7-37.7% in ASSD, and 36.0-43.6% in HD for NTrS=50. For U-net, we observed generally greater improvements in DSC in the ACDC test dataset compared with UKBB, and greater boosts in ASSD/HD in the UKBB test dataset compared with ACDC. Interestingly, while +Morph. and +CRF failed to improve the baseline U-net results, +nDLKC yielded substantial improvements as shown in supplementary Table S3 and S4.

Isensee2017 segmentation results for training and testing on different datasets are shown in supplementary Table S3 and S4. The segmentation performance boosts achieved using +nDLKC are summarized in Table 2. For example for Isensee2017 trained on the UKBB (NTrS=50) and tested on the ACDC dataset, +nDLKC improved baseline DSC by 1.8-3.4%, ASSD by 11.8-19.7%, and HD by 3.6-21.8%. For Isensee2017 trained on the ACDC (NTrS=50) and tested on the UKBB dataset, +nDLKC yielded *PI* of -0.7-2.4% in DSC, 1.0-16.7% in ASSD, and 1.4-14.1% in HD. For Isensee2017 trained on ACDC and tested on UKBB, *PI* was greater for RV compared with LV and Myo..

Figure 8 and 9 depict the distribution of the 6 sets (NTrS=5, 10, 20, 30, 40, 50) of segmentation accuracy yielded by baseline U-net, baseline Isensee2017, and their combination with +nDLKC for CNN trained and tested on different datasets. Supplementary Table S3 and S4 provide the details of segmentation accuracy achieved using U-net, Isensee2017, and the combination with various post-processing methods for CNN trained (NTrS=5, 10, 20, 30, 40, 50) and tested on different datasets. For U-net trained on UKBB (NTrS=50) and tested on ACDC, baseline DSC = {0.644±0.26, 0.538±0.24, 0.564±0.28}, ASSD = {5.98±5.03 mm, 6.18±5.03 mm, 9.88±13.59 mm}, and HD = {22.5±16.8 mm, 26.9±15.9 mm, 28.7±18.9 mm} for {LV, Myo., RV}. The baseline U-net results were substantially improved to DSC = {0.686±0.25, 0.550±0.18, 0.592±0.24}, ASSD = {4.37±4.15 mm, 5.50±5.26 mm, 6.70±4.27 mm}, and HD = {15.6±9.8 mm, 25.6±13.1 mm, 24.4±11.1 mm} for {LV, Myo., RV} with the help of +nDLKC. For Isensee2017 trained on UKBB (NTrS=50) and tested on ACDC, +nDLKC improved the DSC from {0.844±0.20, 0.754±0.16, 0.701±0.28} to {0.859±0.17, 0.779±0.12, and 0.718±0.26}, ASSD from {1.96±1.48 mm, 2.18±1.51 mm, 3.31±2.29 mm} to {1.72±1.28 mm, 1.75±0.94 mm, 2.92±1.67 mm}, and HD from {6.1±5.5 mm, 9.4±8.3 mm, 13.6±8.4 mm} to {5.0±4.6 mm, 7.4±6.3 mm, 13.1±6.4 mm} for {LV, Myo., RV}. Similar results for U-net and Isensee2017 trained on ACDC and tested on UKBB are obtained as shown in supplementary Table S4. For these fairly

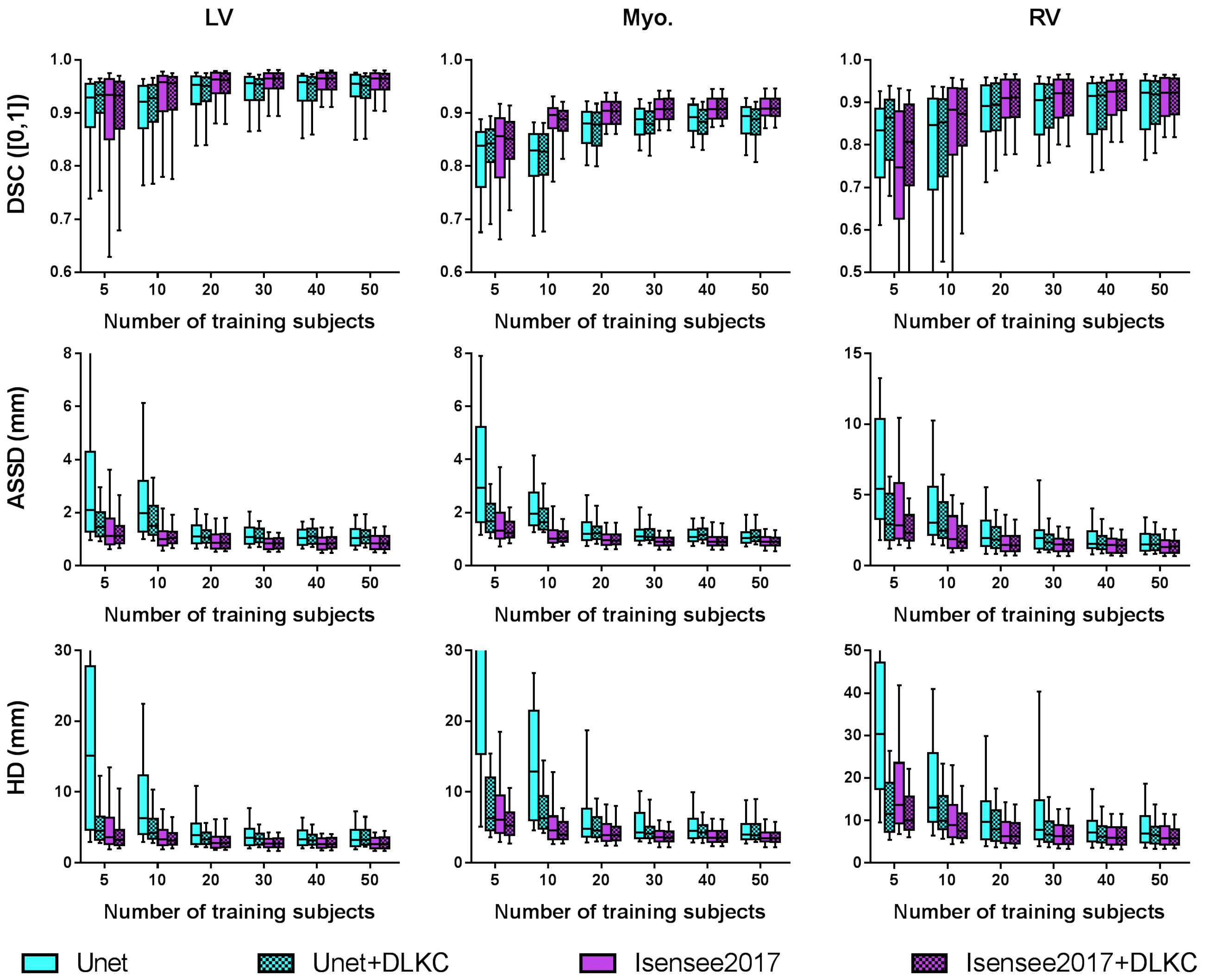

Figure 6: **ACDC** dataset segmentation results combining U-net/Isensee2017 and +DLKC (**U-net and Isensee2017 trained on the ACDC** training subjects). Box and whisker plots (box = 25-75th percentile; horizontal line = median; whisker = 10-90th percentile) from the top to bottom rows are shown for the distributions of DSC, ASSD, and HD for LV, Myo., and RV from left to right.

low baseline CNN accuracies, we achieved much higher *PI* by incorporating +nDLKC.

For all the experiments in Sec. 4.3 and 4.4, we achieved substantially improved segmentation accuracy with much reduced differences in segmentation performance (Figs. 6, 7, 8, and 9) within and between the ACDC and UKBB test datasets using +DLKC/nDLKC. In contrast, we did not observe much benefit using +Morph. and only partial improvements using +CRF that were substantially lower than +DLKC/nDLKC for a U-net trained and tested on the same and different datasets, as shown in supplementary Table S1, S2, S3 and S4.

4.5. Reduced Segmentation Variability

For all the experiments, we achieved reduced segmentation variability (Table 3 and 4) and more consistent results within and between experiments. As shown in Figs. 6, 7, 8, and 9, for U-net the 25-75th and 10-90th percentiles for all the experiments were generally narrower for +DLKC/nDLKC compared with baseline U-net. For Isensee2017, the 25-75th and 10-90th percentiles were narrower for Isensee2017 trained with NTrS=5, 10 and tested on the same dataset (Figs. 6 and 7), and much narrower for Isensee2017 trained and tested on different datasets (Figs. 8 and 9).

Quantitatively, the standard deviation (SD) was lower with the help of continuous kernel cut, as shown in supplementary Table S1, S2, S3 and S4. For the ACDC dataset (supplementary Table S1), SD of Myo. was reduced from 4.07 mm to 0.93 mm and from 3.80 mm to 0.99 mm by +DLKC for U-net trained with 5 and 10 subjects, respectively. Similarly, SD of Myo. was reduced from 1.42 mm to 0.82 mm and from 1.00 mm to 0.60 mm by +DLKC for Isensee2017 trained with 5 and 10 subjects, respectively. Very similar results were observed for ASSD and HD, and for U-net and Isensee2017 trained and tested on the same and other datasets (supplementary Table S1, S2, S3 and S4).

In addition, baseline U-net yielded lower accuracy when more subjects were used and this issue was mitigated using

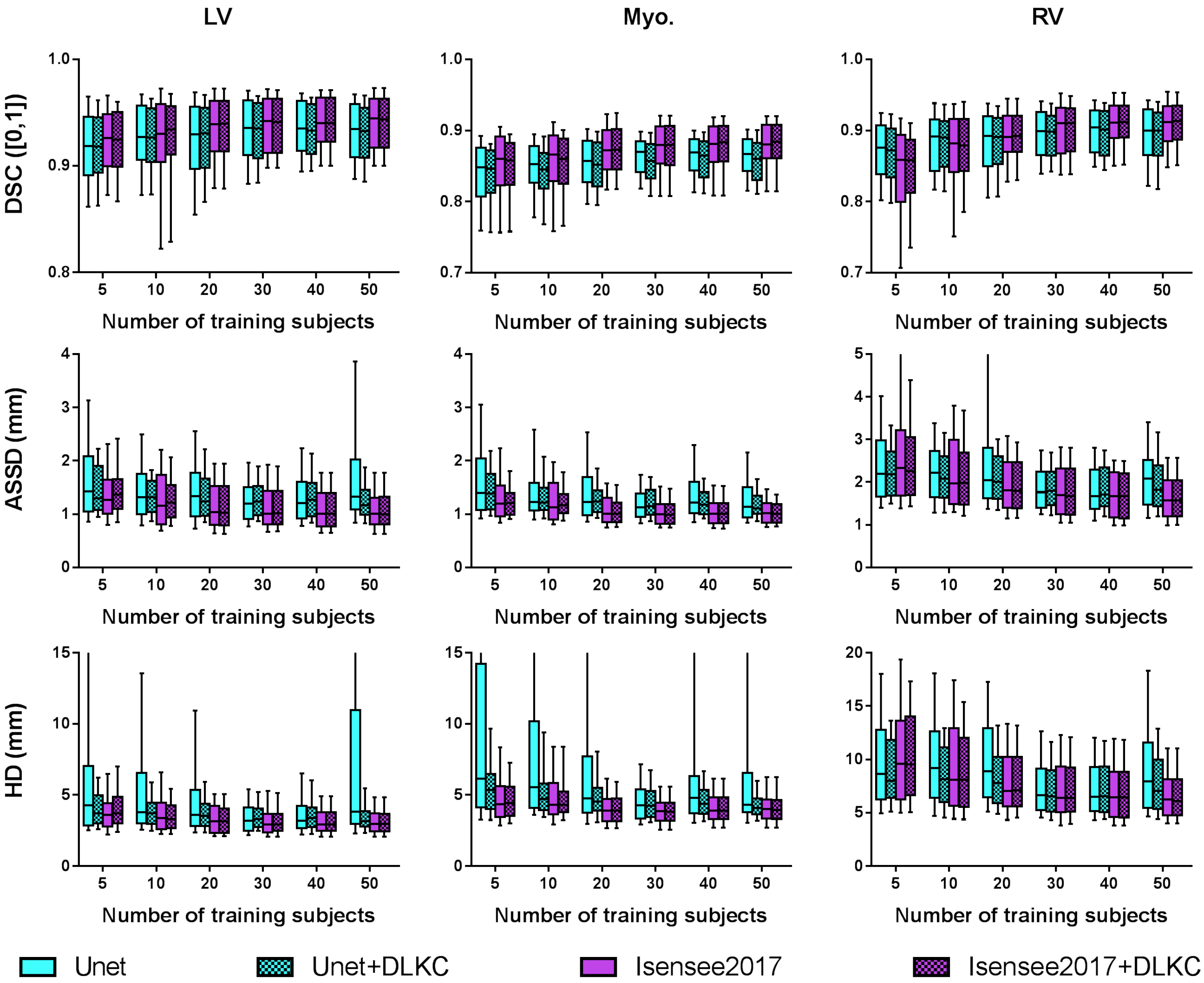

Figure 7: **UKBB** dataset segmentation results combining U-net/Isensee2017 and +DLKC (**U-net and Isensee2017 trained on the UKBB** training subjects). Box and whisker plots (box = 25-75th percentile; horizontal line = median; whisker = 10-90th percentile) from the top to bottom rows are shown for the distributions of DSC, ASSD, and HD for LV, Myo., and RV from left to right.

+DLKC/nDLKC. For example, supplementary Table S1 shows that baseline U-net yielded ASSD of 2.20 mm and HD of 9.8 mm of RV for NTrS=50, which are higher than ASSD of 2.15 mm and HD of 9.3 mm for NTrS=40, but this issue was mitigated by +DLKC. Table S2 shows that ASSD of LV provided by baseline U-nets was reduced from 1.41 mm to 1.35 mm for NTrS=40 and from 1.93 mm to 1.27 mm for NTrS=50. Similarly, HD was improved from 4.1 mm to 3.8 mm for NTrS=40 and from 9.1 mm to 3.6 mm for NTrS=50.

Furthermore, +DLKC/nDLKC led to more consistent segmentation accuracy across the experiments that utilized different numbers of training cases. For U-net trained and tested on the ACDC dataset (supplementary Table S1), ASSD of Myo. was {4.14, 2.90, 1.47, 1.28, 1.21, 1.17} mm (range=[1.17-4.14] mm) for NTrS={5, 10, 20 ,30, 40, 50} subjects. The variability and differences across the six sets of results were reduced as reflected by the refined ASSD of {1.90, 1.98, 1.37, 1.23, 1.24, 1.21} mm (range=[1.21-1.98] mm). For Isensee2017 trained on UKBB and tested on ACDC (supplementary Table S3), ASSD of Myo. was {3.19, 3.24, 2.27, 2.50, 2.37, 2.18} mm (range=[2.18-3.19] mm) for NTrS={5, 10, 20 ,30, 40, 50}. These results were improved to {2.34, 2.22, 1.70, 1.96, 1.85, 1.75} (range=[1.75-2.34] mm), reducing the variability and differences across the experiments using different NTrS. Similar results were observed for ASSD and HD, and for Isensee2017 trained on ACDC and tested on UKBB (supplementary Table S4).

4.6. *Influence of CNN Initial Outputs*

Supplementary Table S9 and S10 show the relationships of segmentation accuracy provided by +DLKC/nDLKC and baseline CNN for CNN trained and tested on the same and different datasets. For the majority of the experiments, we observed significant ($p < 0.05$) correlations between +DLKC/nDLKC and baseline CNN segmentation accuracy measurements. This suggests the influence of CNN initial outputs on the final segmentation, which motivates us to generate better CNN outputs to initialize the continuous kernel cut framework.

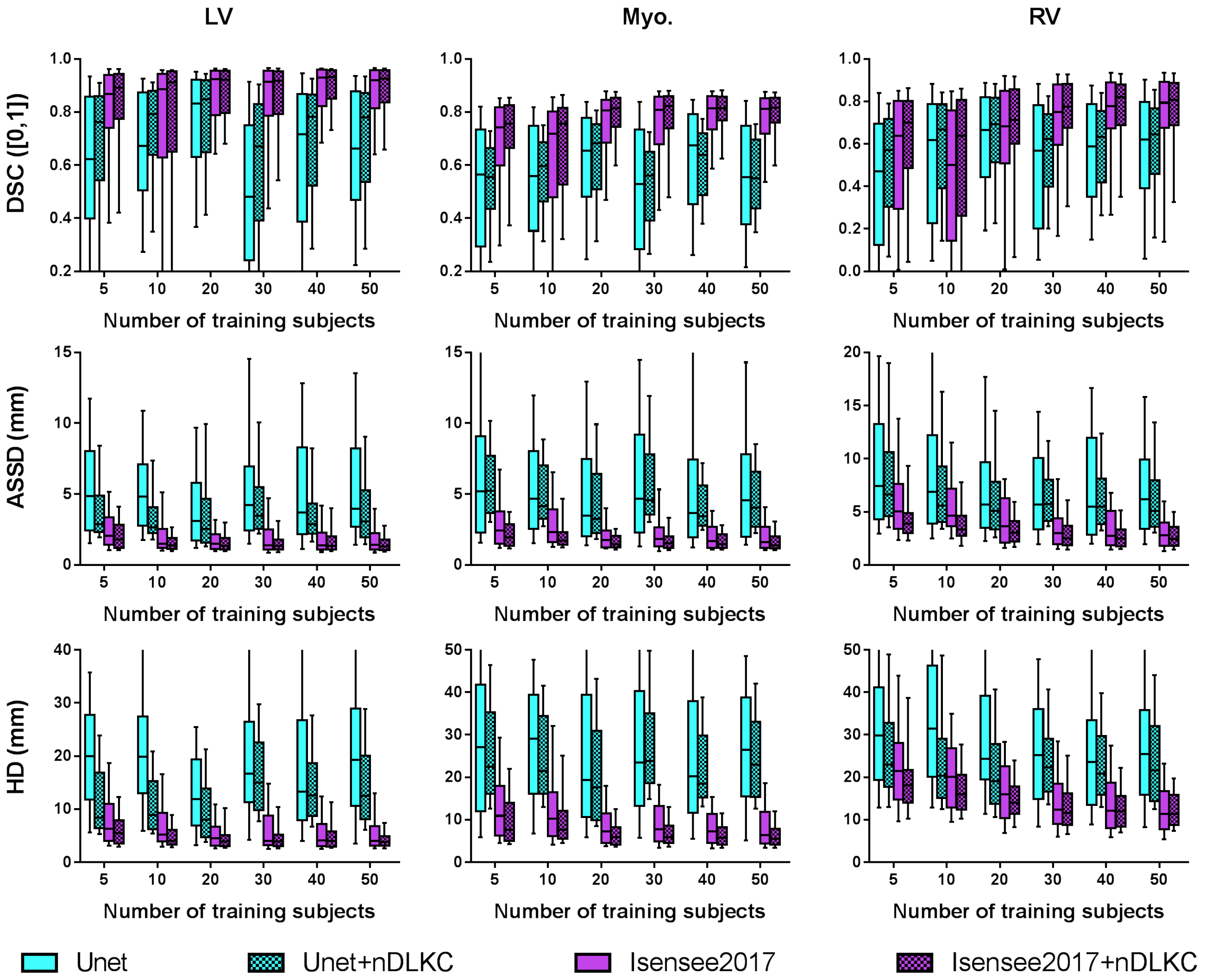

Figure 8: **ACDC** dataset segmentation results combining U-net/Isensee2017 and +nDLKC (**U-net and Isensee2017 trained on the UKBB** training subjects). Box and whisker plots (box = 25-75th percentile; horizontal line = median; whisker = 10-90th percentile) from the top to bottom rows are shown for the distributions of DSC, ASSD, and HD for LV, Myo., and RV from left to right.

4.7. Computational Efficiency

Finally, our approach required approximately 6 hours for U-net training, 2 s for U-net prediction, 0.5 s for CMF and 3.5 s for +DLKC/nDLKC post-processing for each subject. Isensee2017 required about ~24 hours for training each of the 5 U-net models in the ensemble, and 10 s for inference per subject.

5. Discussion

Cardiac MRI provides a wealth of imaging biomarkers with great potential for cardiovascular disease diagnosis and use in novel image-guided cardiac interventions. In this work, we developed and evaluated an automatic cardiac MRI segmentation approach, and demonstrated: 1) improved segmentation accuracy and reduced segmentation variability with minimal extra computational burden; 2) the capacity of training CNN using smaller training datasets and applying trained models to shifted dataset; 3) the advantages of continuous kernel cut within a continuous max-flow segmentation framework and the first investigation of this approach for cardiac MRI segmentation.

5.1. Continuous Kernel Cut vs *Dense CRF and CMF*

We achieved improved segmentation accuracy and reduced segmentation variability with continuous kernel cut compared with baseline U-net, and higher performance than dense CRF as shown in supplementary Table S1, S2 and S11. CNN provides unparalleled capability to learn representative image features and perform pixel-level labeling through feature classification. However, it is well known (Zheng et al., 2015; Chen et al., 2018; Kamnitsas et al., 2017) that CNN provides coarse outputs, including non-sharp/non-smooth segmentation boundaries, object localization errors, small isolated regions or holes, and spatial labeling inconsistency, leading to suboptimal results. A number of post-processing methods have been developed and some of them have demonstrated the capacity of

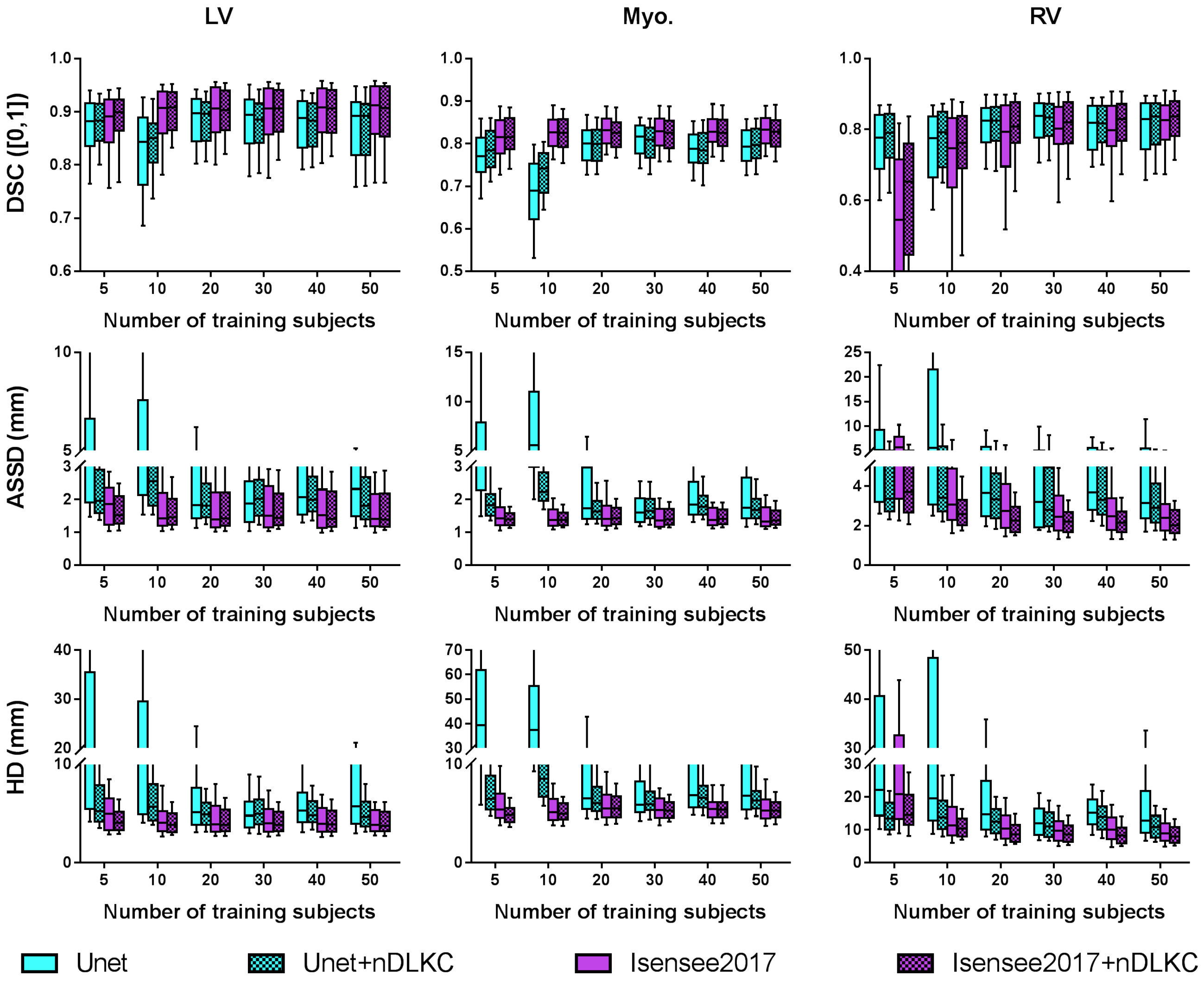

Figure 9: **UKBB** dataset segmentation results combining U-net/Isensee2017 and +nDLKC (**U-net and Isensee2017 trained on the ACDC** training subjects). Box and whisker plots (box = 25-75th percentile; horizontal line = median; whisker = 10-90th percentile) from the top to bottom rows are shown for the distributions of DSC, ASSD, and HD for LV, Myo., and RV from left to right.

improving CNN segmentation performance. We also implemented commonly used morphological operations and dense CRF. Unfortunately, we observed little-to-no and sometimes adverse effects on the U-net initial outputs using morphological operations. Previous studies reported positive (Dou et al., 2017; Zheng et al., 2015; Chen et al., 2018; Kamnitsas et al., 2017) and negative (Myronenko, 2018) effects using dense CRF post-processing, and our observations were consistent with these previous investigations. Recent studies (Veksler, 2019) showed that dense CRF exploited local optimization and the results can be far from optimum; perhaps this helps explain the previous and our current observations. In recognition of this issue, Veksler (2019) proposed to use discrete graph cut to globally optimize the dense CRF model and achieved excellent results. Other studies (Zheng et al., 2015) proposed to integrate a dense CRF as a recurrent layer of a CNN but this requires heavier computational burden (Chen et al., 2018). Perhaps more importantly, the resulting complex network structure may hinder the interpretation of CNNs, which represents one of the major obstacles in regulatory approvals and translation to clinical use (Lee et al., 2017; Pesapane et al., 2018; ESR, 2019).

Here we explored well-established global optimization methods in the continuous settings to improve CNN initial coarse segmentation. In addition, we employed normalized cuts for feature clustering because of the unique properties of balanced feature partitioning without a shrinking bias, which fundamentally differ from Potts and dense CRF methods, with superior performance (Tang et al., 2018; Veksler, 2019). Dangi et al. (2018) also employed deep learning, atlases, and discrete graph cut for cardiac MRI left ventricle segmentation. This approach (Dangi et al., 2018) employed a CNN for LV centre detection for slice misalignment correction, atlas generation and registration to a target image, and iterative discrete graph cut for refining the propagated atlas labels. Our approach substantially differs from the previous work in that: 1) we utilized CNN segmentation probability maps to initialize a continuous kernel cut model that is comprised of normalized cuts and continuous regularization; and 2) we developed a way to efficiently optimize the high-order segmentation model through upper bound optimization and convex relaxation within a spatially continuous max-flow framework, and demonstrated guaranteed algorithm convergence. For the experiments using U-net and Isensee2017 trained and tested on the same and different datasets, we observed overall improved segmentation accuracy provided by +DLKC/nDLKC, compared with baseline CNN. For U-net, the baseline segmentation was improved by +CMF and +DLKC/nDLKC because these methods explicitly model neighbourhood dependencies and encourage alignments between segmentation and image contrast edges, which are favoured to generate clinically meaningful segmentation. For example, +CMF yielded an average improvement of ~8% in ASSD and ~20% in HD for the 6 sets of U-net predictions on the ACDC test dataset, although there was slight decrease in DSC (supplementary Table S1). In addition, we previously combined U-net and CMF for dynamic cardiac MRI series LV and Myo. segmentation and biomarker quantification (MICCAI, 2018). Among all participants in the challenge, our approach (Guo et al., 2018) was one of the top three challenge winners (MICCAI, 2018). Compared with +CMF, here we observed much greater overall improvements over baseline U-net using +DLKC/nDLKC as shown in supplementary Table S1 and S2. This is consistent with the previously-observed superiority of normalized cuts to graph cut or GrabCut (Tang et al., 2018). This may be because of: 1) new image features incorporating voxel spatial location information that compensates for image signal intensity inhomogeneity; 2) the kernel $A(x, y)$ in Eq. (2) in +DLKC/nDLKC that encodes the proximity of features at any two pixels; and 3) the normalized cuts term that favors balanced partitioning without a shrinking effect, all of which were very difficult to obtain with +CMF. Compared with +nDLKC, we achieved up to 2-4% greater improvement in ASSD and HD using +DLKC for U-net trained and tested on the same dataset (supplementary Table S1 and S2). This may be because +DLKC enforced the similarity between U-net initial predictions and algorithm outputs as a way of exploiting CNN-learned deep features while +nDLKC utilized hand-crafted shallow image features with likely less discrimination power. Note that we used +DLKC when the initial predictions were reasonable, e.g., U-net and Isensee2017 trained and tested on the same datasets. For fairly low initial CNN segmentation, e.g., U-net and Isensee2017 trained and tested on different datasets, we implemented +nDLKC and the results are shown in supplementary Table S3 and S4. For all the experiments, we observed overall significant ($p<0.05$) correlations (Table S9 and S10) between algorithm refined and initial CNN results for DSC, ASSD and HD, suggesting the influence of CNN initialization on the final outputs. This motivates us to develop ways to generate better CNN initial outputs. We are planning to employ Monte-Carlo dropout in CNN to generate a "mean" of U-net predictions $\bar{\chi}$ and incorporate CNN uncertainty to further boost the segmentation performance as our next step work.

We achieved greater boosts in ASSD and HD compared with DSC and sometimes slight decrease in DSC when using +DLKC, as shown in Table 1 and 2 (details in Table S1, S2, S3 and S4). This might be because: 1) U-net was optimized for DSC on the validation dataset and Isensee2017 was trained using DSC loss, 2) manual segmentation was inconsistent in the UKBB datasets as we observed and previously reported (Zheng et al., 2018), 3) some slices (i.e., apex and base) were segmented by algorithms but not by expert observers and *vice versa*, 4) the baseline DSC is relatively high and DSC is less sensitive than surface distance metrics (Zheng et al., 2018). Regardless, the greater boosts in ASSD and HD suggest better localization of image edges, spatial labeling consistency, and smoothness of the final segmentation, which were identified as critical by our clinical collaborators for clinically-relevant imaging biomarker measurements such as myocardium wall thickness.

5.2. Influences of Baseline CNN Segmentation Accuracy on Continuous Kernel Cut Improvements

For the experiments in Sec. 4.3 and 4.4, we achieved general trend of lower segmentation accuracy using U-net, U-net+DLKC/nDLKC, compared with Isensee2017 and Isensee2017+DLKC/nDLKC. This is not surprising as the U-net we employed in this study represents a general CNN that is widely used for various medical image segmentation tasks (Isensee et al., 2019) whereas Isensee2017 was specifically optimized for cardiac MRI segmentation. In other words, the proposed +DLKC/nDLKC framework combined with the widely used U-net speaks to the utility of our approach for a variety of applications. The differences between U-net and Isensee2017 segmentation accuracies were substantially minimized with help of +DLKC/nDLKC. For U-net and Isensee2017 trained and tested on the same datasets using relatively large training cases (e.g., NTrS=20, 30, 40, 50) as shown in supplementary Table S1 and S2, we observed lower segmentation accuracy provided by U-net and U-net+DLKC, compared with Isensee2017 and Isensee2017+DLKC. However, these algorithm results are comparable to multi-observer multi-occasion expert manual segmentation (Sec. 5.3), which provides a reference for algorithm performance evaluation. For U-net and Isensee2017 trained and tested on the same dataset using small number of training subjects (e.g., NTrS=5, 10), we did observe higher segmentation accuracy using U-net *vs* Isensee2017 in a number of cases, e.g., DSC of RV for U-net trained (NTrS=5) and tested on ACDC in Fig. 6; DSC, ASSD and HD of RV for U-net trained (NTrS=5) and tested on UKBB in Fig. 7; DSC of RV for U-net trained (NTrS=10) on UKBB and tested on ACDC in Fig. 8; and DSC of RV for U-net trained (NTrS=5, 10) on ACDC and tested on UKBB in Fig. 9. For U-net and Isensee2017 trained (e.g., NTrS=5, 10, 20, 30, 40, 50) and tested on different datasets, which represents the vast majority of applications of CNN in research and clinical settings, we obtained fairly low baseline accuracy towards lower accuracy for U-net *vs* Isensee2017. These baseline U-net and Isensee2017 results were substantially improved and the differences between U-net and Isensee2017 results were much reduced with the help of +DLKC/nDLKC (supplementary Table S5 and S6). The lower segmentation accuracy provided by U-net+DLKC/nDLKC *vs* Isensee2017+DLKC/nDLKC was because of the lower initial performance provided by U-net *vs* Isensee2017, which is evidenced by the strong and significant correlations between +DLKC/nDLKC and baseline CNN accuracy measurements (supplementary Table S9 and S10). These results suggest the respective strengths of U-net and Isensee2017.

Table 1, 2, 3, and 4 show that +DLKC/nDLKC yielded different degrees of improvements in the mean and SD of initial CNN segmentation accuracy. For example, in situations when the initial CNN segmentation accuracy was high (e.g., U-net and Isensee2017 trained and tested on the same datasets with NTrS= 20, 30, 40, 50), we achieved relatively low improvements. We think that the relatively low improvements are because the baseline CNN segmentation accuracy is already high and comparable to manual outcomes (Sec. 5.3), which is difficult to surpass. The greater boost for U-net compared with Isensee2017 may be because of the lower initial segmentation accuracy provided by U-net *vs* Isensee2017. For initial CNN segmentation with moderate accuracy (e.g., U-net and Isensee2017 trained with NTrS=5, 10 training cases and tested on the same datasets), we obtained substantial improvements as shown in Table 1 and 2. In cases of initial CNN segmentation with fairly low accuracy (e.g., U-net and Isensee2017 trained with NTrS=5, 10, 20, 30, 40, 50 and tested on different datasets), we obtained high percent improvements. We believe that the latter two cases represent the vast majority of applications of CNNs in research and clinical settings in recognition of the costs for curating large datasets, the efforts needed for expert manual annotations, and the requirements of high-end computational resources. However, these problems are inherently challenging and, as expected, we achieved low segmentation accuracy using baseline U-net and Isensee2017. Here we developed a continuous kernel cut framework that provides a way to effectively alleviate these critical issues, facilitating broader applications of CNN for research and clinical applications. In particular, we demonstrated substantially improved mean and SD of the baseline U-net and Isensee2017 segmentation accuracy using +DLKC/nDLKC. Although the final segmentation accuracy was lower than U-net/Isensee2017 trained and tested on the same dataset using large training cases (supplementary Table S1 and S2), we think that our approach constitutes an important step towards solving these inherently difficult problems and enhancing the applicability of CNNs. We think that the final segmentation performance may be further improved by: 1) incorporating Monte-Carlo dropout and CNN uncertainty to generate better initial segmentation, 2) ensembling segmentation predictions provided by the same and different CNNs, 3) incorporating shape prior such as the circular shape of LV and Myo., and 4) exploiting the relationships between ED and ES phases, which represent our future research directions.

5.3. Improving Accuracy of CNNs Trained with Small Datasets

We obtained high percent improvements for U-net and Isensee2017 trained with small numbers of subjects (e.g., NTrS=5, 10 as shown in Table 1 and 2 and supplementary Table S1 and S2). It is well recognized that data-driven CNNs require large and diverse datasets for training and do not work well when the training dataset is small, which represents a major obstacle that hampers the widespread use of CNNs (Litjens et al., 2017). Although several large datasets are being made publicly available, acquisition of relevant annotations requires a significant amount of effort, e.g., expert availability, workload, cost, and annotation variability. In addition, the lack of diversity in disease phenotypes represents another issue. In this context, training CNNs with small datasets can be beneficial and our approach represents an important step towards facilitating wider use of CNN for medical image applications. To provide an example, we achieved LV segmentation DSC=0.921, ASSD=1.43 mm, and HD=3.9 mm for U-net trained (NTrS=20) and tested on the UKBB dataset (supplementary Table S2), which account for 98%, 73%, and 81% of the accuracy (DSC=0.940, ASSD=1.04 mm, HD=3.16 mm) provided by a benchmark CNN using 3,975 training subjects (Bai et al., 2018). Similarly, we obtained DSC of 0.870 for RV using U-net+DLKC (NTrS=20), which accounts for 95.8% of the accuracy (DSC= 0.908) yielded by Isensee2017 that used 100 subjects for training.

For the UKBB and ACDC databases, we note that multi-observer multi-occasion manual segmentation was performed by experienced experts as a reference for algorithm performance evaluation. For the UKBB dataset multi-observer multi-repetition manual segmentation (Bai et al., 2018), we computed mean manual DSC={0.930, 0.876, 0.880}, ASSD={1.17 mm, 1.19 mm, 1.88 mm}, and HD={3.13 mm, 3.76 mm, 7.35 mm} for {LV, Myo., RV}. For ACDC (Bernard et al., 2018), we derived mean expert manual segmentation DSC of {0.935, 0.889, 0.912} for {LV, Myo., RV}. The highest algorithm segmentation accuracy on the two datasets is higher than the respective expert manual segmentation accuracy. For example, Bai et al. (2018) reported algorithm DSC=0.940, ASSD=1.04 mm, and HD=3.16 mm for LV, which are higher than manual DSC=0.930, ASSD=1.17 mm, and HD=3.13 mm. Similarly, Isensee2017 obtained algorithm DSC of 0.945 *vs* manual segmentation DSC=0.935 for LV. We think studies that reported algorithm segmentation accuracy that is higher than expert manual outputs should be interpreted with caution. As shown in supplementary Table S1 and S2, for U-net and Isensee2017 trained and tested on the same dataset using relatively large (NTrS=20, 30, 40 ,50) training datasets, we achieved algorithm segmentation accuracy that is slightly lower than the highest algorithm accuracy in the respective dataset. Note that the previous studies (Bai et al., 2018; Isensee et al., 2017) trained and tested CNNs on the same datasets using large numbers of training cases, and the accuracy results were measured on test subjects that were different from our work. Regardless, our results are comparable to expert manual segmentation, suggesting that our approach is useful for research and clinical applications. For CNN trained and tested on the same dataset using small (NTrS=5, 10) training cases and for CNN trained (NTrS=5, 10, 20, 30, 40, 50) and tested on different datasets (supplementary Table S1 and S2), we obtained much improved performance. This is evidenced by the substantial percent improvement in the mean and SD of segmentation accuracy as shown in Table 1, 2, 3 and 4, which represents a main focus of this work. Although the final segmentation accuracy is lower than state-of-the-art algorithm results (CNNs trained and tested on the same datasets using large training cases) and manual outcomes, we want to emphasize that the proposed continuous kernel cut framework constitutes an important step towards solving these inherently challenging problems and we are planning to further optimize our algorithm in the future.

Recent studies contributed to a number of novel CNNs and some of them have demonstrated excellent performance for natural and medical image segmentation tasks, including but not limited to PSPnet, Dense U-net, and DeepLab. We also implemented DeepLabV3+ but achieved much lower segmentation accuracy compared with U-net and Isensee2017 (data not shown). For medical image segmentation, U-net and its variants represent a popular choice and were preferred and widely used by segmentation challenge winners (Bai et al., 2018; Bernard et al., 2018). The proposed continuous kernel cut module provides a simple and clinically-practical framework to improve CNN coarse initial outputs, which narrows the gap towards clinical translation of CNNs. Our approach is generalizable to other types of initializations, including manual, atlas, and any other CNNs. We also anticipate that our algorithm framework is able to alleviate the critical requirements of high quality expert manual annotations, which, once addressed, may further increase the applicability of CNN; this represents another future research direction.

5.4. Reducing CNN Segmentation Variability

We reduced baseline U-net and Isensee2017 segmentation variability within and across the experiments using +DLKC/nDLKC. This is evidenced by the lower SDs of DSC, ASSD and HD (supplementary Table S1, S2, S3 and S4), narrower range of the 25-75th/10-90th percentiles (Figs. 6, 7, 8, 9) that suggest less segmentation outliers. We also observed substantial variations by training the same network on the same dataset multiple times, which represent another limitation of CNN that was highlighted in the Kaggle Diabetic Retinopathy Challenge. In addition, baseline U-net provided lower segmentation accuracy when more subjects were used for training and this problem was alleviated with the help of +DLKC/nDLKC (e.g., ASSD of LV in Fig. 7). Furthermore, the differences in the mean and SD of segmentation accuracy using U-net/Isensee2017 trained with different numbers of training cases were substantially minimized and the final results were more consistent with less variability. In healthcare, there is a critical need for repeatable and reproducible radiological results (Pesapane et al., 2018). Our approach provides a way to mitigate the CNN reproducibility issue and may facilitate the use of clinically-relevant biomarkers in multi-center research and clinical trials.

5.5. Improving CNN Segmentation on Shifted Data

Another observation is the improved capacity of U-net and Isensee2017 segmentation on shifted data, i.e., CNN trained on one dataset and tested on other datasets, as shown in supplementary Table S3 and S4. In recognition of the limitations of CNN (e.g., requirements of large and diverse training datasets, high-quality manual annotations, and computational resources), a common strategy is to adopt pre-trained CNN models for specific clinical applications. However, it is well recognized and we have shown that CNN does not generalize well in the cases of data shift, which represents the vast majority of applications of CNNs in the research and clinical settings. This issue may be alleviated using transfer learning and domain adaptation but still remains an active research direction. Our approach provides an interpretable way to effectively mitigate this critical issue by substantially improving the initial U-net and Isensee2017 segmentation results by ~10% for DSC, ~20-30% for ASSD, and ~10-60% for HD for most of the experiments (supplementary Table S3 and S4). This highlights the practical value of using our approach to increase the applicability of CNNs for research and clinical applications.

Another advantage is that our algorithm provides rapid segmentation with minimal extra computational burden. This feature provides the potential to rapidly refine segmentation errors through simple and effective correction procedures (e.g., user interactions) when the computational burden becomes a major concern. An example (Wang et al., 2018) is image segmentation using CNN with interactive tuning, which requires fast response to user inputs with limited hardware resources.

While deep learning has achieved remarkable success in medical image analysis, this approach presents unique challenges. To address these issues and accelerate the applications for the medical imaging community, researchers are making efforts from many directions, including designing CNN with new architectures, incorporating prior domain knowledge, and investigating data pre/post-processing techniques. Here we focused on combining deep learning and straightforward and interpretable machine learning techniques. We proposed a continuous kernel cut segmentation approach and studied the challenging optimization problem using bound optimization and convex relaxation. We developed a novel iterative continuous max-flow algorithm that differs from previous work and applied our approach for cardiac MRI segmentation. We implemented a U-net that is widely used for various medical image segmentation tasks and Isensee2017 optimized for cardiac MRI segmentation. We utilized the outputs provided by baseline U-net and Isensee2017 to initialize the continuous kernel cut module. We achieved much improved segmentation accuracy, generalizability, and reduced segmentation variability for CNN trained with small training cases and tested on the same and different datasets. We note that it is possible to embed the continuous kernel cut module within a CNN and this represents a future direction for our work.

6. Conclusion

We developed an approach for cardiac MRI segmentation by integrating CNN, kernel cut, and bound optimization in a continuous max-flow framework. We comprehensively evaluated the performance of our approach and demonstrated improved segmentation accuracy and reduced segmentation variability, improved segmentation generalizability and the capacity of using small datasets for CNN training, and high computational efficiency that may be suitable for a broad range of applications.

Acknowledgements

We acknowledge the use of the facilities of the Shared Hierarchical Academic Research Computing Network and Compute Canada. We thank Meng Tang and Dr. Yuri Boykov for fruitful discussions and inputs in this work. This work was funded by Canadian Institutes of Health Research (CIHR) MOP: #93531, Ontario Research Fund and GE Healthcare. FG is supported by a postdoctoral fellowship from Natural Sciences and Engineering Research Council of Canada (NSERC-PDF).

SEP, SN and SKP acknowledge the British Heart Foundation for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5000 CMR scans (www.bhf.org.uk; PG/14/89/31194). SEP acknowledges support from the National Institute for Health Research (NIHR) Cardiovascular Biomedical Research Centre at Barts and from the SmartHeart EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1). SN and SKP are supported by the Oxford NIHR Biomedical Research Centre and the Oxford British Heart Foundation Centre of Research Excellence. This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (www.mrc.ac.uk; MR/L016311/1).

Disclosure

No conflicts of interest, financial or otherwise, are declared by F Guo, M Ng, M Goubran, SK Piechnik, S Neubauer, and G Wright. SEP acts as a paid consultant to Circle Cardiovascular Imaging Inc., Calgary, Canada and Servier.

References

Avendi, M., Kheradvar, A., Jafarkhani, H., 2016. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. Medical image analysis 30, 108–119.

Ayed, I.B., Gorelick, L., Boykov, Y., 2013. Auxiliary cuts for general classes of higher order functionals, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE. pp. 1304–1311.

Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., et al., 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance 20, 65.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE transactions on medical imaging 37, 2514–2525.

Bertsekas, D.P., 1999. Nonlinear programming. Athena scientific Belmont.

Blankstein, R., 2012. Introduction to noninvasive cardiac imaging. Circulation 125, e267–e271.

Boykov, Y., Kolmogorov, V., 2003. Computing geodesics and minimal surfaces via graph cuts, in: null, IEEE. p. 26.

Chambolle, A., 2004. An algorithm for total variation minimization and applications. Journal of Mathematical imaging and vision 20, 89–97.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848.

Dangi, S., Linte, C.A., Yaniv, Z., 2018. Cine cardiac MRI slice misalignment correction towards full 3D left ventricle segmentation, in: Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, International Society for Optics and Photonics. p. 1057607.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.A., 2017. 3D deeply supervised network for automated segmentation of volumetric medical images. Medical image analysis 41, 40–54.

ESR, 2019. What the radiologist should know about artificial intelligence-an ESR white paper. Insights into imaging 10, 44.

Ford, L., Fulkerson, D., 1962. Flows in networks. Princeton university press. Princeton, New Jersey 276.

Guo, F., Capaldi, D.P., McCormack, D.G., Fenster, A., Parraga, G., 2019. A framework for Fourier-decomposition free-breathing pulmonary 1H MRI ventilation measurements. Magnetic resonance in medicine 81, 2135–2146.

Guo, F., Ng, M., Wright, G., 2018. Cardiac MRI left ventricle segmentation and quantification: A framework combining U-net and continuous max-flow, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 450–458.

Guo, F., Svenningsen, S., Eddy, R., Capaldi, D., Sheikh, K., Fenster, A., Parraga, G., 2016. Anatomical pulmonary magnetic resonance imaging segmentation for regional structure-function measurements of asthma. Medical physics 43, 2911–2926.

Guo, F., Svenningsen, S., Kirby, M., Capaldi, D.P., Sheikh, K., Fenster, A., Parraga, G., 2017. Thoracic CT-MRI coregistration for regional pulmonary structure–function measurements of obstructive lung disease. Medical physics 44, 1718–1733.

Guo, F., Yuan, J., Rajchl, M., Svenningsen, S., Capaldi, D.P., Sheikh, K., Fenster, A., Parraga, G., 2015. Globally optimal co-segmentation of three-dimensional pulmonary 1H and hyperpolarized 3He MRI with spatial consistence prior. Medical image analysis 23, 43–55.

Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H., 2017. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features, in: International workshop on statistical atlases and computational models of the heart, Springer. pp. 120–129.

Isensee, F., Petersen, J., Kohl, S.A., Jäger, P.F., Maier-Hein, K.H., 2019. nnU-Net: Breaking the spell on successful medical image segmentation. arXiv preprint arXiv:1904.08128 .

Johnson, M., Duvenaud, D.K., Wiltschko, A., Adams, R.P., Datta, S.R., 2016. Composing graphical models with neural networks for structured representations and fast inference, in: Advances in neural information processing systems, pp. 2946–2954.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis 36, 61–78.

Khened, M., Kollerathu, V.A., Krishnamurthi, G., 2018. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. arXiv preprint arXiv:1801.05173 .

Klodt, M., Schoenemann, T., Kolev, K., Schikora, M., Cremers, D., 2008. An experimental comparison of discrete and continuous shape optimization methods, in: European Conference on Computer Vision, Springer. pp. 332–345.

Lee, J.G., Jun, S., Cho, Y.W., Lee, H., Kim, G.B., Seo, J.B., Kim, N., 2017. Deep learning in medical imaging: general overview. Korean journal of radiology 18, 570–584.

Li, L., Wu, F., Yang, G., Xu, L., Wong, T., Mohiaddin, R., Firmin, D., Keegan, J., Zhuang, X., 2019. Atrial scar quantification via multi-scale cnn in the graph-cuts framework. arXiv preprint arXiv:1902.07877 .

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

MICCAI, 2018. LVQuan18 dataset. https://lvquan18.github.io/ .

Mortazi, A., Burt, J., Bagci, U., 2017. Multi-planar deep segmentation networks for cardiac substructures from MRI and CT. arXiv preprint arXiv:1708.00983 .

Myronenko, A., 2018. 3D MRI brain tumor segmentation using autoencoder regularization, in: International MICCAI Brainlesion Workshop, Springer. pp. 311–320.

Ngo, T.A., Lu, Z., Carneiro, G., 2017. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Medical image analysis 35, 159–171.

Nieuwenhuis, C., Töppe, E., Cremers, D., 2013. A survey and comparison of discrete and continuous multi-label optimization approaches for the Potts model. International journal of computer vision 104, 223–240.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., ORegan, D.P., et al., 2018. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging 37, 384–395.

Organization, W.H., 2017. Cardiovascular diseases (CVDs). http://www.who.int/mediacentre/factsheets/fs317/en/ .

Payer, C., Štern, D., Bischof, H., Urschler, M., 2017. Multi-label whole heart segmentation using CNNs and anatomical label configurations, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 190–198.

Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S.E., Frangi, A.F., 2016. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magnetic Resonance Materials in Physics, Biology and Medicine 29, 155–195.

Pesapane, F., Codari, M., Sardanelli, F., 2018. Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. European radiology experimental 2, 35.

Petersen, S.E., Matthews, P.M., Francis, J.M., Robson, M.D., Zemrak, F., Boubertakh, R., Young, A.A., Hudson, S., Weale, P., Garratt, S., et al., 2015. UK Biobanks cardiovascular magnetic resonance protocol. Journal of cardiovascular magnetic resonance 18, 8.

Petitjean, C., Dacher, J.N., 2011. A review of segmentation methods in short axis cardiac MR images. Medical image analysis 15, 169–184.

Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C., et al., 2015. Right ventricle segmentation from cardiac MRI: a collation study. Medical image analysis 19, 187–202.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Rupprecht, C., Huaroc, E., Baust, M., Navab, N., 2016. Deep active contours. arXiv preprint arXiv:1607.05074 .

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence 22, 888–905.

Tang, M., Marin, D., Ayed, I.B., Boykov, Y., 2018. Kernel cuts: Kernel and spectral clustering meet regularization. International Journal of Computer Vision , 1–35.

Tran, P.V., 2016. A fully convolutional neural network for cardiac segmentation in short-axis MRI. arXiv preprint arXiv:1604.00494 .

Veksler, O., 2019. Efficient graph cut optimization for full CRFs with quantized edges. IEEE transactions on pattern analysis and machine intelligence .

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging 37, 1562–1573.

Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I., 2016. Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease, in: Reconstruction, Segmentation, and Analysis of Medical Images. Springer, pp. 95–102.

Wu, K.C., 2017. Sudden cardiac death substrate imaged by magnetic resonance imaging: From investigational tool to clinical applications. Circulation: Cardiovascular Imaging 10, e005461.

Yang, H., Sun, J., Li, H., Wang, L., Xu, Z., 2016. Deep fusion net for multi-atlas segmentation: Application to cardiac MR images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 521–528.

Yuan, J., Bae, E., Tai, X.C., Boykov, Y., 2010. A continuous max-flow approach to potts model, in: European conference on computer vision, Springer. pp. 379–392.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer. pp. 818–833.

Zheng, Q., Delingette, H., Duchateau, N., Ayache, N., 2018. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE transactions on medical imaging 37, 2137–2148.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1537.

Zhuang, X., 2013. Challenges and methodologies of fully automatic whole heart segmentation: a review. Journal of healthcare engineering 4, 371–407.